

Patents

R.J. Marks II
(1989-1990)

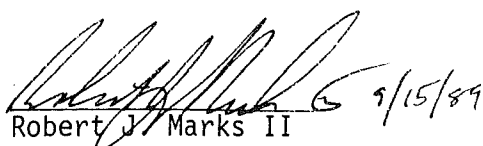
Patent Disclosure: Szasz Series Windows in Signal Processing

Robert J. Marks II

September 15, 1989

There exist a number of signal processing algorithms wherein a window, $\varphi(k)$, shifts across a signal to give an alternate representation of the signal. Included are weighted running averages, spectrograms and zamograms [1, 2, 3]. Conventionally, weighted running averages are computed using the equivalent of a finite impulse response (FIR) filter the taps of which correspond to the window samples. Digitally computed spectrograms are traditionally computed by weighting the signal samples in a interval by the window weights followed by a fast Fourier transform (FFT). Digital zamograms also require the use of FFT's for each point in time in which a spectral line is computed [3].

For windows and that are uniform (*i.e.* *rectangular* or *boxcar* windows), the value of a signal representation generated from a sliding window can be obtained by adding to the current representation new data introduced by the shift and deleting data no longer included in the window. With non-rectangular windows, however, shifting alters the weights of all data and the procedure is no longer applicable. An approach with similar computational advantages occurs when the window is of the form $\varphi(k) = e^{sk}$. Then, since $\varphi(k \pm 1) = e^{\pm s} e^{sk}$, shifting from k to $k \pm 1$ is equivalent to multiplying each data point by $e^{\pm s}$. Unfortunately, there are no useful windows that are exponential except the degenerate case of the rectangular window. There are, however, a number of commonly used windows that are superpositions of weighted exponentials. We refer to a weighted sum of exponentials as a *Szasz series* [4, 5]. Trigonometric polynomials are special cases. The Szasz components of the signal representation can be individually computed using


Robert J. Marks II 9/15/89

the exponential updating approach and the components superimposed to obtain the desired processing output. The generic procedure for the updating using Szasz windows, illustrated in Fig. 1 is:

1. In each Szasz component, subtract the terms that were in the previous window but not that in the current window. Likewise, add the newly introduced terms.
2. Multiply each of the elements common to both windows by the Szasz increment to effect the shift.
3. Add all of the Szasz components to obtain the desired outputs.

Two Szasz components may be complex conjugates of each other. In such cases, it is many times computationally convenient to combine the two components into a single composite component as shown in Fig. 2. Similarly, only the real portion of the output of a Szasz component may be required in certain cases.

In the next section, the Szasz series is reviewed. Application of the Szasz series to weighted running averages, spectrograms and zamograms are then presented.

1 Szasz Series Windows

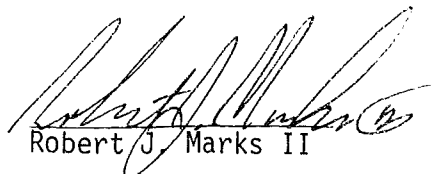
A linear exponent Szasz series can be written as

$$\varphi(k) = \sum_q \alpha_q e^{s_q k} \quad (1)$$

where the $\{\alpha_q\}$'s and the $\{s_q\}$'s are possibly complex. We will assume that there are Q terms in the sum. In certain cases, we require the kernel to be even. We then use the alternate form

$$\varphi_e(k) = \varphi(|k|) \quad (2)$$

Some popularly used windows and their Szasz series representations are in Tables 1 through 4. In each case, the Szasz series is an even trigonometric polynomial so that $\varphi_e(k) = \varphi(k)$. Each window is assumed to be zero for $|k| > L$. Other windows that are not exactly equal to a Szasz series can always be approximated to an arbitrary accuracy by a Szasz series.


Robert J. Marks II 7/15/89

α_q	s_q
0.42	0
0.25	$j\pi/L$
0.25	$-j\pi/L$
0.04	$j2\pi/L$
0.04	$-j2\pi/L$

Table 4: Blackman: $\varphi(k) = 0.42 + 0.5 \cos(\frac{\pi k}{L}) + 0.08 \cos(\frac{2\pi k}{L}), Q = 5$.

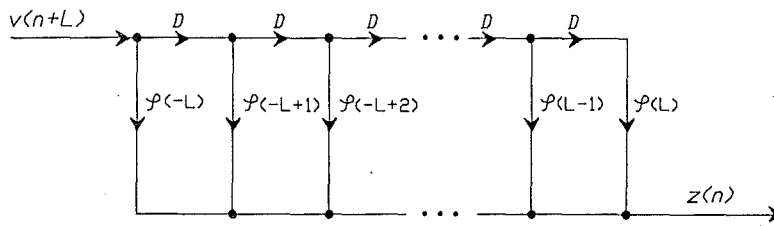


Figure 3: An FIR implementation of the weighted running average filter. The D denotes a unit delay.

2 Weighted Running Averages

The weighted running average, $z(n)$, of a signal, $v(n)$, is

$$z(n) = \sum_{k=-L}^L \varphi(k)v(n-k) \quad (3)$$

As is shown in Fig. 3, this process can be straightforwardly implemented on an FIR filter with $2L + 1$ taps.

If the Szasz series in Eq. 1 is used, we can write Eq. 3 as

$$z(n) = \sum_q z_q(n) \quad (4)$$

where

$$z_q(n) = \alpha_q \sum_{k=-L}^L e^{s_q k} v(n-k) \quad (5)$$

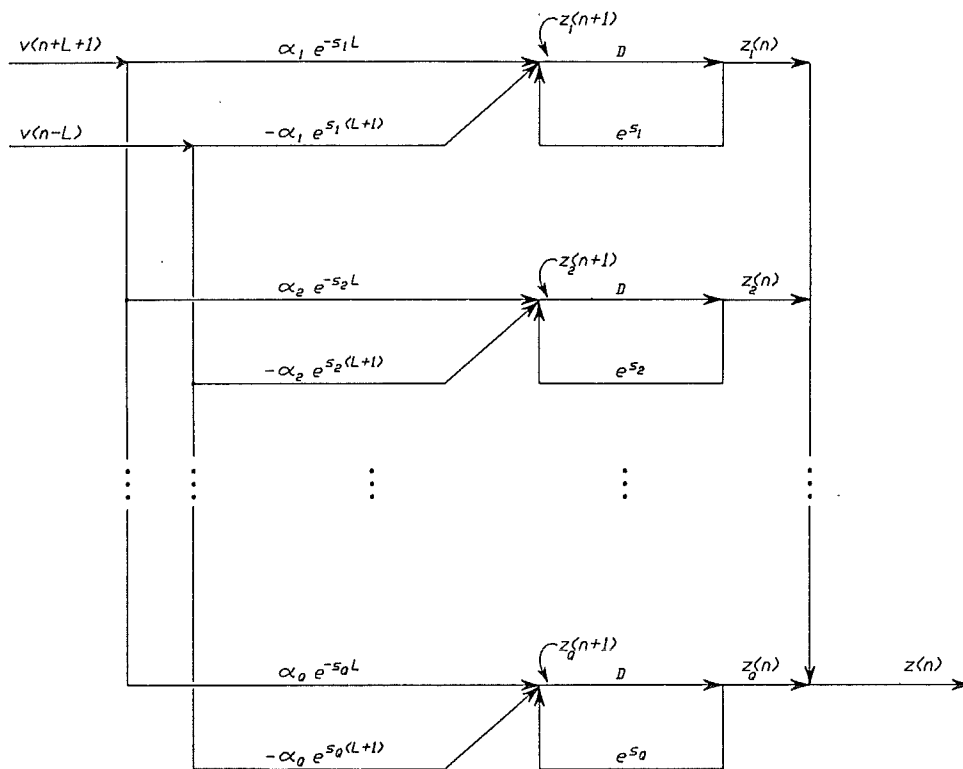


Figure 4: An IIR implementation of the weighted running average filter.

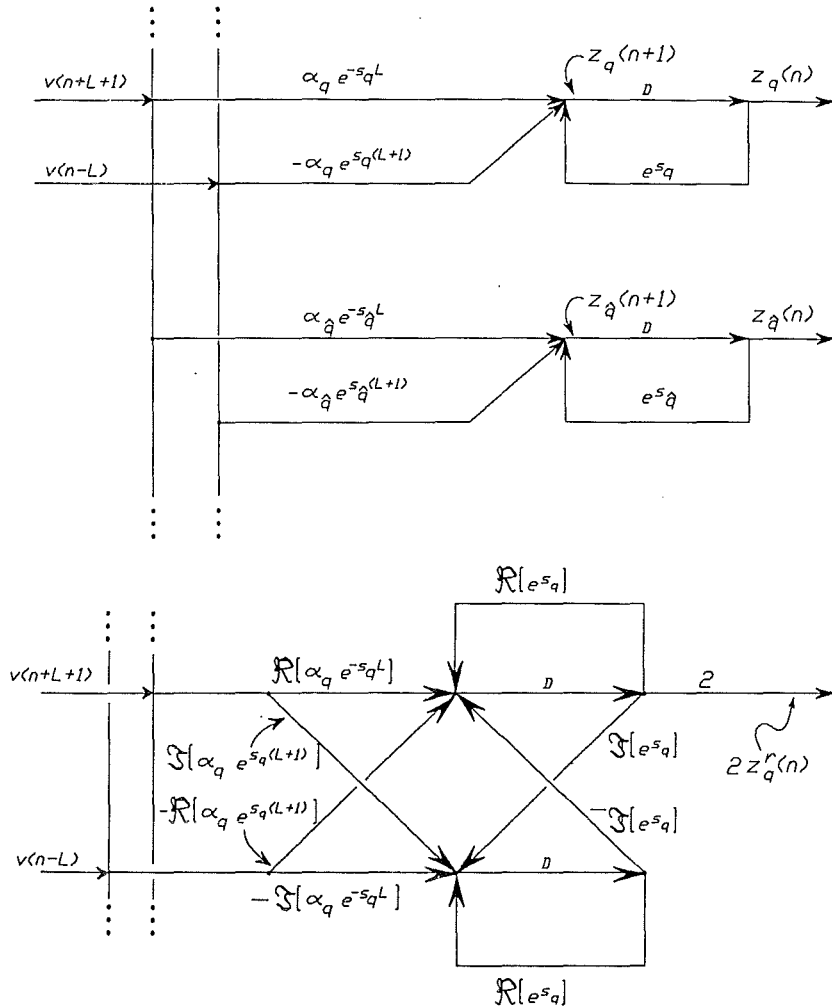


Figure 6: When two Szasz components are related by a complex conjugate, then the two components (shown here at the top) can be replaced by a single one (shown at the bottom).

3.2 Spectrogram computation using Szasz series components

If the window in Eq. 12 is expressed in terms of the Szasz series in Eq. 1, then the spectrogram in Eq. 12 can be written as

$$S(n, p) = \sum_q S_q(n, p) \quad (13)$$

where

$$S_q(n, p) = \alpha_q \sum_{k=-L}^L e^{s_q k} v(n - k) e^{-j2\pi p k / M} \quad (14)$$

The q^{th} Szasz component update is calculated as follows.

$$\begin{aligned} S_q(n+1, p) &= \alpha_q \sum_{k=-L}^L e^{s_q k} v(n+1 - k) e^{-j2\pi p k / M} \\ &= \alpha_q \sum_{\hat{k}=-L-1}^{L-1} e^{s_q(\hat{k}+1)} v(n - \hat{k}) e^{-j2\pi p(\hat{k}+1) / M} \\ &= \alpha_q e^{s_q} e^{-j2\pi p / M} \sum_{\hat{k}=-L-1}^{L-1} e^{s_q \hat{k}} v(n - \hat{k}) e^{-j2\pi p \hat{k} / M} \\ &= e^{s_q} e^{-j2\pi p / M} S_q(n, p) + \alpha_q e^{-L s_q} e^{j2\pi p L / M} v(n + L + 1) \\ &\quad - \alpha_q e^{(L+1)s_q} e^{-j2\pi p(L+1) / M} v(n - L) \end{aligned} \quad (15)$$

We are again following the procedure outlined in Fig.1. The new data is $\alpha_q e^{-L s_q} e^{j2\pi p L / M} v(n + L + 1)$, the old data is $\alpha_q e^{(L+1)s_q} e^{-j2\pi p(L+1) / M} v(n - L)$ and the Szasz factor is $e^{s_q} e^{-j2\pi p / M}$. Implementation of the specific iteration in Fig. 12 iteration is shown in Fig. 9. Since multiplication of the inputs by the arrays $e^{j2\pi p L / M}$ and $e^{-j2\pi p(L+1) / M}$ is common to each of the Q Szasz components, the alternate implementation shown in Fig. 10 is possible.

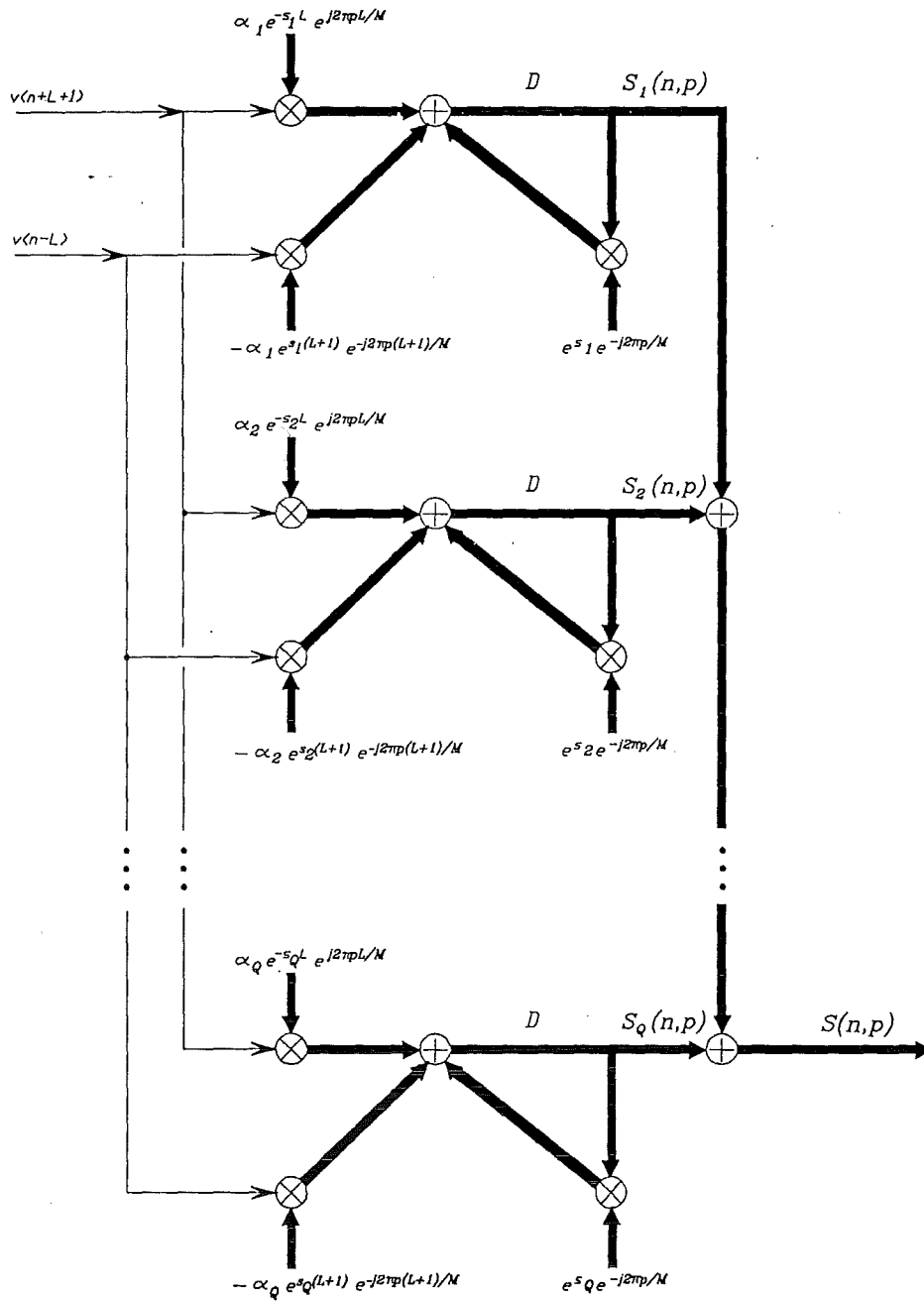


Figure 9: Computation of the spectrogram when the window is represented as a Q th order Szasz series. The thick lines correspond to signal flow directions of vectors parameterized by the frequency variable, p . The thin lines correspond to (possibly complex) scalars.

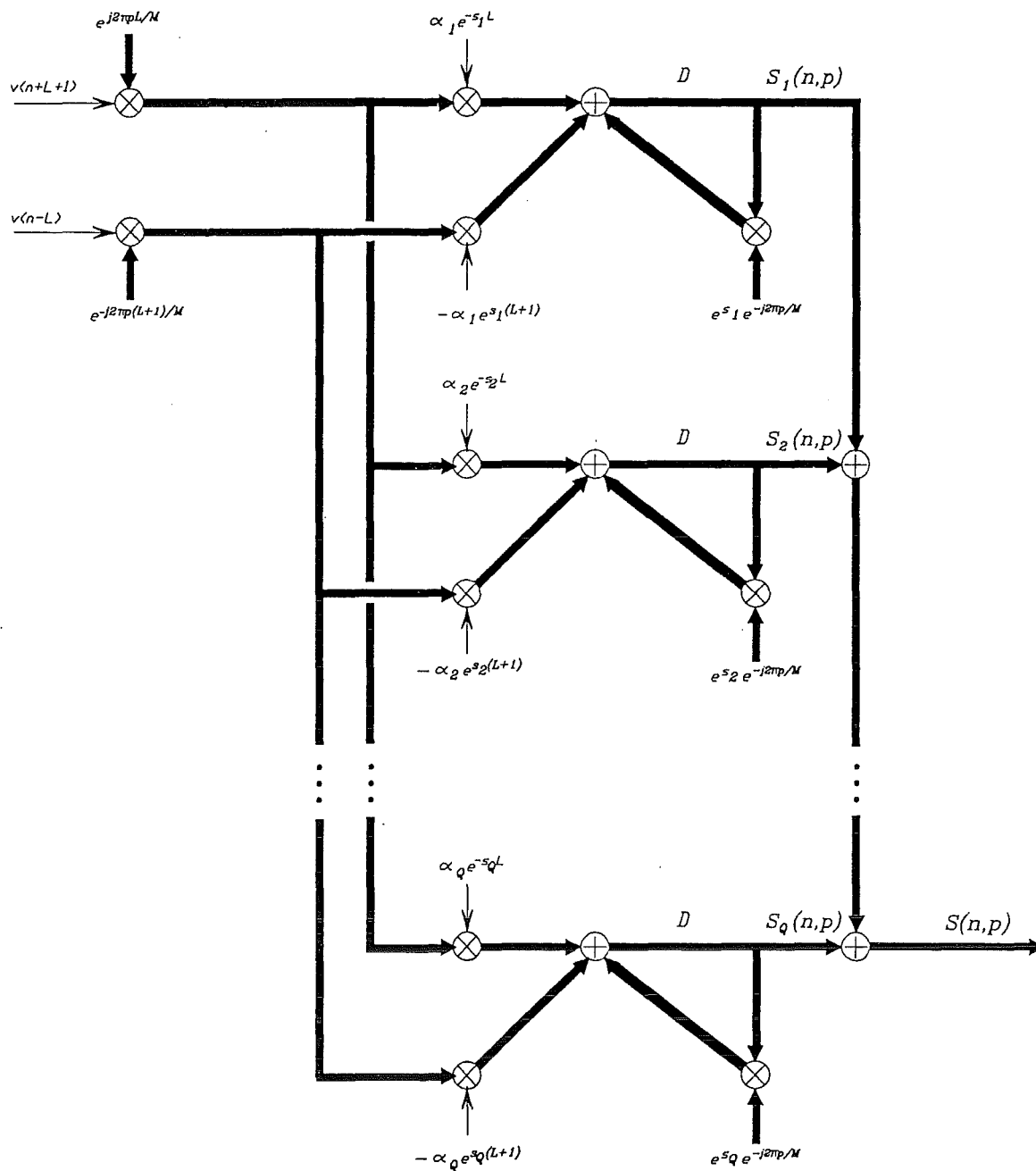


Figure 10: A second technique for computation of the spectrogram when the window is represented as a Q th order Szasz series

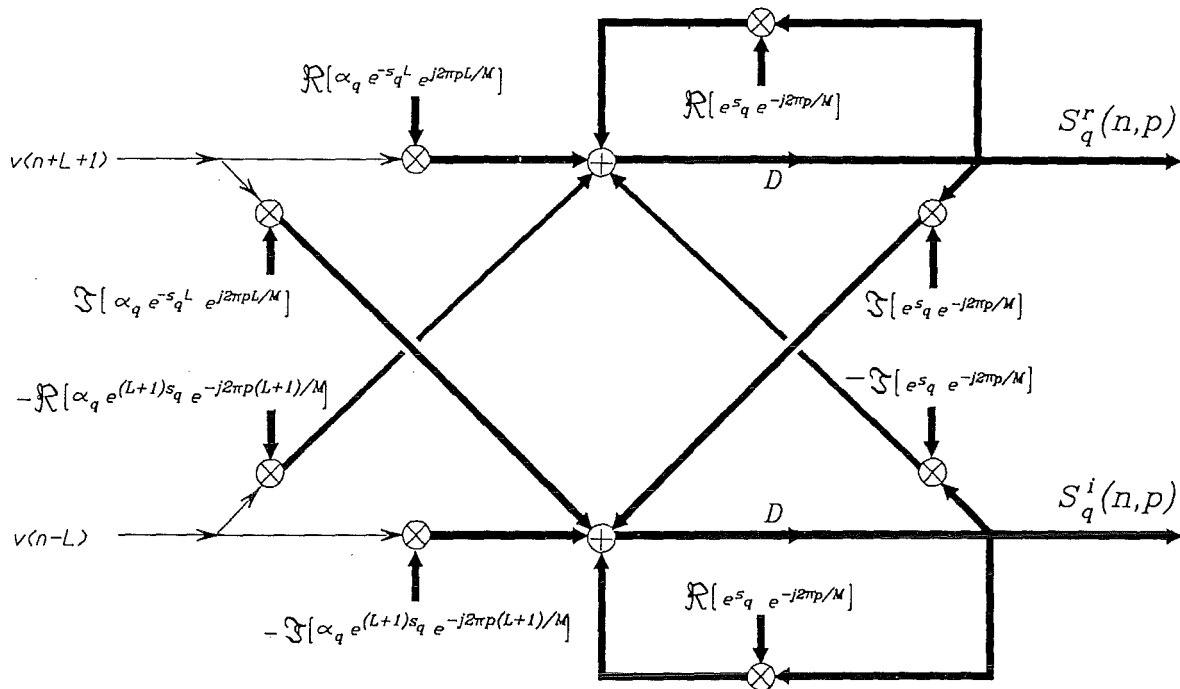


Figure 11: When a Szasz component of a spectrogram is complex, it's real and imaginary components can be realized as shown here. The real and imaginary components of the spectrogram are obtained by summing the real and imaginary components of the Szasz components.

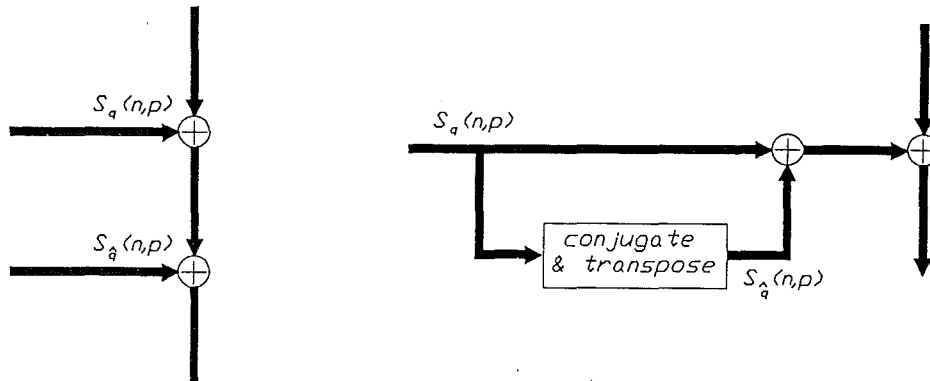


Figure 12: The two Szasz components of a spectrogram indexed by q and \hat{q} shown on the left can be obtained by simple augmentation of the output of the q th Szasz component as shown on the right. Transposition replaces p by $-p$ in the array $S_q(n, p)$.

This relationship, as illustrated in Fig. 12, can be used to obtain the sum of two Szasz components, indexed by q and \hat{q} , by a simple augmentation of the output of the Szasz component with index q . The equivalent operation using the real and imaginary outputs of the Szasz component in Fig. 11 is shown in Fig. 13.

3.2.3 Example: Hanning and Hamming windowed spectrograms

In Fig. 14 we illustrate application of the Szasz series computation of a spectrogram for the a $Q = 3$ case when α_1 is real, $s_1 = 0$, $\alpha_2 = \alpha_3$ and $s_2 = s_3^* = j\pi/L$. The Hanning (Table 2) and Hamming (Table 3) windows are special cases.

4 Zamograms

The zamogram is a display of high resolution time-frequency displays with good resolution in both domains. In the discrete domain, the zamogram of

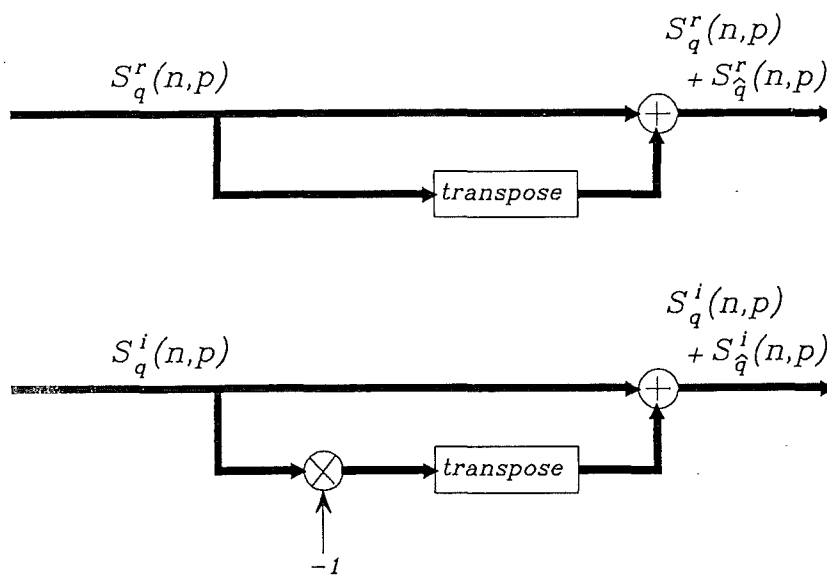


Figure 13: The real and imaginary components of the q th component of a Szasz component can be straightforwardly augmented to give the sum of the real and imaginary parts of two Szasz components.

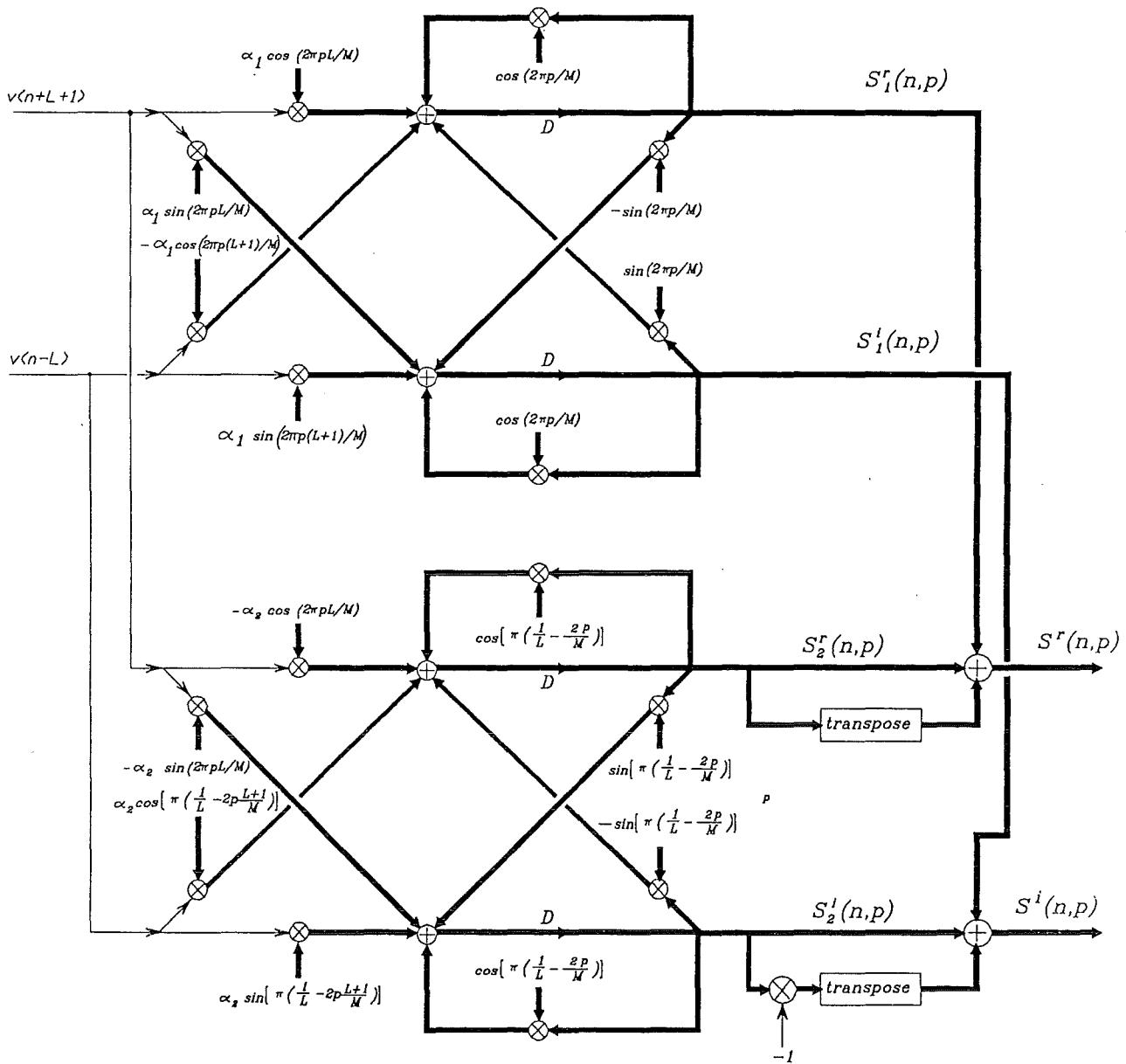


Figure 14: Generation of a spectrogram using Szasz components. Hanning & Hamming windowed spectrograms can both be thusly implemented.

Robert J. Marks II 9/15/85
 Robert J. Marks II

The proof of these equations is straightforward. Let λ_n^- be the set on points in Λ_n but not in Λ_{n+1} . Then

$$\lambda_n^- = \{(m, k) \mid m = \frac{|k|}{2} + n + 1; |k| \leq 2L\} \quad (33)$$

Similarly, let λ_n^+ denote the set of points in Λ_{n+1} that are not in Λ_n . Thus

$$\lambda_n^+ = \{(m, k) \mid m = -\frac{|k|}{2} + n; |k| \leq 2L\} \quad (34)$$

Clearly, then

$$\begin{aligned} C(n+1; p) = & \left[\sum_{(m,k) \in \Lambda_n} + \sum_{(m,k) \in \lambda_n^+} - \sum_{(m,k) \in \lambda_n^-} \right] \varphi(k) \\ & \times x\left(m + \frac{k}{2}\right) x\left(m - \frac{k}{2}\right) e^{-j2\pi pk/M} \end{aligned} \quad (35)$$

or, equivalently,

$$C(n+1; p) = C(n; p) + B_n^+(m) - B_n^-(m) \quad (36)$$

where

$$B_n^\pm(p) = \sum_{(m,k) \in \lambda_n^\pm} \varphi(k) x\left(m + \frac{k}{2}\right) x\left(m - \frac{k}{2}\right) e^{-j2\pi pk/M} \quad (37)$$

Equivalently, we can write

$$B_n^+(p) = 2\Re x^*(n+1)\beta^+(n, p). \quad (38)$$

and

$$B_n^-(p) = 2\Re x^*(n)\beta^-(n, p) \quad (39)$$

Substituting this and Equation(38) into Equation(36) establishes Equation(30) and the proof is complete.

4.1.1 Using Fast Fourier Transforms

We will now present two techniques to evaluate the iterations in Eq. 30.

A signal flow graph at time n is shown in Fig. 15 for direct evaluation of Equation(30). The sample signals are introduced into a shift register as

shown on the left. The shift register is tapped and each of the samples is multiplied by stored weights, $\{\varphi(k)\}$, as shown. The two vectors of the windowed samples are fed into two pipelined FFT processors. Transposition of the output of the lower FFT is required because there is a $e^{j2\pi pk/M}$ term in Equation(32) rather than the $e^{-j2\pi pk/M}$ used in Equation(31). The transposition replaces k with $-k$ to take care of this. The delays in Fig. 15 are required to synchronize the samples $x(n)$ and $x(n+1)$ with the computational delays required in the processing to that point (*e.g.* by the FFT). These two samples are weighted by either ± 2 after which they multiply every element of the output of the FFT processors. The real part of the resulting two vectors are summed. The sum is added to the current zamogram register, and a new spectral line of the zamogram emerges in vector form from the processor. The parameter Δ is the total number of clock cycles required from input to output.

4.1.2 Using a Szasz Window

A second implementation is possible when the zamogram's kernel is expressed as the Szasz series in Eq. 1. The iteration in Equation(30) can be written as

$$\begin{aligned}
 C(n+1; p) = & C(n; p) + [|x(n+1)|^2 - |x(n)|^2]\varphi(0) \\
 & + 2\Re[x^*(n+1) \sum_q b_q^+(n, p) \\
 & - x^*(n) \sum_q b_q^-(n, p)] \quad (40)
 \end{aligned}$$

where the Szasz components, $b_q^\pm(n, p)$, can be updated as

$$\begin{aligned}
 b_q^+(n, p) = & e^{-(s_q - \frac{j2\pi p}{M})} b_q^+(n-1, p) \\
 & - \alpha_q x(n+1) + \alpha_q e^{-2L(s_q - \frac{j2\pi p}{M})} x(n+2L+1) \quad (41)
 \end{aligned}$$

and

$$\begin{aligned}
 b_q^-(n, p) = & e^{(s_q - \frac{j2\pi p}{M})} b_q^-(n-1, p) \\
 & + \alpha_q x(n-1) - \alpha_q e^{2L(s_q - \frac{j2\pi p}{M})} x(n-2L-1) \quad (42)
 \end{aligned}$$

A proof will be presented after some discussion.

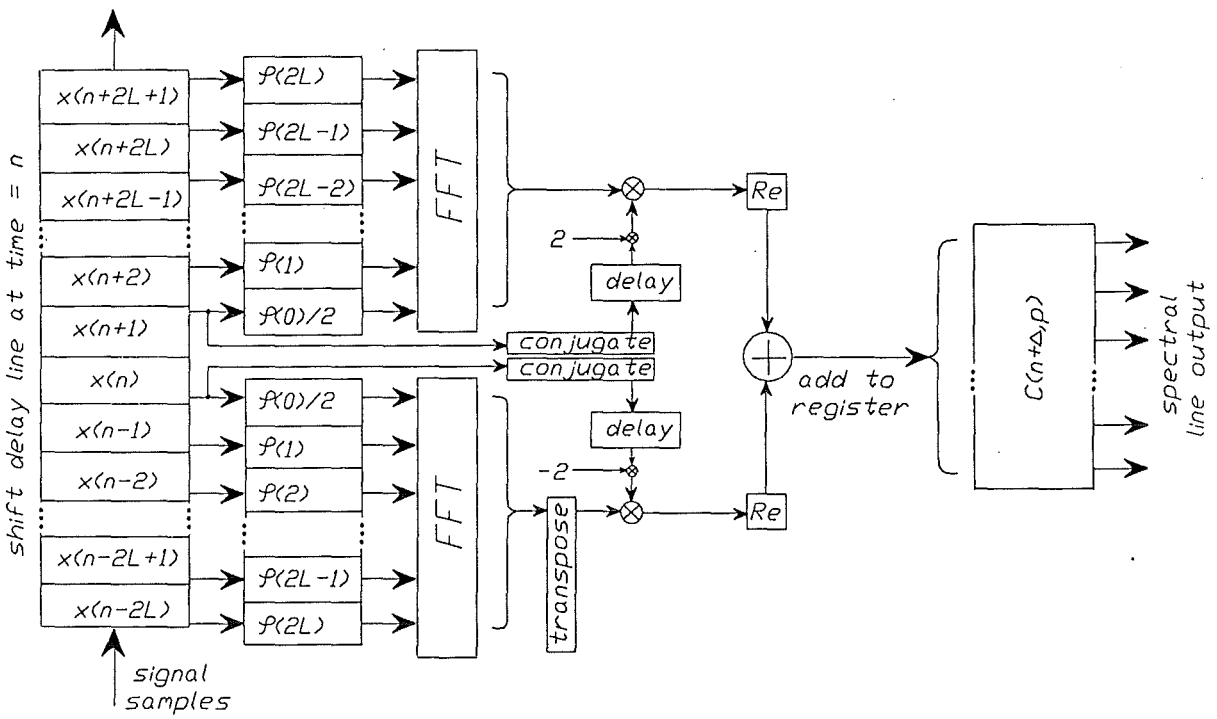


Figure 15: Iterative updating of a zomogram using FFT's.

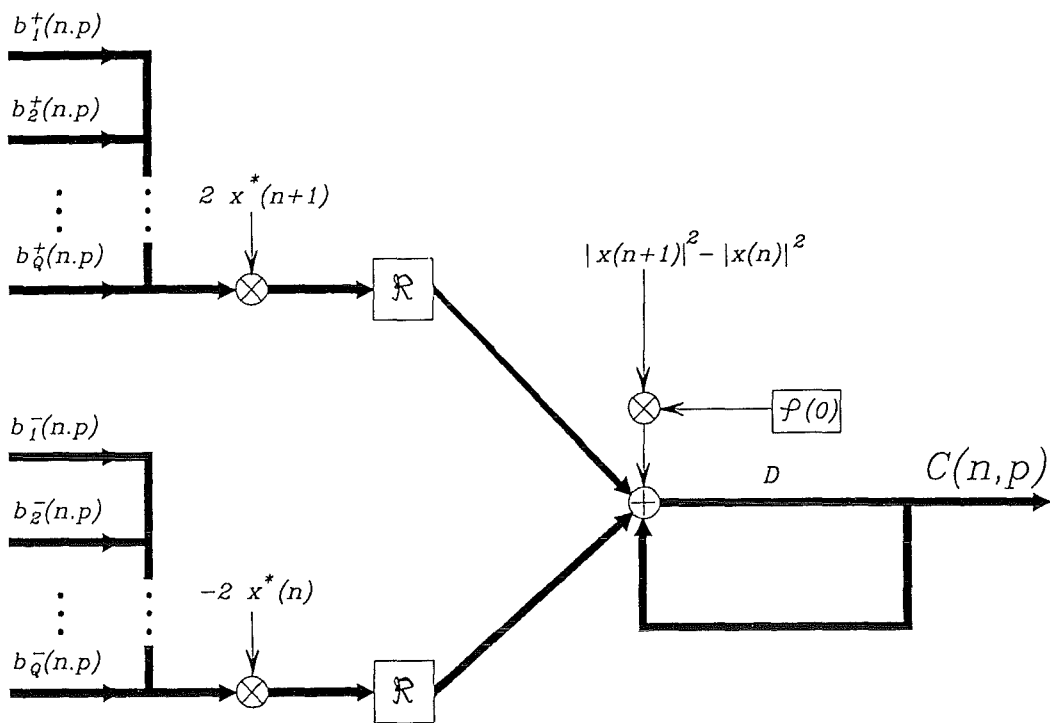


Figure 16: Iterative updating of a zomogram using Szasz components b_q^\pm .

A signal flow diagram for the recursion in Eq. 40 is shown in Fig. 16. Unlike the FFT implementation, we here need to tap the shift register at only five points $x(n - 2L - 1), x(n - 1), x(n), x(n + 1)$ and $x(n + 2L + 1)$.

We can express the complex $b_q^\pm(n, p)$'s in terms of their real and imaginary components as

$$b_q^\pm(n, p) = b_q^{\pm r}(n, p) + j b_q^{\pm i}(n, p) \quad (43)$$

Similarly, let

$$x(n) = x^r(n) + j x^i(n) \quad (44)$$

A corresponding implementation equivalent to that in Fig. 16 is shown in Fig. 17 using real arithmetic.

Note that both Eqs. 41 and 42 are iterations of Szasz components as illustrated in Fig. 1. The Szasz factors are $\exp \pm (s_q - \frac{j2\pi p}{M})$. For Eq. 41, the new data is $\alpha_q \exp[-2L(s_q - \frac{j2\pi p}{M})]x(n + 2L + 1)$ and the old data is $\alpha_q x(n + 1)$. In Eq. 42, the old data is $\alpha_q \exp[2L(s_q - \frac{j2\pi p}{M})]x(n - 2L - 1)$ and the new data is $\alpha_q x(n - 1)$. Implementation of the updates of the b_q^\pm 's in Eqs. 41 and 42 are illustrated in Fig. 18.

Proof: To show Eqs. 40, 41 and 42, we substitute Equation(1) into Eq. 37:

$$B_n^\pm(p) = \sum_{(m,k) \in \lambda_n^\pm} \sum_q \alpha_q e^{s_q |k|} x(m + \frac{k}{2}) x(m - \frac{k}{2}) e^{-j2\pi p k / M} \quad (45)$$

Using the definition in Eq. 33, we find that

$$B_n^+(m) = |x(n + 1)|^2 \varphi(0) + 2\Re x^*(n + 1) \sum_q b_q^+(n, p) \quad (46)$$

where

$$b_q^+(n, p) = \alpha_q \sum_{k=1}^{2L} e^{-s_q k} x(n + k + 1) e^{-j2\pi m k / M} \quad (47)$$

The recursive form in Equation(41) can easily be established from Equation(47).

Similarly,

$$B_n^-(p) = |x(n)|^2 \varphi(0) + 2\Re x^*(n) \sum_q b_q^-(n, p) \quad (48)$$

where

$$b_q^-(n, p) = \alpha_q \sum_{k=1}^{2L} e^{-s_q k} x(n - k) e^{j2\pi m p / M} \quad (49)$$

The recursion in Equation(42) follows and the proof is complete.

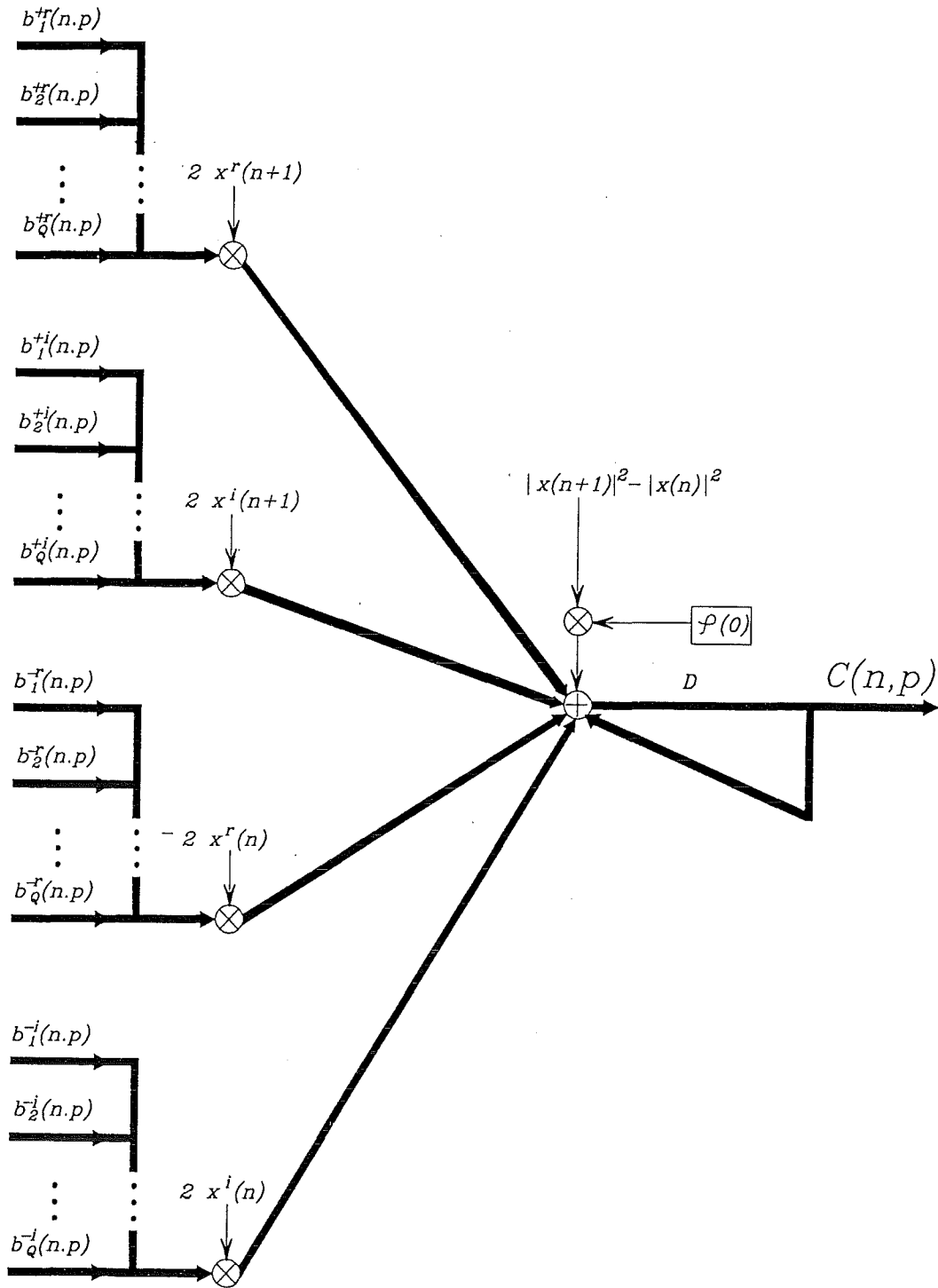


Figure 17: Iterative updating of a zomogram using Szasz components and real arithmetic.

Robert J. Marks II 9/15/89
 Robert J. Marks II

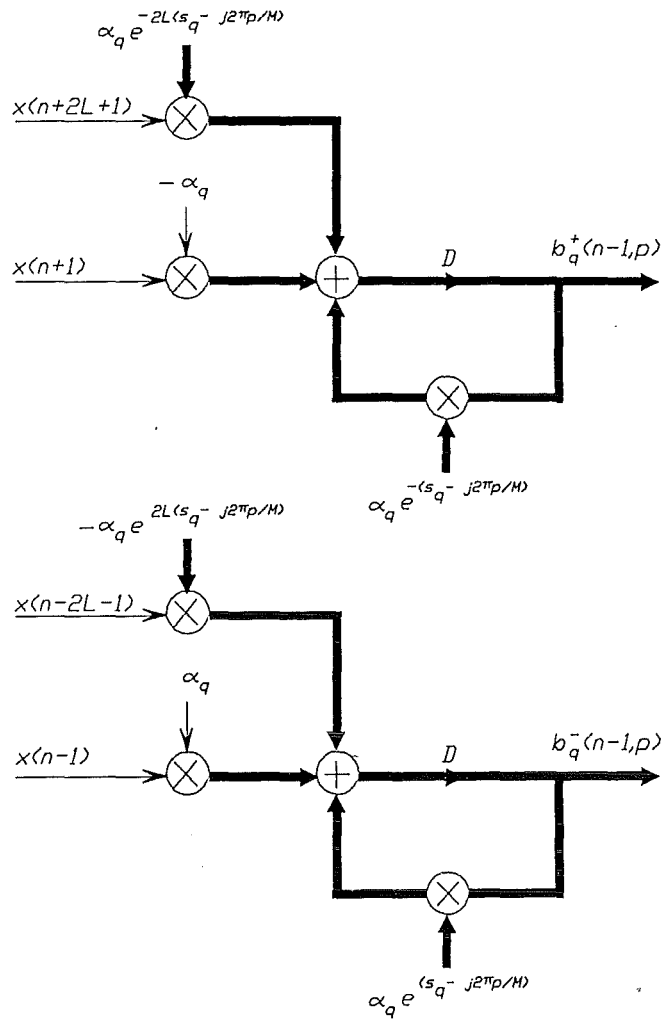


Figure 18: Iterative updating of the Szasz components for the zamogram.

Realizing the real & imaginary parts of a Szasz component of a zamogram: Assume that the signal, $x(n)$, is real. From Eq. 41, the real and imaginary components of $b_q^\pm(n, p)$ follow as

$$b_q^{+r}(n, p) = \Re[e^{-(s_q - \frac{j2\pi p}{M})}]b_q^{+r}(n-1, p) - \Im[e^{-(s_q - \frac{j2\pi p}{M})}]b_q^{+i}(n-1, p) \\ - \Re[\alpha_q]x(n+1) + \Re[\alpha_q e^{-2L(s_q - \frac{j2\pi p}{M})}]x(n+2L+1) \quad (50)$$

and

$$b_q^{+i}(n, p) = \Im[e^{-(s_q - \frac{j2\pi p}{M})}]b_q^{+r}(n-1, p) + \Re[e^{-(s_q - \frac{j2\pi p}{M})}]b_q^{+i}(n-1, p) \\ - \Im[\alpha_q]x(n+1) + \Im[\alpha_q e^{-2L(s_q - \frac{j2\pi p}{M})}]x(n+2L+1) \quad (51)$$

The computational algorithm shown at the top of Fig. 19 implements these equations.

Similarly, from Eq. 42, the real and imaginary components of $b_q^+(n, p)$ are

$$b_q^{-r}(n, p) = \Re[e^{(s_q - \frac{j2\pi p}{M})}]b_q^{-r}(n-1, p) - \Im[e^{(s_q - \frac{j2\pi p}{M})}]b_q^{-i}(n-1, p) \\ + \Re[\alpha_q]x(n-1) - \Re[\alpha_q e^{2L(s_q - \frac{j2\pi p}{M})}]x(n-2L-1) \quad (52)$$

and

$$b_q^{-i}(n, p) = \Im[e^{(s_q - \frac{j2\pi p}{M})}]b_q^{-r}(n-1, p) + \Re[e^{(s_q - \frac{j2\pi p}{M})}]b_q^{-i}(n-1, p) \\ + \Im[\alpha_q]x(n-1) - \Im[\alpha_q e^{2L(s_q - \frac{j2\pi p}{M})}]x(n-2L-1) \quad (53)$$

These two equations are implemented at the bottom of Fig. 19.

If $x(n)$ is real and $\varphi(k)$ is real and even, then an inspection of Eq. 24 reveals that $C(n, p)$ is also real. In this case, Eq. 40 can be written as

$$C(n+1; p) = C(n; p) + [x^2(n+1) - x^2(n)]\varphi(0) \\ + 2x(n+1) \sum_q b_q^{+r}(n, p) \\ - 2x(n) \sum_q b_q^{-r}(n, p) \quad (54)$$

With reference to Fig. 19, the $2b_q^\pm(n, p)$ terms can be generated as shown in Fig. 20.

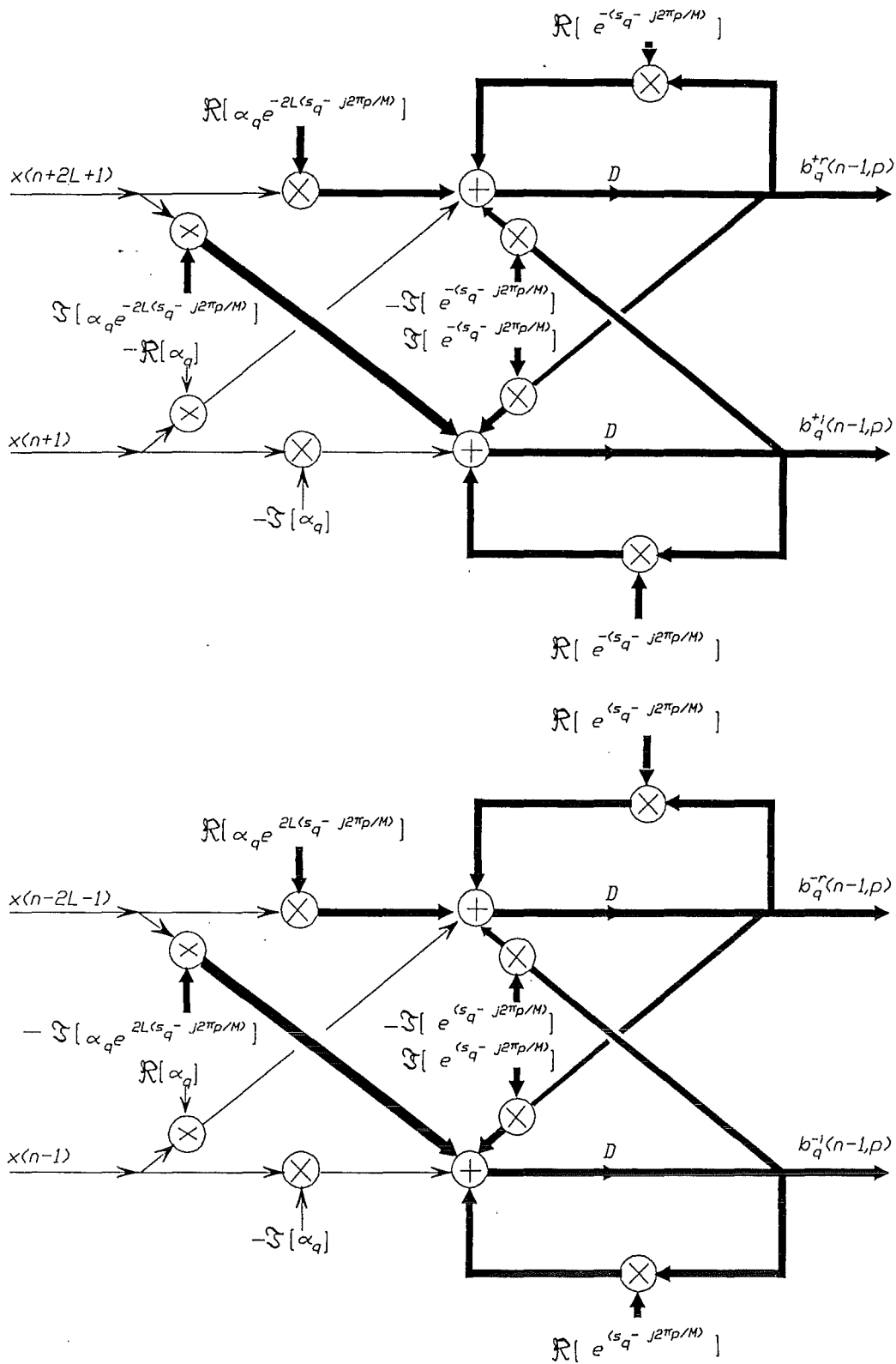


Figure 19: Evaluating the real and imaginary parts of $b_q^+(n,p)$ (top) and $b_q^-(n,p)$ (bottom).

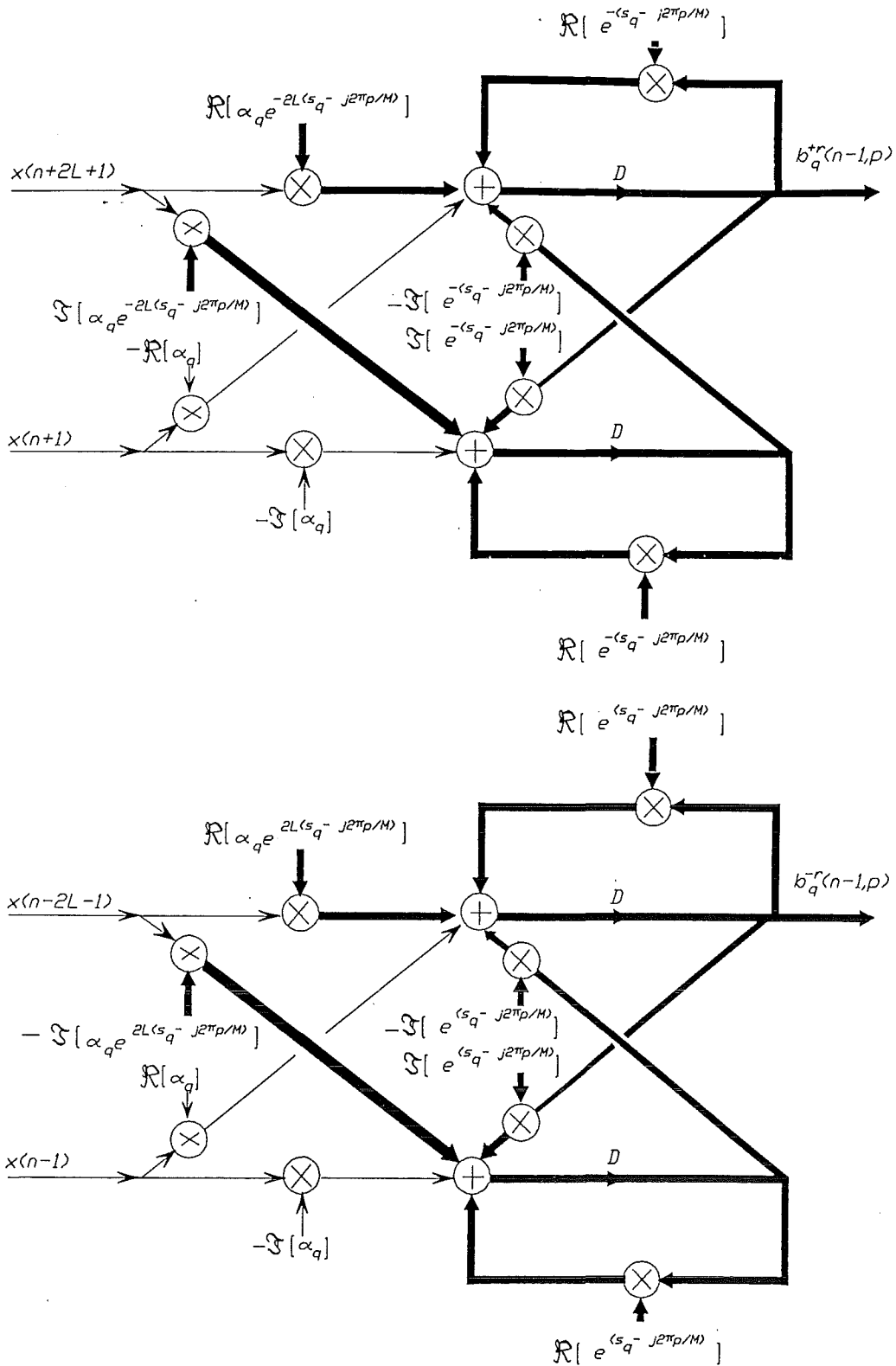


Figure 20: When $\varphi(k)$ and $x(n)$ are real, only $b_q^{\pm r}(n, p)$ contributes to $C(n, p)$. These real components can be generated as shown here.

Robert J. Marks II
 9/15/89
 Robert J. Marks II

Combining conjugately related Szasz components: If two Szasz components with indices q and \hat{q} are related by a complex conjugate as

$$b_{\hat{q}}^{\pm}(n, p) = [b_q^{\pm}(n, p)]^* \quad (55)$$

then, for $\varphi(k)$ and $x(n)$ real, the contribution of the conjugate pair to $C(n, p)$ is simply $2b_q^{\pm}(n, p)$. The implementation follows directly from Fig. 19 and is shown in Fig. 21.

Example- Zamograms with Hanning & Hamming windows: To illustrate computation of zamograms using a Szasz series window, consider again the $Q = 3$ case where α_1 is real and $s_1 = 0$. Let $\alpha_2 = \alpha_3$ and $s_2 = s_3^* = j\pi/L$. The Hanning (Table 2) and Hamming (Table 3) windows are special cases.

Implementation of our running example is shown in Figs. 22, 23 and 24. Figure 22 shows generation of $b_1^{+r}(n-1, p)$ on top and, for the conjugate terms, $2b_2^{+r}(n-1, p)$ on the bottom. The generation of $b_1^{-r}(n-1, p)$ and $2b_2^{-r}(n-1, p)$ is similarly shown in Fig. 23. The terms are gathered as shown in Fig. 24 to produce the zamogram, $C(n, p)$.

Note that in Figs. 22 and 23, the multiplication of $x(n+2L+1)$ and $x(n-2L-1)$, respectively, by the sinusoidal arrays is common to both the $q = 1$ and $q = 2$ stages. As in Fig. 10, the commonality allows a single sinusoidal array multiplication. Such modification of Fig. 22 is shown in Fig. 25. A similar modification is readily applicable to Fig. 23.

5 Notes

Some final remarks follow.

1. The Szasz series window is also potentially applicable to certain other *generalized time-frequency representations* (GTFR's) [6]. Kernels with *Hourglass* and *diamond* shapes [3] in the (m, k) plane can be evaluated by Szasz series windows when, within the shape, the window is $\varphi(k)$. The zamogram has a cone-shaped kernel [3] in the (m, k) plane.
2. In many spectrograms and GTFR's, output spectral lines are not computed at every signal sample point. The Szasz series

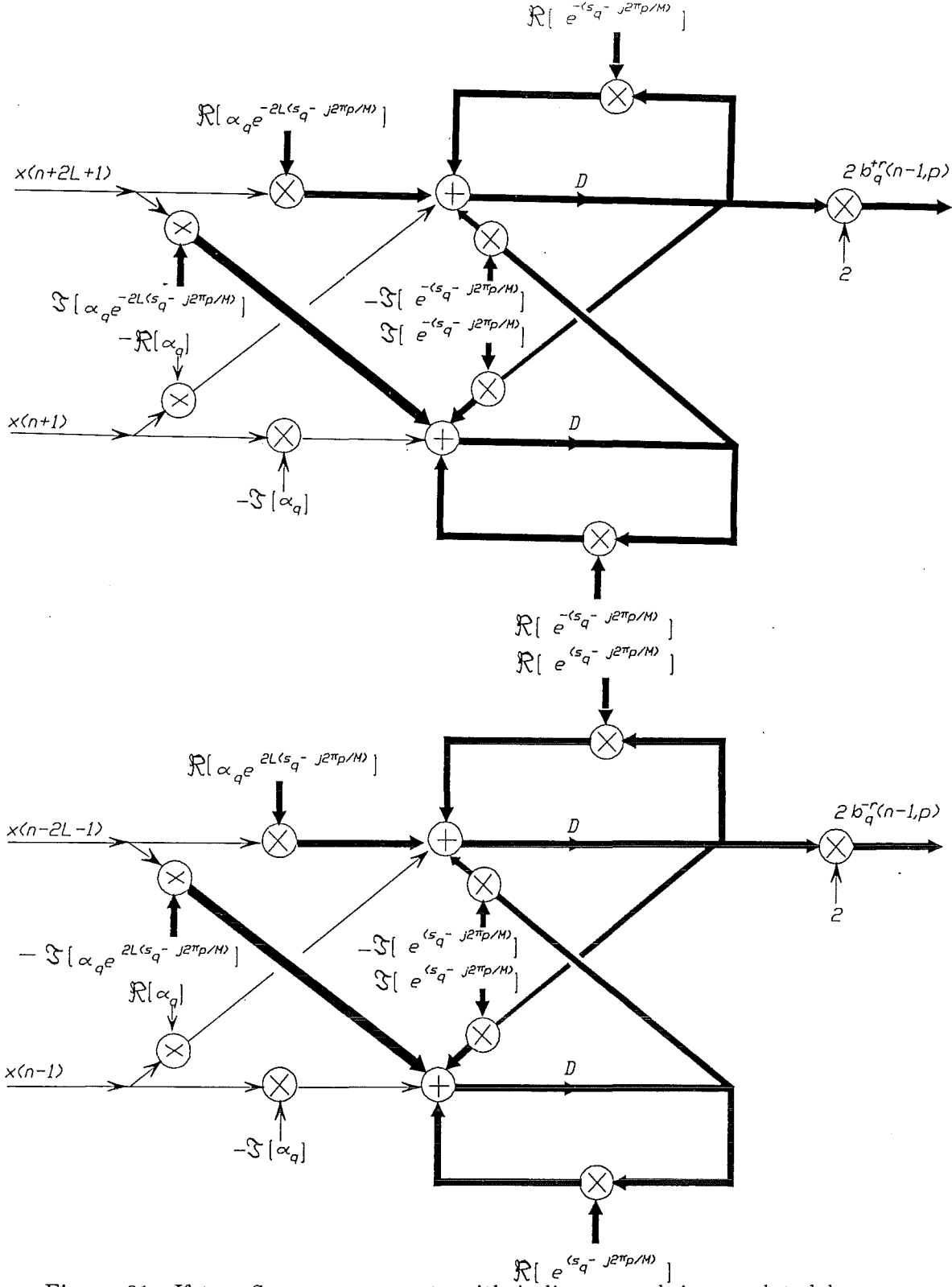


Figure 21: If two Szasz components with indices q and \hat{q} are related by a complex conjugate and $\varphi(k)$ and $x(n)$ are real, then the contributions of both terms to $C(n,p)$ are simply $2b_q^{\pm r}(n,p)$. As shown here, they can be generated as shown here by simply multiplying the outputs in the previous figure by 2.

Robert J. Marks II
 Robert J. Marks II 9/15/89

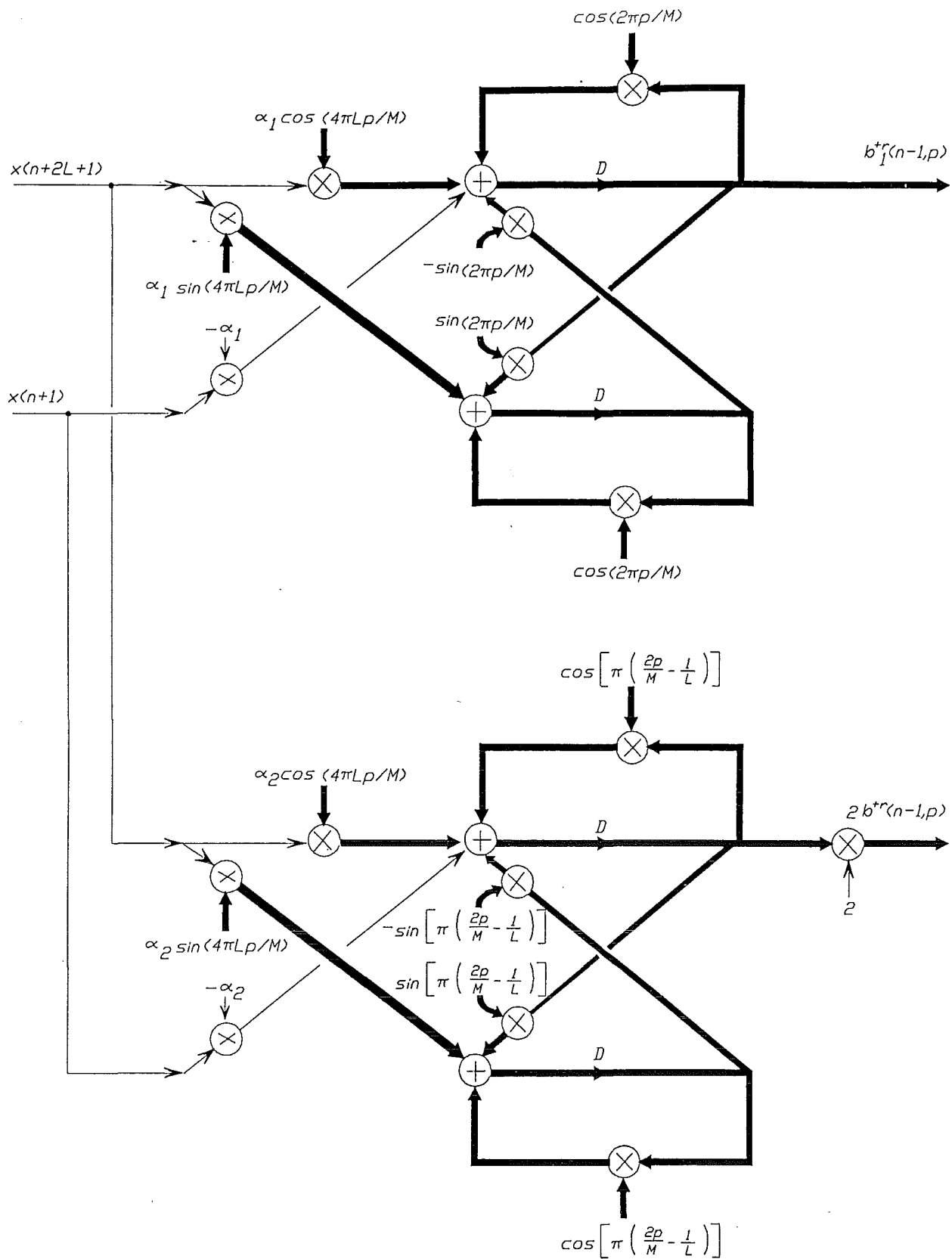


Figure 22: Generation of the $b_2^{+r}(n-1, p)$'s for Hanning and Hamming windows.

Robert J. Marks II
 Robert J. Marks II

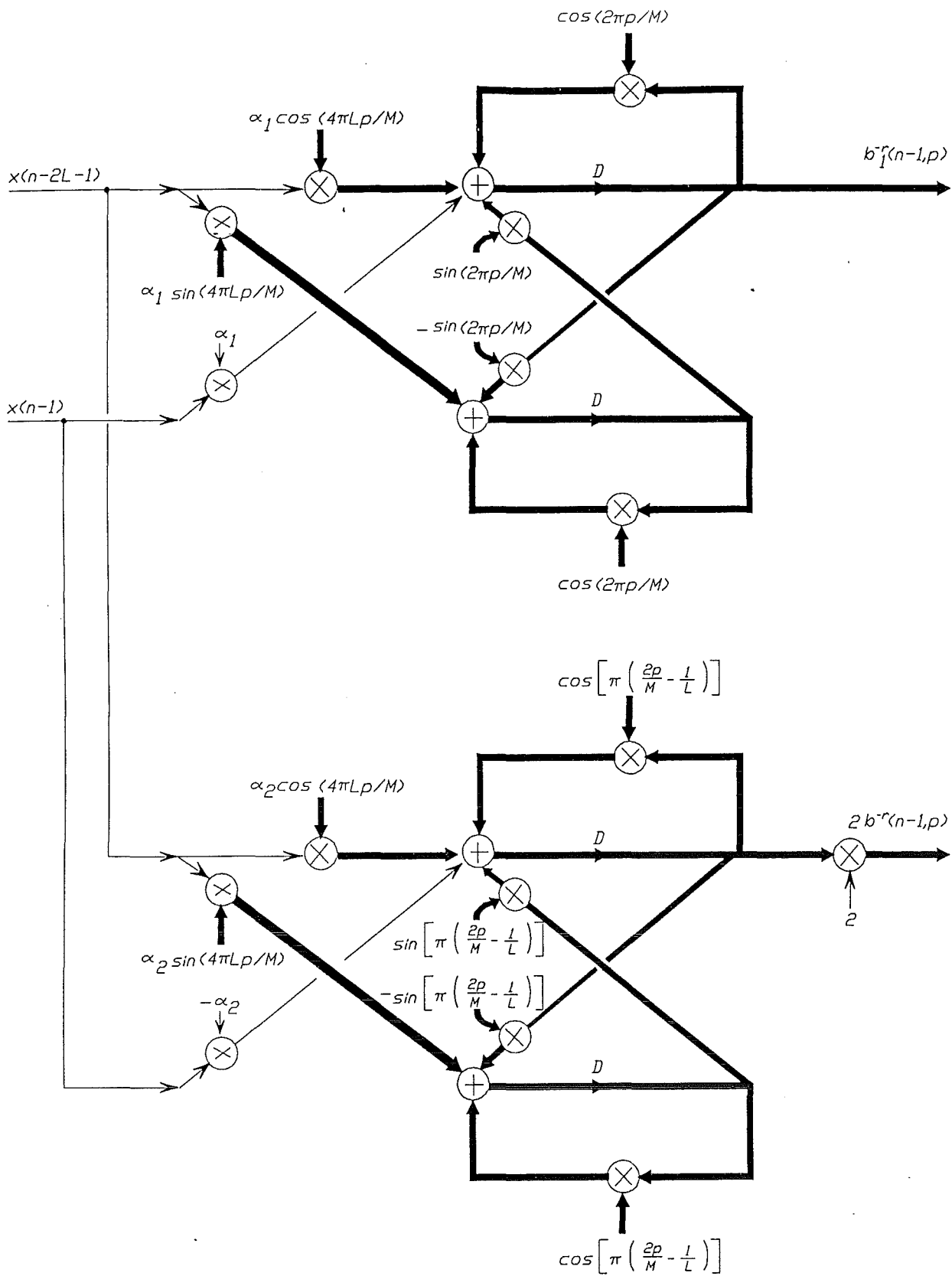


Figure 23: Generation of the $b_2^{-r}(n-1, p)$'s for Hanning and Hamming windows.

Robert J. Marks II
 Robert J. Marks II 9/15/89

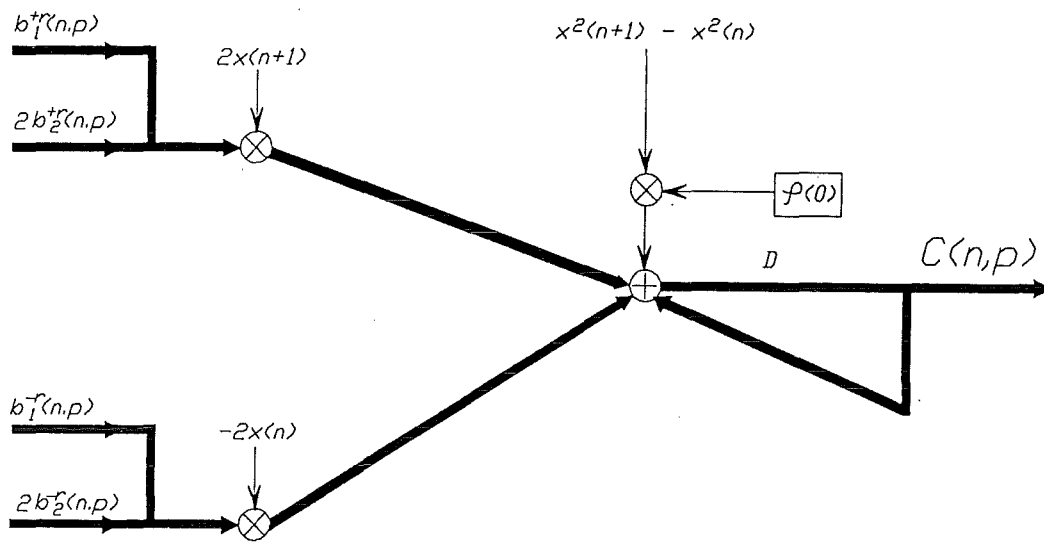


Figure 24: Generation of the zamogram using the inputs generated in the previous two figures.

Robert J. Marks II 9/15/89
 Robert J. Marks II

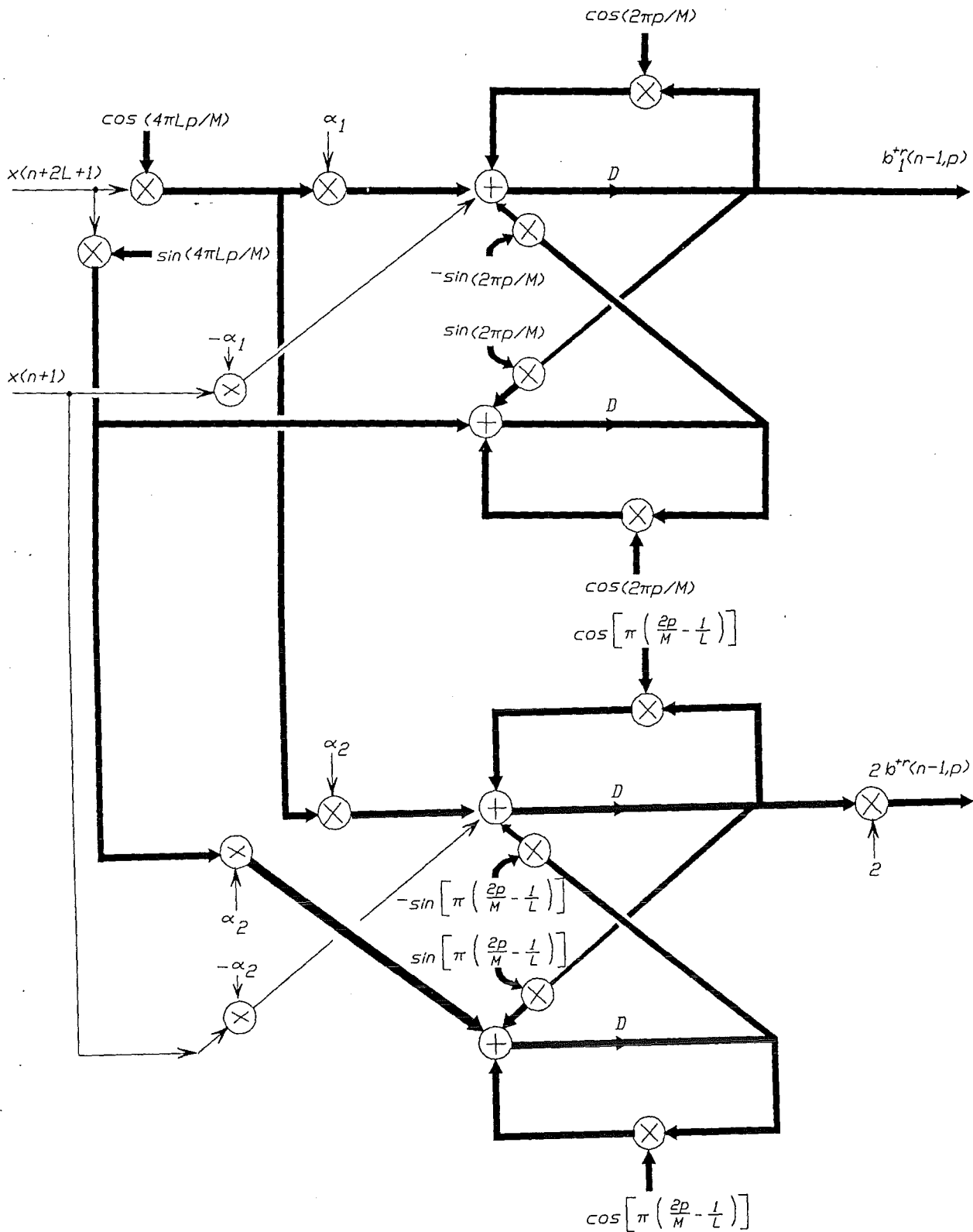


Figure 25: A modification wherein the sinusoidal array common to both components is computed but once.

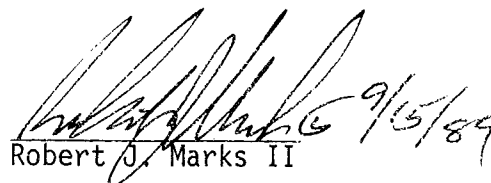
Robert J. Marks II
 Robert J. Marks II

window approach can be adapted to such cases in one of two ways. First, and most obvious, the iteration can proceed at each point with outputs generated periodically. Secondly, the iteration can be modified to the longer period. For example, in the weighted running average example, if there is to be an output at every other input sample point, then, at each iteration, two new samples would be introduced (instead of one) and two old samples would be deleted (instead of one). Each Szasz factor would be squared.

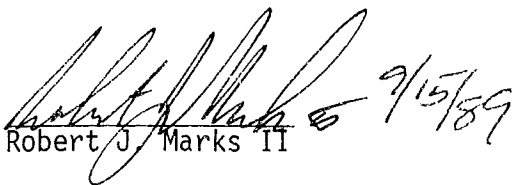
3. For the spectrogram (and the spectrogram component of the zamogram), computation of the output spectral line can be viewed as a number of multiplexed IIR filters parameterized by p . The only time one filter "talks" with another is in the operation of transposition.
4. There exist a number of modifications to the implementation of the Szasz signal processing algorithms that correspond directly to the commutative, distributive and associative laws applied to multiplication and addition. Performing a single sinusoidal array operation in Fig. 25 (compare with Fig. 22) is an example of a variation due to the distributive law.

References

- [1] L.E. Atlas, Y. Zhao and R.J. Marks II "Application of the generalized time-frequency representation to speech signal analysis", *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pp.517-519, Victoria, B.C. Canada, June 4-5, 1987.
- [2] L.E. Atlas, Y. Zhao and R.J. Marks II "The use of cone-shape kernels for generalized time-frequency representations of nonstationary signals", *IEEE Transactions on Acoustics, Speech and Signal Processing*, (in press)
- [3] W.C. Kooiman, "Time-frequency speech displays that are an improvement over the spectrogram", *M.S. Thesis*, Department of Electrical Engineering, University of Washington (1989).


Robert J. Marks II 9/15/89

- [4] R.E.A.C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, Providence, R.I.: Amer. Math. Soc., 1943.
- [5] E. Masry, "An extension of Szasz's theorem and its applications", *IEEE Transactions on Information Theory*, vol.IT-19, pp.184-187 (1973).
- [6] L. Cohen, "Time-frequency distributions - a review", *Proceedings of the IEEE*, vol.77, pp.941-981 (1989).


Robert J. Marks II 9/15/89

CHRISTENSEN
O'CONNOR
JOHNSON
KINDNESS

LAW OFFICES

PATENT, TRADEMARK AND OTHER
INTELLECTUAL PROPERTY MATTERS

2700 WESTIN BUILDING
2001 SIXTH AVENUE
SEATTLE, WASHINGTON 98121

TELEPHONE: (206) 441-8780

TELECOPIER: (206) 441-0516

TELEX: 4938023

CABLE: PATENTABLE

July 26, 1989

Mr. Peter Odabashian
Director, External Affairs
Washington Technology Center
University of Washington
376 Loew Hall, M.S. FH-10
Seattle, WA 98195

Re: Title: OPTICAL NEURAL NET MEMORY
U.S. Patent No. 4,849,940
Issued: July 18, 1989
Patentee: R.J. Marks, II et al.
Your Reference: WTC #87-6
Our Reference: WTCC-1-3835

Dear Peter:

We are pleased to inform you that the subject patent issued on July 18, 1989. Typically, the official Letters Patent comes to us from the United States Patent and Trademark Office several weeks after the stated issue date. We will correspond with you at that time.

While it is not mandatory to use the patent number in marketing embodiments of the invention, the failure to do so may result in an inability to collect damages in the event the patent is infringed. The statute (Title 35, United States Code, Section 287) provides as follows:

Patentees, and persons making or selling any patented article for or under them, may give notice to the public that the same is patented, either by fixing thereon the word "patent" or the abbreviation "pat.", together with the number of the patent, or when, from the character of the article, this cannot be done, by fixing to it or to the package, wherein one or more of them is contained, a label containing a like notice. In the event of such failure so to mark, no damages shall be recovered by the patentee in any action for infringement, except on proof that the infringer was notified of the infringement and continued to infringe thereafter, in which event damages may be recovered only for infringement occurring after such notice.

*Gents —
pls. heed
PRO*

Mr. Peter Odabashian
July 26, 1989
Page 2

If you are in doubt as to the proper use of the patent number on a product or in connection with a process, please let us know.

Very truly yours,

CHRISTENSEN, O'CONNOR,
JOHNSON & KINDNESS



By
Michael G. Toner

MGT/jlm/lkb



PATENT PENDING

The laser was one of the great inventions of this century. The question was, who owned it?

Even before he entered high school, Gordon Gould knew he wanted to be an inventor. His heroes were Marconi, Bell, and Edison. He knew, too, that to invent anything truly significant he'd have to understand the physics of things, how things worked deep down in the invisible quanta. In high school, college, and graduate school he gathered the tools. He wanted to be ready when

the light bulb flickered. On November 9, 1957, a Saturday night just given to Sunday, Gould was unable to sleep. He was 37 years old and a graduate student at Columbia University. The idea came to him, he remembers, about one o'clock. No mere Soft White, this bulb. For the rest of the night and the rest of the weekend, without sleep, Gould wrote down descriptions of his idea, sketched its components, projected its future uses.

On Wednesday morning he hustled two blocks to the neighborhood candy store and had the proprietor, a notary, witness and date his notebook. The pages described a way of amplifying light and of using the resulting beam to cut and heat substances and measure distance. "That notebook is absolutely incredible," says Peter Franken, a professor of physics and optical sciences at the University of Arizona, in Tucson. "It's as if God came down and whispered in Gordon's ear and said, 'Listen, buddy, this is what you're going to do.'"

Gould dubbed the process light amplification by stimulated emission of radiation, or laser, and he knew—he knew, no question—that this was the invention he'd been preparing himself for all along. The invention of a lifetime.

It was indeed, in a way Gould did not anticipate. For it took nearly half a lifetime—the next 30 years—to win the pat-

ents for his ideas. At times the government's resistance to Gould's claims was so stubborn, its behavior so unusual, that he and his allies began to fear a concerted government-industry effort to keep Gould from ever getting a patent.

Gould's vindication came only last year, when he won the last of a series of victories that left him in control of patent rights to perhaps 90% of the lasers used and sold in the United States, lasers that weld auto parts, destroy skin cancers, aim weapons, and register prices at the checkout counter. Gould's patents directly affect some half-

billion dollars in annual sales of lasers; ironically, had they been granted 30 years ago these patents would have expired while the industry was still tiny, and would have captured only a fraction of their current revenue. The company formed to license the Gould

patents, Patlex Corp., now sits atop a rapidly growing mountain of cash,

▶ STORY PROPOSAL ◀

Gordon Gould, inventor of the laser, spent 30 years and \$6 million staking his claim to one of the most spectacular advances of modern science. What stood in his way? Just the twin phalanxes of U.S. government and big business. Many people bet their careers, and their companies, to back the inventor, and the rewards are finally streaming in. Gould's triumph closes the book on a fascinating legal and scientific endeavor. —E.L.

GREG PEASE

BY ERIK LARSON

and last summer it hired Frank Borman, moon pilot and former chief executive of Eastern Air Lines Inc., to be its new boss.

For Gould especially, victory is very, very sweet. Every other day a Federal Express truck arrives at his home in Virginia bearing license contracts to sign. Every quarter a check comes. A grin breaks across Gould's face, a Cheshire cat's grin flecked with canary feathers, as he matter-of-factly estimates that total royalties will be \$46 million. "That's my share of it."

But Gould is 68 years old. He and his partners, men who gambled their futures to back him, spent more than \$6 million fighting both the United States Patent and Trademark Office and the laser industry. The story is not one of courage and perseverance only on Gould's part. Gary Erlbaum liquidated his company and bet the proceeds on Gould. Richard Samuel, a patent attorney, gave up his law partnership to become Gould's master strategist. Gould fought history—and won.

GORDON GOULD, FOR NOW, LIVES IN A small, gray ranch house situated by a creek in Virginia's Northern Neck, two and a half hours from Washington, D.C. The place is modest because that's the way Gould likes to live, not because he can't afford better. He's already a millionaire. At the rear of the house is a huge all-weather porch, and Gould is sitting there in the smoke of an endless chain of cigarettes.

He is a lean, angular man, with heavy-framed glasses and a scalp that has yielded some to the advance of time. There is a war-torn aspect to the room symbolic of the battles so recently won. Smoke. Ragged butts jamming two ashtrays. Gammon, a German shepherd with one blood-fused eye and severe hip dysplasia, moves sideways across

"What'd you say?" Appel asks, squinting through wayward smoke. "That was the only moment? Or the first moment?"

"Well, OK. It was the first moment."

Gould was born on July 17, 1920, in New York City. He was the kid who fixed clocks for neighbors. At Union College, in Schenectady, N.Y., he studied physics and fell in love with light. He went to Yale in 1941 to begin work toward his doctorate, but war forced him to quit. Over the next two years, he worked on the Manhattan Project, the ultimate in applied physics. In 1945, indulging his girlfriend, he began attending meetings of a Marxist

study group in Greenwich Village. The government yanked his security clearance. He took a job at a company that made specialized mirrors and spent the rest of his time trying to develop inventions.

In 1951 Gould resumed his doctoral work at Columbia. He taught part-time at City College of New York until 1954—Senator Joseph R. McCarthy's heyday—when he was called before a special panel of the state board of higher education commissioned to root commies from the halls of academe. Gould spent a day under interrogation but refused to testify against colleagues and friends. He was fired. His faculty adviser at Columbia, incensed by this treatment, got Gould a research assistantship at the university's radiation lab.

Meanwhile, a Columbia physicist, Charles H. Townes, had devised a method of amplifying microwave energy, an advance he dubbed the maser. To do the same with light required a radically different approach, and it was this process Gould conceived that night in 1957. "I almost immediately saw the tremendous potential of this device," Gould says. "It would do for light what the vacuum tube and later the transistor did for radio frequency electronics." He envisioned lasers used to heat, weld, and cut; to machine parts; to measure distance; even to produce the heat necessary for nuclear fusion, technology only today being seriously investigated.

It was then that Gould made the mistake of a lifetime—a mistake that in the grandest of paradoxes promises to make him an extremely wealthy man.

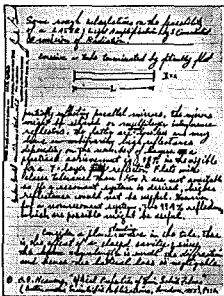
IN COURTROOMS AROUND THE COUNTRY, there are mounds of Gould paper. In the National Archives, a full cart of boxes accounts for a single lawsuit. At one point the patent office set aside a separate room for the Gould patents and took reservations from companies wanting a look.

The patent office lives paper, breathes paper—most of it precise, legal, notarized, certified, a massive white drift of painfully accurate prose. Even with the help of a patent lawyer, few applications succeed on the first round. Two out of three, however, eventually will become patents. In fiscal 1988 this rite of passage—the pendency period—was 19.9 months. To understand why Gould needed 30 years, it's necessary first to know the ritual.



Patent lawyer Richard I. Samuel
Inventive financing: lawsuit went public

MARTHA WOODWARD



The government's resistance to Gould's claims was so stubborn, he says, 'there was a point where I became convinced there was a conspiracy.'

the room, a dog in serious misalignment. Gould lives with his longtime companion, Marilyn Appel. Of dragonish temperament, she is tough, energetic, and blunt, a screener of calls, guardian of the gate. Now and then she charges onto the porch, lights a cigarette, catapults herself into the conversation. Gould sits at rest, a portrait of physical entropy.

What kept him going all these years was sheer, blissful ignorance. "What you have to realize," he says, "is that at no point did I expect it was going to take more than a couple of years to resolve whatever problem existed at a given moment." Only once, he says, did he fear he would never get a patent.

For the inventor, this bureaucracy becomes distilled in a single individual: the patent examiner, the high priest of invention. There are 1,400 examiners, each possessing a startling degree of control over the fate of an idea. On the average, an examiner will spend 17 hours on each case. How does 17 hours become 19.9 months? The initial processing takes a month. An examiner won't get the case for another two to three months. The inventor has three months to respond to each formal action the examiner takes. In a typical case there are two such actions. Throw in another three months for printing and publishing the final patent, and you've spent more than a year.

A challenge to a patent dramatically extends this pendency period. When two applications conflict, the patent office can begin what is called an interference proceeding to determine who was the first inventor. These proceedings can last decades. In another type of proceeding, called reexamination, the challenger can trigger rejection of the patent by producing new evidence of prior art—new evidence that the invention was “not novel or was obvious.”

Gould took the proper first step and consulted a patent attorney for advice on how to proceed. “I was so ignorant of the whole patent procedure that I came away from that meeting with the wrong impression, which was that I had to build a model in order to get a patent,” Gould recalls. This is where he made his big mistake: patent law requires no such thing. Gould needed only to present enough detail to allow someone skilled in the art to build the device.

Gould was so excited about his laser ideas that he left Columbia without finishing his thesis. He joined Technical Research Group Inc. (TRG), a small scientific company on Long Island, hoping to develop laser applications. In 1959 he won TRG a \$1-million contract for laser research, and filed for his patents. But he had lost precious time. Charles Townes and Arthur Schawlow, a Bell Labs physicist, had applied the previous July to patent the optical maser.

Soon after receiving notification that the contract had been awarded, Gould also learned of the government's intent to classify his research as secret. This would not have caused Gould much grief had he been considered your basic loyal American. Officially, however, Gould remained suspect. He was denied clearance to work on the project; his notebooks were confiscated. (Again that Cheshire grin—Gould admits he kept copies of them.)

TRG's president, Lawrence Goldmuntz, spent \$50,000 fighting to win Gould clearance, an effort that culminated in a 1959 hearing during which the ghosts of Gould's past marched before him. An FBI agent testified that he had tapped Gould's phone. The man who had led the Marxist study group revealed he had been an FBI informant. Gould had married and divorced the woman with whom he'd attended the meetings; she too appeared and testified against him. Gould did not get his clearance.

Gould's security troubles put TRG in a bind. In seeking the contract, he'd deliberately kept his proposal vague. Now, that proposal would be used by scientists who did not have his knowledge; they could ask him questions but could not tell him a thing. Largely as a result, Gould contends, TRG failed to build the country's first laser. It was a failure that would haunt Gould for decades to come.

In March 1960 Schawlow and Townes received their optical maser patent. For the next 17 years this and Townes's earlier patent would be considered *the* laser patents. Two months later Theodore Maiman, a scientist at Hughes Research Laboratories, built the first working laser.

Gould's application met its first serious resistance in the early 1960s, when it became mired in the first of five interference proceedings. Although Gould won some claims, he also lost important ground. The patent office ruled that he

Gary Erlbaum liquidated his company and bet the proceeds on Gould. Richard Samuel gave up his law partnership to become Gould's master strategist. Gould fought history—and won.

had not disclosed enough detail to enable anyone to build his laser.

Meanwhile, the costs of these battles had grown too much for TRG. The company's parent, Control Data Corp., sold Gould back his rights. TRG's patent attorney backed Gould, on spec, for another five years, but Gould needed a partner with clout. He thought he had found his knight, REFAC Technology Development Corp., in New York City. In 1975 REFAC agreed to act as Gould's licensing agent in return for 50% of any future royalties. “I believed in Gordon Gould,” says Eugene Lang, REFAC's founder, perhaps best known for guaranteeing the college educations of an entire sixth-grade class in Harlem. “My own associates thought I was nuts.”

Gould signed with REFAC expecting that the company would help him win his patents. But REFAC contended it had agreed only to license Gould's inventions. What Gould needed were big legal guns and the big bucks to pay them.

Gould turned 55 that year. He still did not possess a single significant U.S. laser patent.

RICHARD I. SAMUEL KNOWS FROM KOOKS, the people who troop through a patent attorney's door claiming such minor inventions as the wheel. Samuel is a somber man with a dark, gray-misted beard. When he met Gould, he was a partner in a patent-law firm in Westfield, N.J. “We get a lot of nutcakes coming in,” says Samuel. “You need to figure out whether they are dealing with reality.”

Gordon Gould, referred to the firm by REFAC, quietly explained to Samuel that he had invented the laser. Gould presented his application, all 113 pages and 19 drawings. He presented other official papers, including a document dated only six weeks earlier. This was especially striking. The ritual of the patent process demands that an inventor keep the chain of action and response going. Once this chain is broken, the patent office considers the application abandoned. “With Gould,” Samuel says, “the more I delved, the more I believed he was right.”

Samuel decided Gould's claims indeed had merit, and the firm agreed to pursue them for up to





The Washington Technology Center

376 Loew Hall, FH-10, University of Washington, Seattle, WA 98195

Office of the Executive Director
(206) 545-1920

TO: WTC/UW Principal Investigators

FROM: Edwin B. Stear *EBS*
Executive Director

SUBJECT: Technology Disclosures

This memorandum, along with the enclosed materials, is intended to provide specific guidance on the handling of technology disclosures through the WTC, as well as clarify The Washington Technology Center's Patent and Copyright Policy in general.

As you know, President Gerberding in October 1985 signed Administrative Order No. 17 which exempted the WTC from UW patent and copyright policies and delegated authority to the WTC to have and administer its own Patent and Copyright Policy subject to certain conditions (see the enclosed copy). Subsequently, the WTC Board of Directors approved a WTC Patent and Copyright Policy. Although a copy of this policy was distributed to you some months ago, it is included here to provide a self-contained information packet.

To provide further background, I am enclosing copies of the WTC Principles governing patent and copyright policies and procedures, and the Agreement between The Washington Technology Center and the Washington Research Foundation (WRF).

Finally, in accordance with the documents identified above, the enclosed technology disclosure policy is provided for your information and use in disclosing inventions related to WTC research projects. As noted in the instructions, the disclosure will generally be forwarded to the Washington Research Foundation (or other agent), at the discretion of the WTC, for evaluation of patents and commercial potential.

Please feel free to contact me if you have any questions concerning these policies or procedures.

EBS/bf

Enclosures

INVENTION DISCLOSURE

Washington Technology Center

Instructions

This Invention Disclosure Form is used to report inventions and to record the circumstances under which the invention was made. The Disclosure is a legally important document; care should be taken in its preparation since it provides both the basis for determining patentability and the data for drafting a patent application.

New and potentially useful technology developed by WTC employees with WTC and/or industry grant and contract support should be reported promptly consistent with the Center's Patent and Invention Policy.

The following instructions apply to the correspondingly numbered sections of the form.

1. Use a brief title, sufficiently descriptive to aid in identifying the invention.
2. Provide a brief description, pointing out novel features of the invention. Attach additional material which covers the following points:
 - a. General purpose
 - b. Technical description with references to drawings, schematics, sketches, flow diagrams, etc., as appropriate
 - c. Advantages and improvements over existing methods, devices or materials, and features believed to be new
 - d. Possible variations and modifications
 - e. State-of-the-art prior to invention, and similar or related patents (if known)
3. List all sources of support for the research which led to the conception or actual reduction to practice of the invention. Include WTC personnel, funds or materials as well as those of University or outside agencies, organizations and companies.
4. The invention history is legally important in determining the priority of invention and/or legal "bars" to patenting. The United States Patent law allows submission of a patent application up to one year after an enabling disclosure of the technology. Most foreign countries require a patent application prior to any enabling disclosure (an oral presentation or publication such as an article, abstract or theses, or other communication which would allow a knowledgeable person to duplicate the work).

5. List all reports, abstracts, papers, theses or patent applications which have been or are planned to be submitted by the inventor(s) describing the invention. Give dates of submission and actual or anticipated publication dates. Attach documents, if available. These documents may be used in part to respond to Section 2.
6. List any other known references, patents, patent applications or other publications pertinent to this invention. Attach copies, if available. These documents may also be used in part to respond to Section 2.
7. Describe and date any sale or public use of the invention in the United States. Specify if the use was operational, or for testing purposes, and if there was any effort or intent to maintain invention secrecy after operational use began.
8. List all co-inventors (any individuals who conceived an essential feature of the invention, either independently or jointly with others, during the evolution of the invention). In the event a patent application is filed, inventorship will be verified by the patent attorney.
9. Arrange for two technically qualified witnesses to read and sign this document verifying that they have understood the invention that is disclosed.

Submit the completed Disclosure together with the Transmittal form to Dr. Edwin B. Stear, Executive Director, Washington Technology Center, University of Washington, Mail Stop FH-10, Seattle, Washington 98195. Generally it will then be forwarded to the Washington Research Foundation (or another agent) for evaluation of patentability and commercial potential.

For further information, contact The Washington Technology Center, (206) 545-1920.

WASHINGTON TECHNOLOGY CENTER

INVENTION DISCLOSURE

This invention Disclosure is an important legal document and should be completed carefully. Please refer to the attached instructions.

1. Title of Invention

2. Brief Description

3. Funding Source(s)

4. Invention History	Date	Location and Comments
A. Initial Idea		
B. First description of complete invention, oral or written		
C. Invention development records, notes, drawings (evidence of diligence)		
D. First successful demonstration, if any (first actual reduction to practice)		
E. First publication with full description of invention (may bar patent)		
F. First verbal description to others		

5. List all reports, abstracts, papers, theses or patent applications related to the inventions which have been published or are planned to be submitted by the Inventor(s). Include copies if available.

6. List any other references, patents, patent applications or other publications which may be pertinent to the invention. Include copies if available.

7. Describe and date any sale or public use of the invention in the United States.

8. Inventor or Co-inventors

Signature Date

Signature Date

Name (Print) Title

Name (Print) Title

Address

Address

Telephone

Telephone

Signature Date

Signature Date

Name (Print) Title

Name (Print) Title

Address

Address

Telephone

Telephone

9. Invention disclosed to and understood by (two witnesses required):

Signature Date

Signature Date

Name (Print)

Name (Print)

Submit completed Disclosure to the Washington Technology Center,
University of Washington, 376 Loew Hall, M/S FE-10, Seattle, WA
98195.

Date Received: _____
Washington Technology Center

THE WASHINGTON TECHNOLOGY CENTER

Form to Transmit Invention Disclosure
(For WTC Internal Use Only)

Instructions

Complete this form and the attached Invention Disclosure form and forward to The Washington Technology Center via WTC Program Director, Department Chairperson, and Dean of School/College for approval. If more than one Department is involved, obtain signatures from all Chairpersons and Deans (or their designate).

To: Washington Technology Center Date: _____
Loew Hall 376, FH-10

From: _____
Inventor Name Title Department Mail Stop

Inventor Name Title Department Mail Stop

Inventor Name Title Department Mail Stop

Inventor Name Title Department Mail Stop

Re: Invention entitled: _____

Verified and Approved: _____

Concurrence: _____

WTC Program Director _____

Dean of the School/College _____

Date: _____

Date: _____

Concurrence: _____

Accepted: _____

Department Chairperson _____

Edwin B. Stear, Executive Dir.
WASHINGTON TECHNOLOGY CENTER

Date: _____

Date: _____

ADMINISTRATIVE ORDER NO. 17

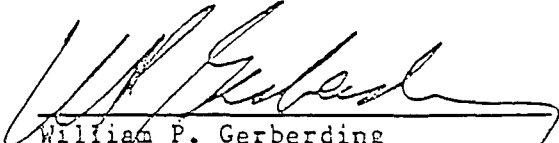
Effective October 18, 1985

SUBJECT: Exemption of the Washington Technology Center from the University of Washington Patent and Copyright Policies and delegation of authority to the WTC to have and administer its own Patent and Copyright Policy subject to certain conditions.

AUTHORITY: University Handbook, Volume II, Part I, Chapter 12, Sections 12-11 and 12-12.

- A. The Washington State Legislature, in Chapter 72, Section 11, Laws of the 1983 1st Extraordinary Session, with the concurrence of the Governor, has established The Washington Technology Center (WTC) at the University of Washington (UW) to be administered by the Board of Regents of the UW. Accordingly, unless otherwise specified, the WTC is subject to UW policies. However, the WTC Board of Directors and the UW Administration, acting under delegated authority from the UW Board of Regents, have agreed that in light of the purposes, goals, objectives and intended nature of the WTC, it should not be fully subject to UW Patent and Copyright Policies but should adopt its own Patent and Copyright Policy.
- B. The WTC is exempted from UW Patent and Copyright Policies subject to certain conditions as follows:
1. the WTC may identify itself as the owner of inventions, patents and copyrights derived from WTC projects;
 2. those inventions, patents and copyrights will be administered under a WTC Patent and Copyright Policy approved by the WTC Board and the UW Administration; and
 3. the WTC will enter into a Technology Administration Agreement (TAA) with the Washington Research Foundation (WRF) that is identical in all substantive respects with the TAA between UW and WRF attached hereto as Exhibit A.

This Administrative Order No. 17 is pursuant to the authority cited above.


William P. Gerberding
President

THE WASHINGTON TECHNOLOGY CENTER

PRINCIPLES GOVERNING
PATENT AND COPYRIGHT POLICIES AND PROCEDURES

1. The Washington Technology Center, hereafter referred to as the WTC, shall own all patents and copyrights arising from WTC sponsored research and technology development programs and projects.
2. The WTC shall negotiate all patent and copyright agreements and licensing arrangements so as to maximize technology transfer for the benefit of the economic development of the State of Washington.
3. Negotiations of patent and copyright agreements and subsequent licensing arrangements shall be the responsibility of the duly appointed individual in charge of the WTC Office at the appropriate participating university in accordance with WTC policies and procedures.
4. The WTC shall develop a Patent and Copyright Policy which will form the basis for negotiation of specific agreements on patents, copyrights, licensing, and distribution of royalty income with each of the participating universities.
5. The WTC shall negotiate up-front patent and copyright agreements, including licensing provisions, with all participating industrial sponsors of WTC programs and projects.
6. The WTC shall negotiate individual up-front patent and copyright agreements with all Industrial Fellows and their employers.
7. All individuals participating in WTC programs and/or projects shall sign an agreement requiring them to be bound by the WTC's Patent and Copyright Policy.
8. When investigators from more than one university work on a WTC project, there shall be a specific up-front agreement among all parties covering patent and copyright issues, including negotiation of agreements with industrial supporters of the project, negotiation of licenses for any intellectual property developed, distribution of royalty income, and ownership of any patents or copyrights in the event the WTC is terminated or ceases to operate for any reason.

THE WASHINGTON TECHNOLOGY CENTER

Patent and Copyright Policy

1. One of the primary missions of The Washington Technology Center (hereinafter referred to as WTC) is to develop new commercializable technology through joint industry-university research and technology development programs. Patents and copyrights are important in this process to:

- (a) protect the economic interests of the WTC and the inventors.
- (b) protect the economic interests of the industrial participants and the licensees.
- (c) provide a firm legal basis for transferring the technology.

It is recognized that the value of the technology may diminish rapidly with time. Therefore, it will often be necessary to transfer technology immediately after disclosure and prior to application for or issuance of patents and copyrights.

Further, it is recognized that it will also be necessary to transfer technology without applying for patents or copyrights in those cases where the technology is not patentable or copyrightable, or where the value of the particular patent or copyright does not justify the expense.

The purpose of this document is to set forth the specific policies adopted by the WTC to assure that these requirements and goals are met.

2. As a condition of participation in WTC research projects, all personnel participating in WTC projects agree to assign their title and rights to all inventions and copyrightable material arising in connection with such research projects to the WTC, to an agent designated by the WTC, or to a sponsor, if required under agreements governing sponsored research. Such personnel shall execute documents of assignment and do everything reasonably required to assist the assignee(s) in obtaining, protecting, and maintaining patents, copyrights or other proprietary rights.

The WTC has no vested interest in inventions or copyrightable material conceived and developed by participants entirely on their own time and without the use of WTC facilities. However, in order to clarify the inventor's or creator's

title to such inventions and/or copyrightable material and to insure compliance with the requirements of any sponsors, all inventions and/or copyrightable material generated during participation in WTC programs and projects shall be reported to the WTC for determination of the degree of WTC interest.

If the WTC, in consultation with the appropriate participating universities, determines that it has no interest in an invention or copyrightable material or decides to forego the patenting, copyrighting, or other commercialization of an invention or copyrightable material, it shall waive its rights to the invention or copyrightable material in writing. Upon receipt of such a waiver, and assuming that no additional WTC or University resources will be invested, the inventor(s) or creator(s) may file a patent or copyright application and/or grant a license of his/her own.

3. WTC research funded wholly or in part by an outside sponsor is subject to this policy as modified by the provisions of negotiated agreement(s) covering such work. It is the general policy of the WTC to negotiate all such agreements, including any special provisions relating to the intellectual property, prior to initiation of the research effort being sponsored. Participants in such sponsored research are bound by the provisions of these agreements.

4. In general, title to any inventions and/or copyrightable material conceived and first reduced to practice in the course of research carried out in the WTC with the support of Federal agencies, industry, or other sponsors shall vest in the WTC. In rare cases, an industrial sponsor may possess a dominant patent or copyright position in a certain technology area so that any patent or copyright the WTC might seek would be of little value. For this or other such reasons, an exception to this WTC title policy may be approved when to do so would honor the general principles of this policy, protect the equities involved, and satisfy the requirements of the parties. In all cases, the granting of such exceptions must be explicitly covered in the agreements referred to above in Paragraph 3.

5. Interaction between the WTC and industry can take any one or more of the following forms: grants, contracts, consortial arrangements, equipment gifts, and appointment of industrial fellows. Industrial firms sponsoring WTC research programs through any one or more of these forms may be assured of at least a non-exclusive license to inventions and copyrights conceived and developed with their support. If necessary for the effective development and marketing of a WTC invention or copyright, an exclusive license may be granted for a limited period of time if the sponsor agrees to finance the cost of the WTC's patent or copyright application and observes due

diligence in bringing the technology involved into public use. In such cases, the patent or copyright costs may be treated as an offset against royalties payable when the invention or copyright is marketed.

Where the sponsor uses the invention or copyright entirely within its own operations, the license may be royalty-free. Where the sponsor, or a third party licensee, manufactures and sells products, services, or processes based on the invention or copyright, reasonable royalty payments to the WTC or its assignee are normally required.

In all cases involving industrial sponsorship of WTC research programs, the specific licensing rights of the sponsor(s) to any patentable and/or copyrightable technology generated in the research programs shall be explicitly covered in the up-front agreements referred to above in Paragraph 3.

6. Although the WTC reserves the right to patent and/or copyright intellectual property itself, it has designated the Washington Research Foundation as its primary patenting, copyrighting, and licensing agent. However, another comparable, mutually-acceptable patenting, copyrighting and licensing agent can be used if so desired by an individual participating university.

7. Both the inventors and/or creators and the WTC are entitled to a share of royalty income from licensed patents and/or copyrights; the WTC on the basis of salary and/or facilities support for the inventor and/or creator and the cost of patent, copyright, and licensing administration; and the inventor and/or creator on the basis of the creative activity, documenting the invention or copyright, and assisting as necessary with commercialization. To recognize creativity and to encourage prompt disclosure of potential patents and copyrights, the WTC allocates the greater share of net early royalty income to the inventor or creator. The remainder is dedicated to further research by allocating shares to the WTC and to the home colleges/departments of the inventors and/or creators as appropriate. Unless amended in an agreement with a participating university, the specific allocation shall be as follows.

After deducting 15% for administrative services, net royalty income received from WTC inventions and/or copyrights handled by an outside agency is distributed as follows:

Cumulative Net Income	Inventor/ Creator	Inventor's University Dept./College	WTC Research Fund
First \$10,000	100%	0%	0%
\$10,000-\$40,000	50%	25%	25%
Above \$40,000	30%	20%	50%

In the event that an invention and/or copyright is administered directly by the WTC, the direct costs of obtaining and maintaining the patent(s) and/or copyright(s) must be recovered in addition to the 15% service fee before distribution of royalty income begins under the above formula.

The royalty derived WTC Research Fund shall be used to promote additional research in areas identified for emphasis by the WTC.

When a proposed WTC program or project involves more than one university, it is the general policy of the WTC to negotiate an up-front agreement with the participating universities covering patent and copyright issues. Including negotiation of agreements with industrial supporters of the project, negotiation of licenses for any intellectual property developed and distribution of royalty income and ownership of any patents and copyrights in the event the WTC is terminated or ceases to operate for any reason.

8. As a public institution, the WTC should undertake sponsored research under conditions which permit timely publication of the research results. However, the WTC reserves the right to defer publication for a reasonable period of time during which the WTC and any sponsor(s) review the feasibility and desirability of patent and/or copyright protection of any intellectual property described in the proposed publication. Likewise, through consultation with appropriate university officials, graduate student theses or dissertations containing invention details may be withheld from the Library shelves for a limited period while this evaluation process is conducted.

Some research agreements may involve WTC access to a sponsor's proprietary data. In all such cases, a clause defining the conditions under which such data will be identified, accepted, used, and controlled shall be included in the up-front agreement referred to in Paragraph 3. or in an amendment thereto. (Where the work is related to a thesis, students must be able to participate in such research in a meaningful way without access to such proprietary data).

When publication of research results based on use of such proprietary data is contemplated, the WTC will agree to provide the sponsor with advance copy of any proposed publication prior to submission for publication to allow the sponsor an opportunity to identify any inadvertent disclosure of its proprietary data.

9. Consultation with commercial enterprises by WTC technical experts can be of significant benefit to the WTC, the employee, the commercial entity and the general public. However, such involvements include the potential for conflicts of interest, for the inhibition of the free exchange of information, and for interference with the experts' allegiance to the WTC and to their university if they also have university affiliations. In order to minimize the potential for such conflicts and as a condition for continued involvement in WTC research projects, all proposed consulting arrangements by WTC staff must be approved by the Executive Director of the WTC, in addition to approval by the appropriate authorities in their respective universities.

Invention clauses in any such consulting agreements must be consistent with the policy of the WTC, with WTC commitments under sponsored research agreements, and, where the consultant is employed by a university, with the policies of that university. Questions concerning potential conflicts should be referred to the Executive Director or Associate Director of the WTC through appropriate university authorities.

10. In the event that the WTC is terminated or ceases to operate for whatever reason, its ownership of inventions, patents and copyrights, whether administered directly by itself or assigned to WRF or another agent, shall revert to the university at which the research leading to the invention, patent or copyright was carried out in accordance with specific agreements when more than one university is involved.

11. The Technology Transfer Committee of the WTC's Board of Directors is responsible for oversight of the WTC Patent and Copyright Policy.

AGREEMENT

AGREEMENT made as of November 12, 1985 between the Washington Research Foundation (the "Foundation") and the Washington Technology Center (the "Center").

RECITALS

The Foundation has been formed to stimulate productive commercial applications of inventions and other technology discovered and developed at the Center as well as other research institutions in the State of Washington. The Center and the Foundation wish to provide for the disclosure to the Foundation of certain technology (the "Technology"), which may presently or hereafter be owned by the Center, for the purpose of development and management of such Technology by the Foundation, including licensing and marketing of such Technology, the pursuit of patent applications, and the development of commercial applications for such Technology.

AGREEMENTS

1. Submission and Evaluation of Technology. The Center may from time to time deliver to the Foundation, at the Center's sole discretion, disclosures of Technology (each such disclosure referred to herein as a "Technology Project"), and the Foundation agrees to evaluate each Technology Project expeditiously. If in the Foundation's judgment the Technology has significant commercial potential, the Foundation will use its best efforts to introduce the Technology Project into commercial use and to secure royalties or other compensation therefrom as it deems appropriate. If the Foundation decides not to pursue the development of the Technology Project, it will so inform the Center in writing no later than ninety (90) days after initial receipt by the Foundation of the Center's disclosure of the Technology Project and, with such notice, shall return to the Center all materials embodying, reflecting or describing the Technology Project. If the Foundation accepts the Technology Project for commercialization, the Foundation will promptly notify the Center of such acceptance in writing. Upon such notification, the Center will assign to the Foundation all rights of the Center in such Technology Project and will execute such instruments as may be necessary to secure the ownership, right, title and interest in the Foundation of such Technology Project, subject to the provisions of this Agreement. The Foundation will thereafter, with due diligence, undertake the commercialization of the Technology Project.

2. Confidentiality. All disclosures made by the Center to the Foundation with respect to Technology shall be treated by the Foundation as confidential in their entirety. It is understood by the Foundation that all disclosures under this Agreement with respect to Technology are made for the exclusive and limited purpose of providing the Foundation with information necessary for it to assess the development potential of the Technology to which such disclosures relate. Until the Foundation has decided to pursue development of a given Technology and until the Center and the Foundation have entered into the agreements contemplated by this Agreement with respect to the assignment of ownership rights in such Technology to the Foundation, the Foundation may not under any circumstances communicate such Technology or such disclosures to any other persons except as may be necessary on a strict need-to-know basis in order to accomplish the evaluations contemplated by this Agreement, nor may the Foundation put such Technology or disclosures to any use other than as provided in this Agreement. Such limited communication is to be restricted to the maximum extent practicable and shall in all cases be restricted to persons who are subject to this Agreement or who enter into equivalent agreements to preserve the secrecy of all such disclosures and Technology. Any agreement entered into between the Center and the Foundation with respect to the conveyance of ownership rights in Technology shall contain provisions adequate to protect the continuing interest of the Center in such Technology in light of any residual or reversionary interest which the Center may retain in such Technology under such conveyance. The provisions of this paragraph and the obligations imposed hereby shall survive the termination of this Agreement for any reason whatsoever.

3. Costs and Expenses. The Foundation will pay all costs and expenses of the evaluation, patenting, licensing or other administration of transfer of each Technology Project but shall be reimbursed therefor out of royalty income from the Technology Project received by the Foundation as set forth in Section 4.

4. Royalties.

4.1 Distribution. The Foundation shall pay to the Center 62.5% of all royalty income from any Technology Project, after reimbursement of all Directly Allocable Costs (as defined in Paragraph 6 hereof). Because of the interest of the Center and the Foundation in the successful development of the Foundation during its formative years, the parties agree that full distribution to the Center of the above-stated share of

royalties with respect to each Technology project shall commence with the 1986 calendar year and shall be payable from January 1, 1986, unless an earlier date for such full distribution of royalties is mutually agreed upon. Until such date as such full distribution becomes payable, the parties agree that 20% of gross royalty income received by the Foundation with respect to each Technology Project shall be paid to the Center.

4.2 Royalty Payments and Accounts. Payments to the Center shall be made annually on a calendar year basis no later than January 31 for the immediately preceding calendar year. Such payment will be accompanied by a full accounting of the previous year's transactions. The Foundation shall keep accounts and records in sufficient detail to enable the royalties to be determined. Upon reasonable notice to the Foundation, such records shall be made available for inspection by an authorized representative of the Center at reasonable times and places to the extent reasonably necessary (i) to verify the accuracy of the annual reports and royalties paid and (ii) to perform at the Center's expense an audit thereof if requested by the Center. If any audit conducted in accordance with the preceding sentence shall have disclosed an underpayment of 10% or more from what had been represented by the Foundation to the Center, the Foundation will pay for the entire cost of such audit and will promptly pay to the Center as royalties an amount equal to the difference between the amount which it paid to the Center and the amount the audit discloses it should have paid to the Center.

5. Review of Foundation Financial Circumstances. A thorough review of the financial circumstances of the Foundation will be made by representatives of the Center and of the Foundation not less often than annually. Such review may also be made at any time upon the request of the Center with reasonable notice to the Foundation. On any such occasion, the Foundation will make available to the Center any financial records the Center may request.

6. Directly Allocable Costs. The term "Directly Allocable Costs" shall mean the Foundation's out-of-pocket expenses and similar costs related to a Technology Project whenever incurred during the term of this Agreement, including without limitation the costs of obtaining patents, consulting fees paid to third parties in respect to the Technology Project, travel expenses and telephone and reproduction costs, but excluding the costs of evaluating the Technology Project pursuant to paragraph 1 hereof. It does not include any portion of general salaries, rent and overhead of the Foundation.

7. Dissolution of Foundation.

In the event the Foundation ceases to operate or takes legal steps to dissolve, the Foundation will accomplish the following prior to dissolution:

7.1 Pay to the Center all cumulative royalty income due to the Center.

7.2 Reassign to the Center all rights, title and interest in all Technology and Technology Project previously assigned to the Foundation and assign to the Center all right, title and interest in any improvements and developments derived from such Technology and Technology Project. Such reassignment to the Center shall also involve a reassignment of any and all license, royalty or other agreements related to any Technology Project.

8. Termination.

8.1 In the event that the Foundation fails in its obligations hereunder either with respect to the payment of royalties or with respect to the prompt and vigorous development of any Technology or Technology Projects assigned to it by the Center as contemplated by this Agreement, the Center may at its option and upon thirty days written notice to the Foundation, terminate this agreement either with respect to the specific Technology Project as to which such failure of payment or development has occurred, or with respect to this Agreement as a whole. Upon such termination, any and all license agreements relating to any Technology Project shall not terminate but the Center shall automatically be substituted for the Foundation as a party to such agreements and all rights and obligations of the Foundation shall thereupon automatically be assigned to and become vested in the Center, provided, however that the Foundation shall continue to receive continuing payments in the same amount as it would have retained pursuant to Paragraph 4 of this Agreement after payment to the Center thereunder. All license, royalty and other agreements with respect to any Technology Project shall expressly identify that such agreement is subject to the terms and conditions of this Agreement and may be assignable to the Center pursuant to the terms of this Agreement.

8.2 Either the Foundation or the Center may terminate this Agreement at any time upon thirty days written notice, but in no event prior to December 31, 1986, with respect to any future assignments of Technology Projects by the Center to the

Foundation. In such event, all rights and obligations hereunder with respect to Technology or Technology Projects earlier assigned to the Foundation shall, subject to Sections 8.1 and 8.3 hereof, continue in full force and effect according to their terms and shall not be affected by a termination under this Section 8.2.

8.3 This Agreement may be terminated at any time by mutual agreement.

9. Miscellaneous.

9.1 This Agreement constitutes the entire agreement between the parties with respect to the subject matter hereof, and supersedes any prior agreements, understandings, promises and representations made by either party to the other concerning the subject matter hereof and the terms applicable hereto. This Agreement may not be amended or modified except by an instrument in writing signed by duly authorized officers or representatives of both parties hereto.

9.2 If any provision of this Agreement is, becomes or is deemed invalid, illegal or unenforceable in any jurisdiction, such provision shall be deemed amended to conform to applicable laws so as to be valid and enforceable or, if it cannot be so amended without materially altering the intention of the parties, it shall be stricken and the remainder of this Agreement shall remain in full force and effect.

9.3 This Agreement shall be governed by and construed in accordance with the laws of the State of Washington.

9.4 No waiver of any right under this Agreement shall be deemed effective unless contained in a writing signed by the party charged with such waiver, and no waiver of any right arising from any breach or failure to perform shall be deemed to be a waiver of any future such right or of any other right arising under this Agreement.

9.5 All notices, reports and other communications required under this Agreement shall be in writing and shall be deemed given when delivered in person or five days after mailing by prepaid first-class mail, addressed as follows:

Center: Executive Director
The Washington Technology Center
376 Loew Hall FH-10
University of Washington
Seattle, WA 98195

Foundation: President
Washington Research Foundation
1107 N.E. 45 TH Street
Suite 322
Seattle, WA 98105

or to such other address as either party may specify by notice to the other.

9.6 Neither this Agreement nor any right or obligation arising hereunder may be assigned by either party in whole or in part, without the prior written consent of the other party, which consent may be withheld in the absolute discretion of the other party. This Agreement shall be binding upon any assignor and, subject to the restrictions on assignment herein set forth, inure to the benefit of the successors and assigns of each of the parties hereto.

IN WITNESS WHEREOF, the parties have executed this Agreement on the date first set forth above.

THE WASHINGTON
TECHNOLOGY CENTER

By: Edwin B. Stear
DR. EDWIN B. STEAR

TITLE: EXECUTIVE DIRECTOR
DATED: NOVEMBER 12, 1985

WASHINGTON RESEACH FOUNDATION

By: Patrick Y. Tam
DR. PATRICK Y. TAM

TITLE: PRESIDENT
DATED: NOVEMBER 12, 1985

December 30, 1987

Mr. Robert J. Marks, II
Dept. of Electrical Engineering, FT-10
Electrical Engineering Building
University of Washington
Seattle, WA 98195

Re: U.S. Patent Application
Serial No: 131,012
OPTICAL NEURAL NET MEMORY - Marks et al.
WTC 87-6
Our Reference: WTCC-1-3835

Dear Bob:

The following five documents are included in our files relating to the above-referenced patent application.

1. Optical Processor Architectures for a Class of Continuous Level Neural Nets
2. An Introduction to Neural Networks for Solving Combinatorial Search Problems
3. Alternating Projection Neural Networks
4. Content Addressable Memories: A Relationship Between Hopfield's Neural Net and an Iterative Matched Filter
5. An All Optical Iterative Neural Net Recall Memory

Copies of the first one or two pages of each document are enclosed for your reference.

With respect to each of these documents, please let us know if either of the following conditions applies:


1. The document was published or otherwise made available to the public in printed form more than one year prior to the application filing date, i.e., before December 10, 1986; or
2. The document was published or made available to the public in printed form prior to the application filing date (December 10, 1987), and the document includes subject matter that is pertinent to the invention and that was contributed by a non-inventor, i.e., by Judson McDonnell, J.A. Ritcey or Qwan F. Cheung.

Mr. Robert J. Marks, II
December 30, 1987
Page Two

If the first condition applies to any document, then the document is "prior art" for patent examination purposes, and must be cited to the United States Patent and Trademark Office. If the second condition applies, then a more detailed analysis will be required to determine the status of the document.

Yours very truly

CHRISTENSEN, O'CONNOR,
JOHNSON & KINDNESS


By
Michael G. Toner

MGT/mrw
Enclosure

cc: Mr. Peter Odabashian

ISDL REPORT

ALTERNATING PROJECTION NEURAL NETWORKS

R.J. Marks II, S. Oh, L.E. Atlas and J.A. Ritcey

submitted to

IEEE Trans on CAS

Report 11587

*Interactive System Design Lab
Mail Stop FT-10
University of Washington
Seattle, Washington 98195*

11-5-87

ISDL REPORT

AN INTRODUCTION TO NEURAL NETWORKS FOR
SOLVING COMBINATORIAL SEARCH PROBLEMS

AN INTRODUCTION TO NEURAL NETWORKS FOR SOLVING COMBINATORIAL SEARCH PROBLEMS

University of Washington
Seattle, WA 98195

INTRODUCTION

Birds have a mass density gr to appear in *IEEE Expert* an observation motivated early twentieth century inventors to create air flying machines and, ultimately, to invent the airplane. Recent research in artificial neural networks (ANN) research is similarly motivated between our ears that nature's neural networks work quite well.

An ANN can be loosely defined as a connected array of electronic processors. The processors, or neurons, can be homogeneously or can be partitioned into layers. The neural interconnects can be or otherwise improve some performance. *Report 7787* Interactive System Design Lab Mail Stop FT-10 University of Washington Seattle, Washington 98195

ANNs have intrigued researchers from anthropology containing their papers on neural networks.

7-7-87

digital computer and solution of combinatorial

Optical Processor Architectures for a Class of Continuous Level Neural Nets

Robert J. Marks II, Les E. Atlas and Kwan F. Cheung
Interactive Systems Design Laboratory
FT-10 University of Washington, Seattle, Washington 98195

ABSTRACT

Optical processing architectures are presented for a recently proposed class of continuous level neural networks. Both the feed forward and feedback paths are optical, i.e. no electronics or phase conjugators are used.

INTRODUCTION

Optical neural network architectures have been proposed by a number of researchers [1-5]. Neural net architectures are highly redundant in a distributed manner. As a result, they are resilient to computational inexactitude.

Based on the continuous level neural network (CLNN) model in Ref. [6], we present similar architectures wherein no electronics or phase conjugation is required in the forward or feedback paths. After a review of the basic CLNN model, these architectures are discussed in detail. Potential implementation problems and their solutions are also explored.

A MEMORY EXTRAPOLATION NET

Consider a set of N continuous level linearly independent vectors of length $L > N$: $\{\vec{f}_n | 0 \leq n \leq N\}$. We form the library matrix

$$\underline{F} = [\vec{f}_1 | \vec{f}_2 | \dots | \vec{f}_N]$$

and the interconnect matrix

$$\underline{T} = \underline{F} (\underline{F}^T \underline{F})^{-1} \underline{F}^T \quad (1)$$

AN ALL OPTICAL ITERATIVE

NEURAL NET RECALL MEMORY

Robert J. Marks II

Interactive Systems Design Lab
Department of Electrical Engineering
University of Washington
Seattle, WA 98195

11-3-86

CONTENT ADDRESSABLE MEMORIES:
A RELATIONSHIP BETWEEN HOPFIELD'S NEURAL NET
AND AN ITERATIVE MATCHED FILTER *

Robert J. Marks II
Les E. Atlas
Interactive Systems Design Lab
Department of Electrical Engineering
University of Washington
Seattle, Washington 98195

ABSTRACT

Hopfield's neural net content addressable memory (CAM) is shown to be algorithmically equivalent to an iterative matched filter (IMF) CAM. The IMF CAM can be implemented with fewer operations per iteration. Hopfield's CAM, however, can operate asynchronously and is highly fault tolerant. The algorithms are described in a signal space setting where, for orthogonal library elements, each iteration corresponds to two successive projections -- one onto the subspace spanned by the library elements and the other onto a vertex of a hypercube.

ISDL REPORT

**CONTENT ADDRESSABLE MEMORIES:
A RELATIONSHIP BETWEEN HOPFIELD'S
NEURAL NET AND
AN ITERATIVE MATCHED FILTER**

Report 51887

*Interactive System Design Lab
Mail Stop FT-10
University of Washington
Seattle, Washington 98195*

5-18-87

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

*Interactive Systems Design Laboratory
Department of Electrical Engineering, FT-10
Telephone: (206) 543-6990 or 543-2150*

1-6-88

Micheal Toner
Christensen, O'Conner, Johnson & Kindness
2700 Westin Building
2001 Sixth Avenue
Seattle, WA. 98121

Dear Micheal:

I write in response to your letter of Dec. 30, 1987 concerning the OPTICAL NEURAL NET MEMORY PATENT. All except paper #5 was made available to the public prior to Dec. 12, 1987.

1. The paper:

R.J. Marks II, L.E. Atlas and K.F. Cheung "Optical processor architectures for a class of continuous level networks"

was submitted for publication to **Optics Letters** in 1987. The paper discusses the first design of the processor described in the subject patent. Cheung's contribution was a comparative literature search to assure that our effort did not overlap published reports of other neural network implementations. He contributed neither to the algorithm development nor to the processor architecture.

2. The paper:

J.G. McDonnel, R.J. Marks II and L.E. Atlas "An introduction to neural networks for solving combinatorial search problems", **IEEE Expert**, (in press) ... invited paper.

was also submitted for publication in 1987. It is a tutorial of other works and deals neither with the algorithm nor the processor of the subject patent.

3. The paper:

R.J. Marks II, S. Oh, L.E. Atlas and J.A. Ritcey "Alternating projection neural networks" **ISDL Report 11587** (submitted for publication to **IEEE Trans. CAS**)

was submitted for publication on November 8, 1987. It discusses in detail the algorithm implemented by the subject processor but does not address implementation. Ritcey's contribution was analysis of the algorithm convergence properties.

4. The paper:

R.J. Marks II and L.E. Atlas "Content addressable memories: a relationship between Hopfield's neural net and an iterative matched filter"

was submitted for publication in prior to December 1986. The paper, however, is a tutorial introduction to neural networks previously proposed by others and does not impact on our Application.

5. The paper:

"An All Optical Iterative Neural Net Recall Memory"

was submitted to the Boeing High Technology Center as an internal document in November 1986. To my knowledge, no copies were made available to the public. The paper served as the first draft for paper #1 above.

I hope this is the information you need.

Best personal regards,

A handwritten signature in black ink, appearing to read "Robert J. Marks II", with a long horizontal flourish extending to the right.

Robert J. Marks II
Professor

cc: *Peter Odabashian*
Les Atlas
Seho Oh


University of Washington Correspondence

INTERDEPARTMENTAL

Interactive Systems Design Laboratory, Department of Electrical Engineering, FT-10

12-12-87

PROPRIETARY CONFIDENTIAL

TO: Les E. Atlas and Seho Oh
From: Bob Marks 

Attached is a copy of the patent application for the optical APNN. According to the WTC, we should keep the fact that there is a patent quiet. We can, however, talk about the technology in papers and at meetings.

A period of about a year and a half is the time typically taken to process the application.

cc. (memo only) Peter Odabastian, WTC

CHRISTENSEN
O'CONNOR
JOHNSON
KINDNESS

LAW OFFICES

PATENT, TRADEMARK AND OTHER
INTELLECTUAL PROPERTY MATTERS

2700 WESTIN BUILDING
2001 SIXTH AVENUE
SEATTLE, WASHINGTON 98121

TELEPHONE: (206) 441-8780

TELECOPIER: (206) 441-0516

TELEX: 4938023

CABLE: PATENTABLE

December 10, 1987

VIA FAR WEST TAXI

Robert J. Marks, II
Dept. of Electrical Engineering, FT-10
Electrical Engineering Building
University of Washington
Seattle, WA 98195

Re: U.S. Patent Application
For: OPTICAL NEURAL NET MEMORY
Our Reference: WTCC-1-3835

Dear Bob:

Enclosed please find a final draft of the above-referenced patent application, together with an attached three page document entitled Combined Declaration and Power of Attorney in Patent Application. Also enclosed is an Assignment of the invention to the Washington Technology Center.


Please arrange for final review of the application by yourself, Mr. Atlas and Mr. Oh. If the application is satisfactory, each of you should sign and date the Combined Declaration in the spaces provided on page 3 of that document. The Combined Declaration should at all times remain attached to the patent application. On the same day that each inventor signs the Combined Declaration, each inventor must also execute the Assignment before a notary public.

Once these steps have been completed, please arrange to have the patent application, attached Combined Declaration and Assignment returned to our office for filing later today in the United States Patent and Trademark Office. In order to accomplish filing today, we should receive the above-listed documents from you no later than 4 p.m.

In a copy of this letter sent to Peter Odabashian, we have enclosed a further document entitled Verified Statement Claiming Small Entity Status - Nonprofit Organization. This document should be executed by an authorized representative of the Washington Technology Center, and then returned to us no later than 4 p.m. for filing with the application.

Yours very truly,

CHRISTENSEN, O'CONNOR,
JOHNSON & KINDNESS

By 
Michael G. Toner

MGT/mrw
Enclosure
cc: Peter Odabashian

CHRISTENSEN
O'CONNOR
JOHNSON
KINDNESS

LAW OFFICES

PATENT, TRADEMARK AND OTHER
INTELLECTUAL PROPERTY MATTERS

2700 WESTIN BUILDING
2001 SIXTH AVENUE
SEATTLE, WASHINGTON 98121

TELEPHONE: (206) 441-8780

TELECOPIER: (206) 441-0516

TELEX: 4938023

CABLE: PATENTABLE

December 8, 1987

VIA FAR WEST TAXI

Robert J. Marks, II
Dept. of Electrical Engineering, FT-10
Electrical Engineering Building
University of Washington
Seattle, WA 98195

Re: U.S. Patent Application
For: OPTICAL NEURAL NET MEMORY
Our Reference: WTCC-1-3835

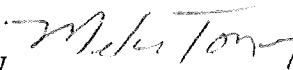
Dear Bob:

Enclosed please find a draft of the above-referenced patent application. During your review of the application, please keep in mind that the application must provide enough information to enable a person of ordinary skill in the art to make and use the invention, and must disclose the best mode known to the inventors at this time for carrying out the invention.

When you have completed your review, please call with your comments. If you will be sending us a marked-up copy of the enclosed draft, we will arrange for a delivery service if you wish. Once we have received and incorporated your corrections, we will then place the application in final form for the signatures of all inventors, and then transmit the application to the United States Patent and Trademark Office. The application must be transmitted no later than Friday, December 11.

Yours very truly,

CHRISTENSEN, O'CONNOR,
JOHNSON & KINDNESS

By 
Michael G. Toner


MGT/mrw
Enclosure

cc: Peter Odabashian, w/ encl.

University of Washington Correspondence

INTERDEPARTMENTAL

11-17-87

To: Peter A. Odabashian
WTC, mail stop FH-10
From: Robert J. Marks II 
Subject: APNN Patent

I talked with Mike Toner on the phone about some further developments on the patent. We decided that the best procedure is to write this memo with a copy to Mike.

The new issues are due, in part, to my colleague Les Atlas and student Seho Oh. Both were supported to some extent by the Boeing money. I understand from Mike that, although the Patent Office does not partition contributions in per cent, such can be done by us and kept on file at the Patent Office. I suggest the following partition:

Les. E. Atlas.....	15%
Seho Oh.....	10%
Robert J. Marks II.....	75%

(Oh is not a US citizens if that matters.) -I have not spoken to these men, but have little doubt that they will agree.

cc: *Mr. Mike Toner*
2700 Westin Bldg.
2001 6th Ave.
Seattle, WA 98195

ADDENDA TO OPTICAL IMPLEMENTATIONS OF ALTERNATING PROJECTION NEURAL NETWORKS

Reference: The APNN paper refers to:

R.J. Marks II, S. Oh, L.E. Atlas and J.A. Ritcey "Alternating projection neural networks", ISDL report 11587 (submitted to IEEE Trans. CAS)

1. Hidden Layers

The number of input-output relationships that can be stored in an APNN is equal to the number of clamped input neurons. The number of input neurons (and thus the capacity of the APNN) can be increased artificially by establishing a hidden layer of neurons. (See section 6 on p.17 of the APNN paper.) The states of the hidden layer neurons can be nearly any nonlinear combination of the states imposed on the hidden layers. The nonlinearity from one hidden neuron to the next, however, must be different. In the optical APNN, this is done by using arbitrary nonlinear electronics prior to the input source array to generate the states of these hidden neurons which, in turn, are used to intensity modulate the input light source corresponding to that hidden neuron. In contrast to the Hopfield model, we are, in essence, placing nonlinearities prior to the input rather than in the feedback path.

2. Binary Outputs

If there is a single output neuron in an APNN and the neural state is known to be either 1 or -1, then the sign of the output state is the correct result after one iteration. As is outlined in the APNN paper (Case 1 on p.12 and remark *f* on pp. 22-23), by superposition, this result can be extended to an arbitrary number of output neurons as long as each output neuron was trained on only plus and minus ones. The implication for the optical APNN architecture is that no feedback is required. Furthermore, the problem with absorptive losses is no longer an issue in this case.

3. Learning

Learning addresses the matter in which the interconnect matrix transmittance is updated when new library vectors are to be stored in the neural network. The Gram-Schmidt learning procedure (Section 5 on p.16 of the APNN paper and remark *c* on p.22) can be directly applied to the optical APNN architectures by making the following changes:

- (a) The entire transmittance matrix must be available (i.e. T instead of just T_Q).
- (b) The source array must be extended to include those neurons whose state, in playback, is determined by the fibers, i.e. the output neurons. Similarly, the output detector array must be extended to include sensing those states normally associated with the floating (input) and hidden layer neurons.

The entire input array is excited corresponding to the new library vector. The error vector, ϵ , is read by the output array. The neural interconnects are updated in accordance to the equation on p.17 of the APNN paper. This can be done with conventional electronics.



The Washington Technology Center

376 Loew Hall, FH-10, University of Washington, Seattle, WA 98195

Office of the Executive Director
(206) 545-1920

November 10, 1987

TO: James B. Wilson, Senior Assistant Attorney General
University of Washington, AG-50

FROM: Lynn M. Fleming, Director of Administration

Subject: Appointment of Special Assistant Attorney General for
Patent Counsel

This is to request the appointment of Mr. Mike Toner of Christensen, O'Conner, Johnson, and Kindness, 2701 Westin Building, 2001 Sixth Avenue, Seattle, WA 98121 as special assistant attorney general to assist The Washington Technology Center (WTC) at the University of Washington in the application for a patent covering an invention entitled **An Optical Continuous Level Neural Network** by R.J. Marks, II (WTC #87-6.) The inventor is a member of the University's Department of Electrical Engineering and developed the invention through a WTC research project supported by a WTC contract. Consistent with the UW/WTC Memorandum of Agreement and the WTC Patent and Copyright Policy, the inventor is in the process of assigning his rights to this invention to The Washington Technology Center, retaining certain rights to royalties as provided by the policy. Patent coverage is deemed to be essential for the effective commercialization of this invention.

Mr. Toner will be the responsible attorney for filing and prosecution of the patent application with reimbursement of actual hourly services at the rate of \$145.00 per hour. Periodic billings will be based on the hourly rate multiplied by the number of hours expended on the case plus other actual out-of-pocket expenses. The total estimated cost for this patent application is \$8,000 which will be charged to the Center's Technology Transfer budget account. The appointment should be for a period of four years.

It is in the best interest of the inventors, the Center, and the state to secure the patent rights to this property as soon as possible. Accordingly, it is requested that this request be processed expeditiously, thereby authorizing the services required for early filing of the patent application.

Thank you for your assistance.

cc. R.J. Marks, II

ADMINISTRATIVE ORDER NO. 17

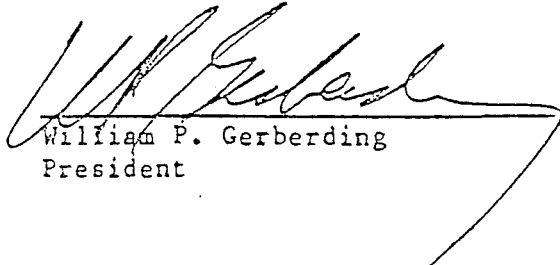
Effective October 18, 1985

SUBJECT: Exemption of the Washington Technology Center from the University of Washington Patent and Copyright Policies and delegation of authority to the WTC to have and administer its own Patent and Copyright Policy subject to certain conditions.

AUTHORITY: University Handbook, Volume II, Part I, Chapter 12, Sections 12-11 and 12-12.

- A. The Washington State Legislature, in Chapter 72, Section 11, Laws of the 1983 1st Extraordinary Session, with the concurrence of the Governor, has established The Washington Technology Center(WTC) at the University of Washington(UW) to be administered by the Board of Regents of the UW. Accordingly, unless otherwise specified, the WTC is subject to UW policies. However, the WTC Board of Directors and the UW Administration, acting under delegated authority from the UW Board of Regents, have agreed that in light of the purposes, goals, objectives and intended nature of the WTC, it should not be fully subject to UW Patent and Copyright Policies but should adopt its own Patent and Copyright Policy.
- B. The WTC is exempted from UW Patent and Copyright Policies subject to certain conditions as follows:
1. the WTC may identify itself as the owner of inventions, patents and copyrights derived from WTC projects;
 2. those inventions, patents and copyrights will be administered under a WTC Patent and Copyright Policy approved by the WTC Board and the UW Administration; and
 3. the WTC will enter into a Technology Administration Agreement (TAA) with the Washington Research Foundation(WRF) that is identical in all substantive respects with the TAA between UW and WRF attached hereto as Exhibit A.

This Administrative Order No. 17 is pursuant to the authority cited above.


William P. Gerberding
President

THE WASHINGTON TECHNOLOGY CENTER
PRINCIPLES GOVERNING
PATENT AND COPYRIGHT POLICIES AND PROCEDURES

1. The Washington Technology Center, hereafter referred to as the WTC, shall own all patents and copyrights arising from WTC sponsored research and technology development programs and projects.
2. The WTC shall negotiate all patent and copyright agreements and licensing arrangements so as to maximize technology transfer for the benefit of the economic development of the State of Washington.
3. Negotiations of patent and copyright agreements and subsequent licensing arrangements shall be the responsibility of the duly appointed individual in charge of the WTC Office at the appropriate participating university in accordance with WTC policies and procedures.
4. The WTC shall develop a Patent and Copyright Policy which will form the basis for negotiation of specific agreements on patents, copyrights, licensing, and distribution of royalty income with each of the participating universities.
5. The WTC shall negotiate up-front patent and copyright agreements, including licensing provisions, with all participating industrial sponsors of WTC programs and projects.
6. The WTC shall negotiate individual up-front patent and copyright agreements with all Industrial Fellows and their employers.
7. All individuals participating in WTC programs and/or projects shall sign an agreement requiring them to be bound by the WTC's Patent and Copyright Policy.
8. When investigators from more than one university work on a WTC project, there shall be a specific up-front agreement among all parties covering patent and copyright issues, including negotiation of agreements with industrial supporters of the project, negotiation of licenses for any intellectual property developed, distribution of royalty income, and ownership of any patents or copyrights in the event the WTC is terminated or ceases to operate for any reason.

THE WASHINGTON TECHNOLOGY CENTER

Patent and Copyright Policy

1. One of the primary missions of The Washington Technology Center (hereinafter referred to as WTC) is to develop new commercializable technology through joint industry-university research and technology development programs. Patents and copyrights are important in this process to:

- (a) protect the economic interests of the WTC and the inventors.
- (b) protect the economic interests of the industrial participants and the licensees.
- (c) provide a firm legal basis for transferring the technology.

It is recognized that the value of the technology may diminish rapidly with time. Therefore, it will often be necessary to transfer technology immediately after disclosure and prior to application for or issuance of patents and copyrights.

Further, it is recognized that it will also be necessary to transfer technology without applying for patents or copyrights in those cases where the technology is not patentable or copyrightable, or where the value of the particular patent or copyright does not justify the expense.

The purpose of this document is to set forth the specific policies adopted by the WTC to assure that these requirements and goals are met.

2. As a condition of participation in WTC research projects, all personnel participating in WTC projects agree to assign their title and rights to all inventions and copyrightable material arising in connection with such research projects to the WTC, to an agent designated by the WTC, or to a sponsor, if required under agreements governing sponsored research. Such personnel shall execute documents of assignment and do everything reasonably required to assist the assignee(s) in obtaining, protecting, and maintaining patents, copyrights or other proprietary rights.

The WTC has no vested interest in inventions or copyrightable material conceived and developed by participants entirely on their own time and without the use of WTC facilities. However, in order to clarify the inventor's or creator's

title to such inventions and/or copyrightable material and to insure compliance with the requirements of any sponsors, all inventions and/or copyrightable material generated during participation in WTC programs and projects shall be reported to the WTC for determination of the degree of WTC interest.

If the WTC, in consultation with the appropriate participating universities, determines that it has no interest in an invention or copyrightable material or decides to forego the patenting, copyrighting, or other commercialization of an invention or copyrightable material, it shall waive its rights to the invention or copyrightable material in writing. Upon receipt of such a waiver, and assuming that no additional WTC or University resources will be invested, the inventor(s) or creator(s) may file a patent or copyright application and/or grant a license of his/her own.

3. WTC research funded wholly or in part by an outside sponsor is subject to this policy as modified by the provisions of negotiated agreement(s) covering such work. It is the general policy of the WTC to negotiate all such agreements, including any special provisions relating to the intellectual property, prior to initiation of the research effort being sponsored. Participants in such sponsored research are bound by the provisions of these agreements.

4. In general, title to any inventions and/or copyrightable material conceived and first reduced to practice in the course of research carried out in the WTC with the support of Federal agencies, industry, or other sponsors shall vest in the WTC. In rare cases, an industrial sponsor may possess a dominant patent or copyright position in a certain technology area so that any patent or copyright the WTC might seek would be of little value. For this or other such reasons, an exception to this WTC title policy may be approved when to do so would honor the general principles of this policy, protect the equities involved, and satisfy the requirements of the parties. In all cases, the granting of such exceptions must be explicitly covered in the agreements referred to above in Paragraph 3.

5. Interaction between the WTC and industry can take any one or more of the following forms: grants, contracts, consortial arrangements, equipment gifts, and appointment of industrial fellows. Industrial firms sponsoring WTC research programs through any one or more of these forms may be assured of at least a non-exclusive license to inventions and copyrights conceived and developed with their support. If necessary for the effective development and marketing of a WTC invention or copyright, an exclusive license may be granted for a limited period of time if the sponsor agrees to finance the cost of the WTC's patent or copyright application and observes due

diligence in bringing the technology involved into public use. In such cases, the patent or copyright costs may be treated as an offset against royalties payable when the invention or copyright is marketed.

Where the sponsor uses the invention or copyright entirely within its own operations, the license may be royalty-free. Where the sponsor, or a third party licensee, manufactures and sells products, services, or processes based on the invention or copyright, reasonable royalty payments to the WTC or its assignee are normally required.

In all cases involving industrial sponsorship of WTC research programs, the specific licensing rights of the sponsor(s) to any patentable and/or copyrightable technology generated in the research programs shall be explicitly covered in the up-front agreements referred to above in Paragraph 3.

6. Although the WTC reserves the right to patent and/or copyright intellectual property itself, it has designated the Washington Research Foundation as its primary patenting, copyrighting, and licensing agent. However, another comparable, mutually-acceptable patenting, copyrighting and licensing agent can be used if so desired by an individual participating university.

7. Both the inventors and/or creators and the WTC are entitled to a share of royalty income from licensed patents and/or copyrights; the WTC on the basis of salary and/or facilities support for the inventor and/or creator and the cost of patent, copyright, and licensing administration; and the inventor and/or creator on the basis of the creative activity, documenting the invention or copyright, and assisting as necessary with commercialization. To recognize creativity and to encourage prompt disclosure of potential patents and copyrights, the WTC allocates the greater share of net early royalty income to the inventor or creator. The remainder is dedicated to further research by allocating shares to the WTC and to the home colleges/departments of the inventors and/or creators as appropriate. Unless amended in an agreement with a participating university, the specific allocation shall be as follows.

After deducting 15% for administrative services, net royalty income received from WTC inventions and/or copyrights handled by an outside agency is distributed as follows:

Cumulative Net Income	Inventor/ Creator	Inventor's University Dept./College	WTC Research Fund
First \$10,000	100%	0%	0%
\$10,000-\$40,000	50%	25%	25%
Above \$40,000	30%	20%	50%

In the event that an invention and/or copyright is administered directly by the WTC, the direct costs of obtaining and maintaining the patent(s) and/or copyright(s) must be recovered in addition to the 15% service fee before distribution of royalty income begins under the above formula.

The royalty derived WTC Research Fund shall be used to promote additional research in areas identified for emphasis by the WTC.

When a proposed WTC program or project involves more than one university, it is the general policy of the WTC to negotiate an up-front agreement with the participating universities covering patent and copyright issues. Including negotiation of agreements with industrial supporters of the project, negotiation of licenses for any intellectual property developed and distribution of royalty income and ownership of any patents and copyrights in the event the WTC is terminated or ceases to operate for any reason.

8. As a public institution, the WTC should undertake sponsored research under conditions which permit timely publication of the research results. However, the WTC reserves the right to defer publication for a reasonable period of time during which the WTC and any sponsor(s) review the feasibility and desirability of patent and/or copyright protection of any intellectual property described in the proposed publication. Likewise, through consultation with appropriate university officials, graduate student theses or dissertations containing invention details may be withheld from the Library shelves for a limited period while this evaluation process is conducted.

Some research agreements may involve WTC access to a sponsor's proprietary data. In all such cases, a clause defining the conditions under which such data will be identified, accepted, used, and controlled shall be included in the up-front agreement referred to in Paragraph 3. or in an amendment thereto. (Where the work is related to a thesis, students must be able to participate in such research in a meaningful way without access to such proprietary data).

When publication of research results based on use of such proprietary data is contemplated, the WTC will agree to provide the sponsor with advance copy of any proposed publication prior to submission for publication to allow the sponsor an opportunity to identify any inadvertent disclosure of its proprietary data.

9. Consultation with commercial enterprises by WTC technical experts can be of significant benefit to the WTC, the employee, the commercial entity and the general public. However, such involvements include the potential for conflicts of interest, for the inhibition of the free exchange of information, and for interference with the experts' allegiance to the WTC and to their university if they also have university affiliations. In order to minimize the potential for such conflicts and as a condition for continued involvement in WTC research projects, all proposed consulting arrangements by WTC staff must be approved by the Executive Director of the WTC, in addition to approval by the appropriate authorities in their respective universities.

Invention clauses in any such consulting agreements must be consistent with the policy of the WTC, with WTC commitments under sponsored research agreements, and, where the consultant is employed by a university, with the policies of that university. Questions concerning potential conflicts should be referred to the Executive Director or Associate Director of the WTC through appropriate university authorities.

10. In the event that the WTC is terminated or ceases to operate for whatever reason, its ownership of inventions, patents and copyrights, whether administered directly by itself or assigned to WRF or another agent, shall revert to the university at which the research leading to the invention, patent or copyright was carried out in accordance with specific agreements when more than one university is involved.

11. The Technology Transfer Committee of the WTC's Board of Directors is responsible for oversight of the WTC Patent and Copyright Policy.

WTC
10/23/85

AGREEMENT

AGREEMENT made as of November 12, 1985 between the Washington Research Foundation (the "Foundation") and the Washington Technology Center (the "Center").

RECITALS

The Foundation has been formed to stimulate productive commercial applications of inventions and other technology discovered and developed at the Center as well as other research institutions in the State of Washington. The Center and the Foundation wish to provide for the disclosure to the Foundation of certain technology (the "Technology"), which may presently or hereafter be owned by the Center, for the purpose of development and management of such Technology by the Foundation, including licensing and marketing of such Technology, the pursuit of patent applications, and the development of commercial applications for such Technology.

AGREEMENTS

1. Submission and Evaluation of Technology. The Center may from time to time deliver to the Foundation, at the Center's sole discretion, disclosures of Technology (each such disclosure referred to herein as a "Technology Project"), and the Foundation agrees to evaluate each Technology Project expeditiously. If in the Foundation's judgment the Technology has significant commercial potential, the Foundation will use its best efforts to introduce the Technology Project into commercial use and to secure royalties or other compensation therefrom as it deems appropriate. If the Foundation decides not to pursue the development of the Technology Project, it will so inform the Center in writing no later than ninety (90) days after initial receipt by the Foundation of the Center's disclosure of the Technology Project and, with such notice, shall return to the Center all materials embodying, reflecting or describing the Technology Project. If the Foundation accepts the Technology Project for commercialization, the Foundation will promptly notify the Center of such acceptance in writing. Upon such notification, the Center will assign to the Foundation all rights of the Center in such Technology Project and will execute such instruments as may be necessary to secure the ownership, right, title and interest in the Foundation of such Technology Project, subject to the provisions of this Agreement. The Foundation will thereafter, with due diligence, undertake the commercialization of the Technology Project.

2. Confidentiality. All disclosures made by the Center to the Foundation with respect to Technology shall be treated by the Foundation as confidential in their entirety. It is understood by the Foundation that all disclosures under this Agreement with respect to Technology are made for the exclusive and limited purpose of providing the Foundation with information necessary for it to assess the development potential of the Technology to which such disclosures relate. Until the Foundation has decided to pursue development of a given Technology and until the Center and the Foundation have entered into the agreements contemplated by this Agreement with respect to the assignment of ownership rights in such Technology to the Foundation, the Foundation may not under any circumstances communicate such Technology or such disclosures to any other persons except as may be necessary on a strict need-to-know basis in order to accomplish the evaluations contemplated by this Agreement, nor may the Foundation put such Technology or disclosures to any use other than as provided in this Agreement. Such limited communication is to be restricted to the maximum extent practicable and shall in all cases be restricted to persons who are subject to this Agreement or who enter into equivalent agreements to preserve the secrecy of all such disclosures and Technology. Any agreement entered into between the Center and the Foundation with respect to the conveyance of ownership rights in Technology shall contain provisions adequate to protect the continuing interest of the Center in such Technology in light of any residual or reversionary interest which the Center may retain in such Technology under such conveyance. The provisions of this paragraph and the obligations imposed hereby shall survive the termination of this Agreement for any reason whatsoever.

3. Costs and Expenses. The Foundation will pay all costs and expenses of the evaluation, patenting, licensing or other administration of transfer of each Technology Project but shall be reimbursed therefor out of royalty income from the Technology Project received by the Foundation as set forth in Section 4.

4. Royalties.

4.1 Distribution. The Foundation shall pay to the Center 62.5% of all royalty income from any Technology Project, after reimbursement of all Directly Allocable Costs (as defined in Paragraph 6 hereof). Because of the interest of the Center and the Foundation in the successful development of the Foundation during its formative years, the parties agree that full distribution to the Center of the above-stated share of

royalties with respect to each Technology project shall commence with the 1986 calendar year and shall be payable from January 1, 1986, unless an earlier date for such full distribution of royalties is mutually agreed upon. Until such date as such full distribution becomes payable, the parties agree that 20% of gross royalty income received by the Foundation with respect to each Technology Project shall be paid to the Center.

4.2 Royalty Payments and Accounts. Payments to the Center shall be made annually on a calendar year basis no later than January 31 for the immediately preceding calendar year. Such payment will be accompanied by a full accounting of the previous year's transactions. The Foundation shall keep accounts and records in sufficient detail to enable the royalties to be determined. Upon reasonable notice to the Foundation, such records shall be made available for inspection by an authorized representative of the Center at reasonable times and places to the extent reasonably necessary (i) to verify the accuracy of the annual reports and royalties paid and (ii) to perform at the Center's expense an audit thereof if requested by the Center. If any audit conducted in accordance with the preceding sentence shall have disclosed an underpayment of 10% or more from what had been represented by the Foundation to the Center, the Foundation will pay for the entire cost of such audit and will promptly pay to the Center as royalties an amount equal to the difference between the amount which it paid to the Center and the amount the audit discloses it should have paid to the Center.

5. Review of Foundation Financial Circumstances. A thorough review of the financial circumstances of the Foundation will be made by representatives of the Center and of the Foundation not less often than annually. Such review may also be made at any time upon the request of the Center with reasonable notice to the Foundation. On any such occasion, the Foundation will make available to the Center any financial records the Center may request.

6. Directly Allocable Costs. The term "Directly Allocable Costs" shall mean the Foundation's out-of-pocket expenses and similar costs related to a Technology Project whenever incurred during the term of this Agreement, including without limitation the costs of obtaining patents, consulting fees paid to third parties in respect to the Technology Project, travel expenses and telephone and reproduction costs, but excluding the costs of evaluating the Technology Project pursuant to paragraph 1 hereof. It does not include any portion of general salaries, rent and overhead of the Foundation.

7. Dissolution of Foundation.

In the event the Foundation ceases to operate or takes legal steps to dissolve, the Foundation will accomplish the following prior to dissolution:

7.1 Pay to the Center all cumulative royalty income due to the Center.

7.2 Reassign to the Center all rights, title and interest in all Technology and Technology Project previously assigned to the Foundation and assign to the Center all right, title and interest in any improvements and developments derived from such Technology and Technology Project. Such reassignment to the Center shall also involve a reassignment of any and all license, royalty or other agreements related to any Technology Project.

8. Termination.

8.1 In the event that the Foundation fails in its obligations hereunder either with respect to the payment of royalties or with respect to the prompt and vigorous development of any Technology or Technology Projects assigned to it by the Center as contemplated by this Agreement, the Center may at its option and upon thirty days written notice to the Foundation, terminate this agreement either with respect to the specific Technology Project as to which such failure of payment or development has occurred, or with respect to this Agreement as a whole. Upon such termination, any and all license agreements relating to any Technology Project shall not terminate but the Center shall automatically be substituted for the Foundation as a party to such agreements and all rights and obligations of the Foundation shall thereupon automatically be assigned to and become vested in the Center, provided, however that the Foundation shall continue to receive continuing payments in the same amount as it would have retained pursuant to Paragraph 4 of this Agreement after payment to the Center thereunder. All license, royalty and other agreements with respect to any Technology Project shall expressly identify that such agreement is subject to the terms and conditions of this Agreement and may be assignable to the Center pursuant to the terms of this Agreement.

8.2 Either the Foundation or the Center may terminate this Agreement at any time upon thirty days written notice, but in no event prior to December 31, 1986, with respect to any future assignments of Technology Projects by the Center to the

Foundation. In such event, all rights and obligations hereunder with respect to Technology or Technology Projects earlier assigned to the Foundation shall, subject to Sections 8.1 and 8.3 hereof, continue in full force and effect according to their terms and shall not be affected by a termination under this Section 8.2.

8.3 This Agreement may be terminated at any time by mutual agreement.

9. Miscellaneous.

9.1 This Agreement constitutes the entire agreement between the parties with respect to the subject matter hereof, and supersedes any prior agreements, understandings, promises and representations made by either party to the other concerning the subject matter hereof and the terms applicable hereto. This Agreement may not be amended or modified except by an instrument in writing signed by duly authorized officers or representatives of both parties hereto.

9.2 If any provision of this Agreement is, becomes or is deemed invalid, illegal or unenforceable in any jurisdiction, such provision shall be deemed amended to conform to applicable laws so as to be valid and enforceable or, if it cannot be so amended without materially altering the intention of the parties, it shall be stricken and the remainder of this Agreement shall remain in full force and effect.

9.3 This Agreement shall be governed by and construed in accordance with the laws of the State of Washington.

9.4 No waiver of any right under this Agreement shall be deemed effective unless contained in a writing signed by the party charged with such waiver, and no waiver of any right arising from any breach or failure to perform shall be deemed to be a waiver of any future such right or of any other right arising under this Agreement.

9.5 All notices, reports and other communications required under this Agreement shall be in writing and shall be deemed given when delivered in person or five days after mailing by prepaid first-class mail, addressed as follows:

Center: Executive Director
The Washington Technology Center
376 Loew Hall FH-10
University of Washington
Seattle, WA 98195

Foundation: President
Washington Research Foundation
1107 N.E. 45 TH Street
Suite 322
Seattle, WA 98105

or to such other address as either party may specify by notice to the other.

9.6 Neither this Agreement nor any right or obligation arising hereunder may be assigned by either party in whole or in part, without the prior written consent of the other party, which consent may be withheld in the absolute discretion of the other party. This Agreement shall be binding upon any assignor and, subject to the restrictions on assignment herein set forth, inure to the benefit of the successors and assigns of each of the parties hereto.

IN WITNESS WHEREOF, the parties have executed this Agreement on the date first set forth above.

THE WASHINGTON
TECHNOLOGY CENTER

WASHINGTON RESEACH FOUNDATION

By: Edwin B. Stear
DR. EDWIN B. STEAR

By: Patrick Y. Tam
DR. PATRICK Y. TAM

TITLE: EXECUTIVE DIRECTOR
DATED: NOVEMBER 12, 1985

TITLE: PRESIDENT
DATED: NOVEMBER 12, 1985

5. List all reports, abstracts, papers, theses or patent applications which have been or are planned to be submitted by the inventor(s) describing the invention. Give dates of submission and actual or anticipated publication dates. Attach documents, if available. These documents may be used in part to respond to Section 2.
6. List any other known references, patents, patent applications or other publications pertinent to this invention. Attach copies, if available. These documents may also be used in part to respond to Section 2.
7. Describe and date any sale or public use of the invention in the United States. Specify if the use was operational, or for testing purposes, and if there was any effort or intent to maintain invention secrecy after operational use began.
8. List all co-inventors (any individuals who conceived an essential feature of the invention, either independently or jointly with others, during the evolution of the invention). In the event a patent application is filed, inventorship will be verified by the patent attorney.
9. Arrange for two technically qualified witnesses to read and sign this document verifying that they have understood the invention that is disclosed.

Submit the completed Disclosure together with the Transmittal form to Dr. Edwin B. Stear, Executive Director, Washington Technology Center, University of Washington, Mail Stop FH-10, Seattle, Washington 98195. Generally it will then be forwarded to the Washington Research Foundation (or another agent) for evaluation of patentability and commercial potential.

For further information, contact The Washington Technology Center, (206) 545-1920.

WASHINGTON TECHNOLOGY CENTER

INVENTION DISCLOSURE

This invention Disclosure is an important legal document and should be completed carefully. Please refer to the attached instructions.

1. Title of Invention

An Optical Continuous Level Neural Network

2. Brief Description

A library of continuous level object vectors is stored in two dimensions on an optical transmittance. When a portion of a library vector is input into the optical processor by an array of point light sources, the remainder of the vector is iteratively recovered at light speed.

3. Funding Source(s)

Boeing High Technology Center

4. Invention History	Date	Location and Comments
A. Initial Idea	Oct '86	The optical processor is an implementation of an algorithm proposed by the inventor [1]*
B. First description of complete invention, oral or written	Nov '86	Reference [2]*
C. Invention development records, notes, drawings (evidence of diligence)	Feb '87	Addition of Optical Switches [3]*
D. First successful demonstration, if any (first actual reduction to practice)	none	Some simulations are in [1]*
E. First publication with full description of invention (may bar patent)	12/12/86 2/10/87	Boeing High Tech Center Seminar [4]* U.W. Seminar [5]*
F. First verbal description to others	10/86	to Donald C. Wunsch (Boeing Electronics) at OSA meeting in Seattle.

5. List all reports, abstracts, papers, theses or patent applications related to the inventions which have been published or are planned to be submitted by the Inventor(s). Include copies if available.

***See attached reference list**

6. List any other references, patents, patent applications or other publications which may be pertinent to the invention. Include copies if available.

See attached reference list

7. Describe and date any sale or public use of the invention in the United States.

none

8. Inventor or Co-inventors

Robert J. Marks II 4/27/87
Signature Date

Signature Date

Robert J. Marks II, Assoc. Prof
Name (Print) Title

Name (Print) Title

UWEE Dept, FT-10, 98195
Address

Address

(206) 543-6990
Telephone

Telephone

Signature Date

Signature Date

Name (Print) Title

Name (Print) Title

Address

Address

Telephone

Telephone

9. Invention disclosed to and understood by (two witnesses required):

Signature Date

Signature Date

Name (Print)

Name (Print)

Submit completed Disclosure to the Washington Technology Center,
University of Washington, 376 Loew Hall, M/S FB-10, Seattle, WA
98195.

Date Received: _____
Washington Technology Center

THE WASHINGTON TECHNOLOGY CENTER

**Form to Transmit Invention Disclosure
(For WTC Internal Use Only)**

Instructions

Complete this form and the attached Invention Disclosure form and forward to The Washington Technology Center via WTC Program Director, Department Chairperson, and Dean of School/College for approval. If more than one Department is involved, obtain signatures from all Chairpersons and Deans (or their designate).

To: Washington Technology Center Date: _____
Loew Hall 376, FH-10

From: _____
Inventor Name Title Department Mail Stop

Inventor Name Title Department Mail Stop

Inventor Name Title Department Mail Stop

Inventor Name Title Department Mail Stop

Re: Invention entitled: _____

Verified and Approved:

Concurrence:

WTC Program Director

Dean of the School/College

Date: _____

Date: _____

Concurrence:

Accepted:

Department Chairperson

Edwin B. Stear, Executive Dir.
WASHINGTON TECHNOLOGY CENTER

Date: _____

Date: _____

REFERENCES (and further comments)

1. R.J. Marks II "A Class of Continuous Level Associative Memory Neural Nets" to appear in the 15 May issue of Applied Optics.
(this paper contains the description of the algorithm performed by the processor).
2. R.J. Marks II "An All Optical Iterative Neural Net Recall Memory" (this paper, sent to the BHTC, was an internal document. It first presents optical feedback in an optical neural net architecture. Others have used (slow) feedback electronics).
3. R.J. Marks II "A class of continuous level neural nets and their optical implementation" (these are copies of the slides used at the BHTC seminar on 12-12-86).
4. R.J. Marks II "Optical architectures for a continuous level neural net" (to date, this has been an internal report but will soon be submitted for publication to Applied Optics. The use of optical switches is first suggested here).
5. R.J. Marks II "A continuous level neural net and its optical implementation". (copies of the slides used at a 2-10-87 U.W. seminar. Optical switches were included in the proposed architecture).
6. K.F. Cheung, R.J. Marks II and L.E. Atlas "Neural Net Associative Memories Based on Convex Set Projections" to be presented at the First Annual Conference on Neural Nets in San Diego, June 1987.

Other literature:

7. "Optoelectronics builds viable neural net memory" Electronics June, '86. (a trade journal explanation of optical neural nets).
8. "Optics and Neural Nets" Computer Design, March '87. (a similar but more recent paper).
9. Psaltis and Farhat, Optics Letters 10, Feb. '85. (the first journal paper on optical neural nets. As with other designs, slow electronics is used in the feedback path).
10. Farhat, et.al., Applied Optics 24, 15 May 1985. (a continuation of reference 9 above).

A Class of Continuous Level Associative

Memory Neural Nets

Robert J. Marks II
Interactive Systems Design Lab
University of Washington, FT-10
Seattle, WA 98195

to appear in Applied Optics

ABSTRACT

A neural net capable of restoring continuous level library vectors from memory is considered. As with Hopfield's neural net content addressable memory, the vectors in the memory library are used to program the neural interconnects. Given a portion of one of the library vectors, the net extrapolates the remainder. Necessary and sufficient conditions for convergence are stated. Effects of processor inexactitude and net faults are discussed. A more efficient computational technique for performing the memory extrapolation (at the cost of fault tolerance), is derived. The special case of table-look-up memories is addressed specifically.

INTRODUCTION

Hopfield's neural net content addressable memory (CAM) [1] has stirred great interest in the signal processing community. The net has been implemented both optically [2-5] and electronically [6]. For optical implementation, intensive neural interconnects are possible since light paths can cross without interference. Planar VLSI implementations, on the other hand, are restricted to nearest neighbor interconnects. The interconnects in Hopfield's CAM are programmed by a set of binary library vectors. Given a noisy subset of one of the library vectors, the neural net ideally converges to the library vector closest to the initialization. The net can operate asynchronously or synchronously. It is also tolerant of both lumped and distributed faults [3,6]. Thus, analog optical processor inexactitude is of less significance than usual.

The neural net introduced in this paper allows for library vectors with continuous elements. The interconnects are determined analogous to Hopfield's recipe. The net can also operate asynchronously and is fault tolerant. It differs from Hopfield's in that the initially known neural states are imposed on the net each iteration. That is, the known states act as the net stimulus and the remaining nodes catalog the response. A human memory analogy is our ability to recall a well known painting by continuously viewing only a portion of it.

After a brief introduction to the mathematics of the neural net, we specifically define the extrapolation neural net. Borrowing from some recent results in iterative signal recovery and synthesis [7-11], important insights into the net's performance are generated. These include sufficient conditions for convergence to the proper library vector and

effects of known state perturbations. A short section on fault tolerance contains empirical evidence that the net still works "well" for both quantized and deleted interconnects. A table look-up net is one where the same P nodes are always used as the net stimulus. Neural net architectures for these specific memory extrapolation problems are presented. Some final remarks tying the net's operation to some other well known iterative algorithms are made in the conclusions.

PRELIMINARIES

Consider a neural net of L nodes. The transmission from the k^{th} to the i^{th} node is t_{ik} . We will assume a symmetric net ($t_{k1} = t_{1k}$) and will allow for autointerconnects ($t_{kk} \neq 0$). The state, s_k , of the k^{th} node, will be assumed to be a function of the sum of its inputs. For synchronous operation (i.e., all delays between node pairs are identical), we have at time M

$$\vec{i}_M = \underline{T} \vec{s}_M \quad (1)$$

where \vec{s}_M is a vector of the L neural states at time M, \vec{i}_M is the vector of the L input sums at time M and \underline{T} is the matrix of t_{ik} 's. Let \underline{N} denote the node operator that determines the next set of states from the input sum:

$$\vec{s}_{M+1} = \underline{N} \vec{i}_M \quad (2)$$

Since the state of the k^{th} node depends only on its input sum, \underline{N} must be a pointwise operator. That is, the k^{th} element of \vec{s}_{M+1} depends only on the k^{th} element of \vec{i}_M .

Substituting (1) into (2) gives the state iteration equation:

$$\vec{s}_{M+1} = \underline{N} \underline{T} \vec{s}_M \quad (3)$$

We illustrate with two short examples, saving our memory extrapolation net for a more detailed treatment.

Solving Simultaneous Equations

Consider the L linear equations

$$\vec{g} = \underline{K} \vec{f}$$

Given \vec{g} and \underline{K} , we wish to find \vec{f} . Design a neural net with

$$\underline{I} = \underline{I} - \underline{K}$$

and let the neural operator be defined for an arbitrary vector \vec{i} , by (see Fig. 1a)

$$\underline{N} \vec{i} = \vec{i} + \vec{g}$$

Thus, the k^{th} node adds g_k to the sum of the node's inputs. Then with initialization $\vec{s}_0 = \vec{g}$, (3) can be inductively shown to be equivalent to

$$\vec{s}_M = \sum_{m=0}^M \underline{I}^m \vec{g}$$

If $\|\underline{I}\| < 1$, we can use a generalized geometric series and write:

$$\begin{aligned} \vec{s}_\infty &= [\underline{I} - \underline{I}]^{-1} \vec{g} \\ &= \vec{f} \end{aligned}$$

The net thus ideally converges to our desired result. [12]

Hopfield's Neural Net

Let $\{ \vec{b}_n \mid 1 \leq n \leq N \}$ denote N library vectors each with only ± 1 elements.

Define the library matrix

$$\underline{B} = [\vec{b}_1 : \vec{b}_2 : \dots : \vec{b}_N]$$

From this, we form the interconnect matrix

$$\underline{I} = \underline{B} \underline{B}^T - N \underline{I}$$

where the superscript T denotes transformation. (Note that $t_{kk} = 0$). Let the node operator be (see Fig. 1b):

$$\underline{N} = \text{sgn}$$

where sgn performs a signum operation on each vector element. The resulting neural net is Hopfield's CAM. For an initialization, \vec{g} , and $N \ll L$, the

net's state many times will converge to the library vector closest to \vec{g} in the Hamming sense.

A MEMORY EXTRAPOLATION NET

Consider a set F of N continuous level linearly independent vectors of length $L \geq N$:

$$F = \{ \vec{f}_n \mid 1 \leq n \leq N \}$$

and the corresponding library matrix:

$$\underline{F} = [\vec{f}_1 : \vec{f}_2 : \dots : \vec{f}_N]$$

We form a neural net with interconnects*[5]

$$\underline{T} = \underline{F} (\underline{F}^T \underline{F})^{-1} \underline{F}^T \quad (4)$$

Given a portion of one of the library vectors, a memory extrapolator, using the library, will reconstruct the remainder of that vector. For our net, we will divide the nodes into two sets: one in which states are known and the remainder, in which the states are unknown. This node partition may change from application to application. That is, any node may be used to stimulate or to respond. Without loss of generality, assume that states 1 through $P < L$ (corresponding to the first P elements in some given $\vec{f} \in F$) are known for a given application. Define the node operator by

$$\begin{aligned} \underline{N} \vec{f} &= \underline{N} [i_1 \ i_2 \ \dots \ i_p : i_{p+1} \ \dots \ i_L]^T \\ &= [\delta_1 \ \delta_2 \ \dots \ \delta_p : i_{p+1} \ \dots \ i_L]^T \end{aligned} \quad (5)$$

where δ_k is the k^{th} element of \vec{f} (Fig. 1c). That is, for $1 \leq k \leq P$, the node state is kept at δ_k . Otherwise, the node state is the input sum. The P

* If \underline{F} is not full rank, then we use

$$\underline{T} = \underline{F}^* (\underline{F}^{*T} \underline{F}^*)^{-1} \underline{F}^{*T}$$

where \underline{F}^* is a full rank matrix obtained from discarding appropriate redundant columns from F .

known states thus act as the input or stimulus to the net and the remaining steady state node states are the response.

In summary, the algorithm is this:

1. Initialize with all unknown states set to zero.* The known states are equated to the known portion of the library vector.
2. Multiply the state vector by \underline{I} in (4).
3. Replace states 1 through P with their known values.
4. Go to step 2 and repeat.

In many cases of interest, we claim that this iterative procedure will converge to the desired library vector. The uniqueness of convergence to the proper library element is addressed in the next section.

PERFORMANCE ANALYSIS

In this section, we derive important convergence properties of the memory extrapolation net and analyze the effects of input uncertainty on the net's performance. Some empirical results on the net's fault tolerance are also discussed.

Insight into the net's performance is gained by viewing the corresponding iterative algorithm in an L dimensional Hilbert space, H . Consider first, the N dimensional subspace**, T , spanned by the N library vectors (i.e., T is the closure of F). The matrix \underline{I} in (4) (orthogonally) projects any vector onto that subspace [13]. That is, for any $\vec{h} \in H$,

$$\inf_{\vec{f} \in T} \|\vec{h} - \vec{f}\| = \|\vec{h} - \underline{I}\vec{h}\|$$

*If convergence is unique, any initialization will converge to the correct result.

**Also called a closed linear manifold.

where $\|\vec{a}\|^2 = \vec{a}^T \vec{a}$. Specifically, note that $\underline{T}^2 = \underline{T}$, $\underline{T}\underline{F} = \underline{F}$ and that, for any element \vec{b} orthogonal to T , $\underline{T}\vec{b} = \vec{0}$ where $\vec{0}$ is the zero vector.

To similarly analyze the \underline{N} operator in (5), we adopt the vector partitioning notation

$$\vec{h} = \begin{bmatrix} \vec{h}_p \\ \vec{h}_q \end{bmatrix}$$

where \vec{h}_p is a P and \vec{h}_q is a $Q = L - P$ dimensional vector. Then, for example, the zero vector can be written as $\vec{0} = [\vec{0}_p; \vec{0}_q]^T$ and (5) becomes

$$\underline{N}\vec{h} = [\vec{\delta}_p; \vec{h}_q]^T$$

Note that the operator

$$\underline{S}\vec{h} = [\vec{0}_p; \vec{h}_q]^T \quad (6)$$

(orthogonally) projects \vec{h} onto the Q dimensional subspace, S , spanned by the unit vectors

$$\vec{e}_k = [\vec{0}_p; \vec{\delta}_k]^T; 1 \leq k \leq Q$$

where the vector $\vec{\delta}_k$ is 1 in its k^{th} position and is otherwise zero. Thus, our operator

$$\underline{N}\vec{h} = [\vec{\delta}_p; \vec{0}_q]^T + \underline{S}\vec{h}$$

projects \vec{h} onto the linear variety, N , which is the translation of S by the vector $[\vec{\delta}_p; \vec{0}_q]^T$.

Algorithm Convergence

As illustrated in Fig. 2, by alternately projecting between the subspace T and linear variety N , one expects convergence to a common point to both [6]. Of principal concern is whether our net's iteration:

$$\vec{s}_{M+1} = \underline{N}\underline{T}\vec{s}_M \quad (7)$$

will converge to $\vec{s} \in F$. A sufficient condition for unique convergence is that

$$P \geq N \quad (8)$$

and the matrix

$$E_p = [\vec{f}_{1p} \ ; \ \vec{f}_{2p} \ ; \ \dots \ ; \ \vec{f}_{Np}] \quad (9)$$

is full rank.

Proof: A fundamental contribution of Youla and Webb [9] states that alternating projections between two (or more) convex sets* converge to a point common to both (all) sets. Since both N (a linear variety) and T (a subspace) are convex, the theorem is applicable here. Furthermore, since both of these sets are linear varieties, convergence is strong [9]. That is, there exists a vector \vec{h} in both sets (i.e., $\vec{h} \in T$ and $\vec{h} \in N$) such that

$$\lim_{M \rightarrow \infty} \| \vec{s}_M - \vec{h} \| = 0$$

Clearly, we would like to have $\vec{h} = \vec{\delta}$. We can be assured of this if T and N intersect only at a single point. Let's explore this notion. If $\vec{h} \in T$, then there exists an N dimensional vector, \vec{a} , such that

$$\vec{h} = \underline{F} \vec{a}$$

Similarly, if $\vec{h} \in N$, then $\vec{h}_p = \vec{\delta}_p$. Any \vec{h} common to both sets must then satisfy

$$\underline{F}_p \vec{a} = \vec{\delta}_p \quad (10)$$

If $P < N$, there are a continuum of solutions. If

$\rightarrow P \geq N$, there is at least one solution. If $\vec{\delta} = \vec{f}_m$, the solution is:

$$\vec{a} = \vec{\delta}_m$$

A sufficient condition for this to be the unique solution is that \underline{F}_p be full rank.

* A set C is convex if $a\vec{a} + (1-a)\vec{b} \in C$ for all $\vec{a}, \vec{b} \in C$ and $0 \leq a \leq 1$.

A more general approach to the question of the degree of subspace intersection, in which our theorem is subsumed, is given by Youla [7-8].

Relaxation Parameters

The speed of convergence of the net iteration can be painfully slow. (Consider, for example, when the angle between T and N in Fig. 1 is very small.) One technique to offset this slow convergence is use of relaxation parameters [9, 14-15]. Specifically, we select two constants, λ_T and λ_N , both of which lie on the interval $[0, 2]$ and redefine the interconnect and node operators by

$$\underline{I}_r = (1 - \lambda_T) \underline{I} + \lambda_T \underline{I}$$

and

$$\underline{N}_r = (1 - \lambda_N) \underline{N} + \lambda_N \underline{N}$$

The autointerconnects are now

$$(t_r)_{kk} = \lambda_T (t_{kk} + 1) - 1$$

and the remaining interconnects become

$$(t_r)_{jk} = \lambda_T t_{jk} ; k \neq j$$

Effects of Input Node Operator Error

Consider the perturbed node operator \underline{N}_e defined by

$$\underline{N}_e \vec{h} = [\vec{\delta}_p + \vec{\Delta}_p : \vec{h}_q]^T$$

where $\vec{\Delta}_p$ is a P dimensional error vector corresponding to faulty library information or processor inexactitude. Define $\vec{\Delta} = [\vec{\Delta}_p : \vec{0}_q]^T$. If $\vec{\Delta} \in T$, then a perturbed fixed point is clearly at $\vec{\delta} + \vec{\Delta}$. Otherwise, we ask whether the linear variety \underline{N}_e intersects T . If it does, then convergence will be to a common point in each set. If not, we can appeal to a result

of Goldberg and Marks [10] who proved that iteration between two non-intersecting finite dimensional convex sets strongly converges to a cycle between two points in each set -- each, a closest point in its set to the other convex set. In either case, the fixed point of iteration is not affected by translation of the linear variety in a direction orthogonal to both sets.

Fault Tolerance

To obtain an empirical feel for the fault tolerance of the extrapolation net, we used $N=5$ orthogonal sampled sine wave vectors of length $L = 40$. Each vector had norm $\|\vec{f}_n\| = \sqrt{20}$. In all cases, we deleted half of a library vector's elements. With only single precision computing error, the mean square error

$$e_M = \|\vec{s}_M - \vec{f}\|^2$$

reduced in 10 iterations from $e_0 = 10.5$ to $e_{10} = 0.3$. Quantizing each element of the I matrix to

seven quantization levels yielded surprisingly similar results. Doubling the quantization interval resulted in divergence.

A number of simulations were performed wherein a percentage of the elements in I were randomly set to zero. Convergence was strongly dependent upon the chosen library vector. Under the scenario above, for example, for 10% of I set to zero, e_{10} typically varied from 0.4 to 0.7. For 20%, $0.7 < e_{10} < 2.8$. A more exhaustive analysis of the fault tolerance is in order.

Tradeoff of Fault Tolerance with Operations per Iteration

The extrapolation net requires L^2 multiplications per iteration. Note, however, that

$$\underline{R} = \underline{F}^T \underline{F}$$

is a non-negative definite (correlation) matrix, and thus its inverse can be written as:

$$\underline{R}^{-1} = \underline{D}^T \underline{\Lambda} \underline{D}$$

where the diagonal matrix $\underline{\Lambda}$ contains the eigenvalues of \underline{R}^{-1} and \underline{D} is the corresponding matrix of eigenfunctions. Therefore, (4) can be written as

$$\underline{I} = \underline{\phi} \underline{\phi}^T$$

where

$$\underline{\phi} = \underline{F} \underline{D}^T \sqrt{\underline{\Lambda}} \quad (11)$$

is an $L \times N$ matrix. As was done by Marks and Atlas [16], one iteration can be performed by first, multiplying \vec{s}_M by $\underline{\phi}^T$ and second, multiplying this vector result by $\underline{\phi}$. Each step costs NL multiplies and, if $N \ll L$, a significant number of multiplies per iteration is saved using this outer product technique at, of course, the loss of fault tolerance and the neural net structure.

TABLE LOOK-UP

An assumption thus far is that any set of P known values in a vector $\vec{s}_E \in F$ can be used to drive the remaining Q nodes. Due to this generality, every node must be connected to every other node. If, on the other hand, the same P nodes are always used as inputs, then the number of interconnects can be reduced. Indeed, the states of the P input nodes are not determined by their inputs. Thus, the interconnects to these nodes can be discarded. As we shall see, such table look-up nets can be reconfigured to $Q < L$ nodes. As with the extrapolation net, the number of operations per iteration can be reduced at the cost of fault tolerance.

A Table Look-Up Net

Again, without loss of generality, assume that the first P elements of some \vec{f} are our input. Since the first P elements of \vec{s}_M and \vec{f} are the same, (1) can be written as:

$$\vec{i}_M = \begin{bmatrix} \vec{i}_{M,p} \\ \vec{i}_{M,q} \end{bmatrix} = \begin{bmatrix} I_2 & | & I_1 \\ \hline I_3 & | & I_4 \end{bmatrix} \begin{bmatrix} \vec{f}_p \\ \vec{s}_{M,q} \end{bmatrix} \quad (12)$$

where we have partitioned the I matrix. For the node operator in (5), we need not be concerned with $\vec{i}_{M,p}$ since the node will transform it to \vec{f}_p . Thus, the I_1 and I_2 partitions have no contribution to the final result. Such "don't care" portions in extrapolation matrices have been noted elsewhere [17]. Setting $\vec{s}_{M+1,q} = \vec{i}_{M,q}$, the informational part of (12) is

$$\begin{aligned} \vec{s}_{M+1,q} &= [I_3 \quad | \quad I_4] \begin{bmatrix} \vec{f}_p \\ \vec{s}_{M,q} \end{bmatrix} \\ &= \vec{g} + I_4 \vec{s}_{M,q} \end{aligned} \quad (13)$$

where

$$\vec{g} = I_3 \vec{f}_p$$

can be computed from the library and the memory address, \vec{f}_p . A net for this operation using Q nodes can be formed akin to that discussed in the Preliminaries section. Our interconnect matrix is I_4

and the node operator is defined by

$$\hat{N} \vec{i} = \vec{i} + \vec{g}$$

If the sufficient criteria in (8) and (9) are applicable, then $\vec{s}_{\infty,q} = \vec{f}_q$ with Q^2 multiplications per iteration. The node used in this net is that in Fig. 1a.

Outer Product Equivalent

The matrix in (11) can be partitioned as:

$$\underline{\phi} = \begin{bmatrix} \underline{\phi}_p \\ \underline{\phi}_q \end{bmatrix}$$

where $\underline{\phi}_p$ contains the first P rows of $\underline{\phi}$ and $\underline{\phi}_q$ the remaining Q . Then

$$\underline{I}_A = \underline{\phi}_q \underline{\phi}_q^T$$

and (13) can be written

$$\vec{s}_{M+1,q} = \vec{g} + \underline{\phi}_q \underline{\phi}_q^T \vec{s}_{M,q}$$

Performing the iteration in this non-net format requires $2NQ$ multiplications per iteration.

FINAL REMARKS

1. A summary of the operations per iteration for each of the four extrapolation techniques are in Table 1.
2. The analysis of the extrapolation net drew strongly from results previously derived for signal synthesis and recovery purposes [7-12, 14-15]. In these cases, the equivalent of a library set was chosen either due to a design or constraint motivation rather than for memory purposes. The celebrated Papoulis-Gerchberg algorithm [7-8, 12, 14, 17-20] (in discrete form), for example, used a similar \underline{N} as ours, but chose as a "library" those vectors whose DFT's were identically zeros in specified bins. The extrapolation net performs this algorithm when the library vectors are the corresponding complimentary rows of the DFT matrix. The continuous form of the Papoulis-Gerchberg algorithm has been performed optically [12, 21-23].
3. We have applied the powerful results of convex set projection in our analysis. Any net with a correspondingly convex \underline{N} can be similarly analyzed. Also, two or more convex operations can be combined at a node. If, for example, we knew that the library vector's elements were between minus and plus one, then the output nodes

could perform an additional convex operation which for $P + 1 \leq k \leq L$, is defined by

$$s_k = \begin{cases} 1 & ; i_k > 1 \\ i_k & ; |i_k| \leq 1 \\ -1 & ; i_k < -1 \end{cases}$$

For $1 \leq k \leq P$, N is as before. One can view this as a projection onto a (convex) hypercube centered at the origin.

4. One advantage of the Hopfield CAM net is that a finite number of iterations can result in the exact correct answer, whereas the extrapolation net generally only gets iteratively closer and closer. A step towards a multilevel net, however, can be obtained from the extrapolation net by requiring each library vector to contain only integers. In lieu of (7), we perform the iteration

$$\vec{s}_{M+1} = \underline{I} \underline{N} \underline{I} \vec{s}_M \quad (14)$$

where the vector operator \underline{I} rounds each vector element to the nearest integer. Geometrically, \underline{I} projects onto the nearest vector with all integer components. Although (14) generally converges in a finite number of iterations and gets us "close" to the desired library element, convergence can be to an element not contained in our library. Consider, for example, Fig. 3 where, as in Fig. 2, the subspaces T and N are shown. The lattice of dots denote vectors with integer components. Beginning with the \vec{s}_0 in the lower right corner, in accordance to (14), we project onto N and then onto T and finally onto the nearest lattice point. Continuing, we eventually converge to \vec{s}_∞ shown as the vertex of the steady state $(\vec{s}_\infty, \vec{b}, \vec{c})$ triangle in

Fig. 3. Although the process has converged in a finite number of iterations, the result is not our desired \vec{f} . Note similar steady state triangles (e.g., t in Fig. 3) exist closer to \vec{f} .

ACKNOWLEDGEMENTS

The author greatly acknowledges the support of this work by the SDIO/IST'S Ultra High Speed Computing Program administered through the U.S. Office of Naval Research in conjunction with the Optical Systems Lab at Texas Tech University and, in part, by the Boeing High Technology Center. Significant contributions to the clarification of result interpretation were made by Dziem Nguyen and Fred Holt at the Boeing High Technology Center. Also appreciated are the stimulating discussions with the author's ISDL colleagues: Les Atlas, Kwan Cheung and Jim Ritcey.

References

1. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities" Proc. Natl. Acad. Sci. USA 79 pp 2554-2558 (1982).
2. D. Psaltis and N. Farhat "Optical Information Processing Based on an Associative-Memory Model of Neural Nets with Thresholding and Feedback" Optics Letters 10 pp 98-100 (1985).
3. N. Farhat, D. Psaltis, A. Prata and E. Paek, "Optical Implementation of the Hopfield Model" Applied Optics 24 pp 1469-1475 (1985).
4. D. W. Tank and J. J. Hopfield, "Simple Neural" Optimization Networks: An A/D converter, Signal Decision Circuit and a Linear Programming Circuit," IEEE Trans Circuits and Systems, CAS-33, pp 533-541 (1986).
5. A.D. Fisher and C.L. Giles "Optical Adaptive Associative Computer Architectures", Proceedings of the IEEE 1985 Compcon Spring, IEEE Computer Society Press, pp.342-344.
6. C. Brown, "Hopfield's Neural Nets Realize Biocomputing," Electronic Engineering Times, (April 7, 1986).
7. D. C. Youla, "Generalized image restoration by method of alternating orthogonal projections," IEEE Trans. Circuits and Systems, CAS-25, pp 694-702 (1978).
8. H. Stark, D. Cahana and H. Webb, "Restoration of an Arbitrary Finite-Energy Optical Objects from Limited Spatial and Spectral Information," J. Opt. Soc. Am. 71, pp. 635-642 (1981).
9. D. C. Youla and H. Webb, "Image Restoration by the Method of Convex Projections: Part 1-Theory" IEEE Trans. Med. Imaging MI-1, pp 81-94 (1982).
10. M. Goldberg and R. J. Marks II, "Signal Synthesis in the Presence of an Inconsistent Set of Constraints" IEEE Trans. Circuits and Systems CAS-32, pp 647-663 (1985).

References

1. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities" Proc. Natl. Acad. Sci. USA 79 pp 2554-2558 (1982).
2. D. Psaltis and N. Farhat "Optical Information Processing Based on an Associative-Memory Model of Neural Nets with Thresholding and Feedback" Optica Letters 10 pp 98-100 (1985).
3. N. Farhat, D. Psaltis, A. Prata and E. Paek, "Optical Implementation of the Hopfield Model" Applied Optics 24 pp 1469-1475 (1985).
4. D. W. Tank and J. J. Hopfield, "Simple Neural Optimization Networks: An A/D converter, Signal Decision Circuit and a Linear Programming Circuit," IEEE Trans. Circuits and Systems, CAS-33, pp 533-541 (1986).
5. A.D. Fisher and C.L. Giles "Optical Adaptive Associative Computer Architectures", Proceedings of the IEEE 1985 Comcon Spring, IEEE Computer Society Press, pp.342-344.
6. C. Brown, "Hopfield's Neural Nets Realize Biocomputing," Electronic Engineering Times, (April 7, 1986).
7. D. C. Youla, "Generalized image restoration by method of alternating orthogonal projections," IEEE Trans. Circuits and Systems, CAS-25, pp 694-702 (1978).
8. H. Stark, D. Cahana and H. Webb, "Restoration of an Arbitrary Finite-Energy Optical Objects from Limited Spatial and Spectral Information," J. Opt. Soc. Am. 71, pp. 635-642 (1981).
9. D. C. Youla and H. Webb, "Image Restoration by the Method of Convex Projections: Part 1-Theory" IEEE Trans. Med. Imaging MI-1, pp 81-94 (1982).
10. M. Goldberg and R. J. Marks II, "Signal Synthesis in the Presence of an Inconsistent Set of Constraints" IEEE Trans. Circuits and Systems CAS-32, pp 647-663 (1985).

11. D. C. Youla and V. Velasco, "Extensions of a Result on the Synthesis of Signals in the Presence of Inconsistent Constraints" IEEE Trans. Circuits and Systems CAS-33 pp 465-468 (1986).
12. R. J. Marks II and D. K. Smith, "Gerchberg-type linear deconvolution and extrapolation algorithms," Transformations in Optical Signal Processing, Proc. SPIE, vol. 373, pp. 161-178, 1981.
13. Gilbert Strang, Linear Algebra and Its Applications, 2nd ed., (Academic Press, NY, 1980) p. 116.
14. R. W. Schafer, R. M. Mersereau and M. A. Richards, "Constrained iterative restoration algorithms," Proc. IEEE 69, 432-450 (1981).
15. J. R. Fienup, "Reconstruction and Synthesis Applications of an Iterative Algorithm," Transformations in Optical Signal Processing, Proc. SPIE, V. 373 1981, pp. 147-160.
16. R. J. Marks II and L. E. Atlas, "Content Addressable Memories: A Relation between Hopfield's Neural Net and an Iterative Matched Filter." Submitted for publication.
17. D. Kaplan and R. J. Marks II, "Noise Sensitivity of Interpolation and Extrapolation Matrices," Applied Optics, 21, pp. 4489-4492 (1982).
18. R. W. Gerchberg, "Super-Resolution Through Error Energy Reduction," Optica Acta, 21, pp. 709-720 (1974).
19. A. Papoulis, "A New Algorithm in Spectral Analysis and Bandlimited Signal Extrapolation," IEEE Trans. Circuits and Systems, CAS-22, pp. 735-742 (1975).
20. A. Papoulis, Signal Analysis, (McGraw-Hill, New York, 1977) pp. 234-251.

21. T. Sato, S. J. Norton, M. Linzer, O. Ikeda, and M. Hirama, "Tomographic image reconstruction from limited projections using iterative revisions in image and transform spaces." Appl. Opt. 20, 395-399 (1981).
22. R. J. Marks II, "Coherent optical extrapolation of 2-D band-limited signals: processor theory," Appl. Opt. 19, 1670-1672 (1980).
- 23 . R. J. Marks II and D. K. Smith, "An iterative coherent processor for bandlimited signal extrapolation," Proc. SPIE 231, 106-111 (1980).

Figure and Table Captions

Figure 1: The three types of nodes used in this paper. The input into the k^{th} node, i_k , is the sum of the contributions of all L nodes through transmittances t_{ik} . (a) A node useful for linear equation solution and table look-up nets. (b) The node used in Hopfield CAM nets. (c) Nodes useful for our extrapolation net.

Figure 2: Illustration of the iterative convergence to the library vector. Beginning with $\vec{s}_0 = [\vec{s}_p; \vec{0}_q]$, we alternately orthogonally project between T and N as shown with the dashed lines. Note that \vec{s}_0 is orthogonal to the subspace.

Figure 3: When rounding the states to the nearest integer, the iteration converges in a finite number of steps -- but not to the desired integer vector, \vec{f} .

Table 1: Multiplies per iteration for four memories. For each, there are N library vectors of length L . P elements of one of these elements are used to regenerate the remaining $Q = L - P$. Each memory scheme executes the same restoration algorithm. Thus, in the absence of processor inexactitude, all perform identically.



The Washington Technology Center

376 Loew Hall, FH-10, University of Washington, Seattle, WA 98195

Office of the Executive Director
(206) 545-1920

October 13, 1986

TO: WTC/UW Principal Investigators

FROM: Edwin B. Stear *EBS*
Executive Director

SUBJECT: Technology Disclosures

This memorandum, along with the enclosed materials, is intended to provide specific guidance on the handling of technology disclosures through the WTC, as well as clarify The Washington Technology Center's Patent and Copyright Policy in general.

As you know, President Gerberding in October 1985 signed Administrative Order No. 17 which exempted the WTC from UW patent and copyright policies and delegated authority to the WTC to have and administer its own Patent and Copyright Policy subject to certain conditions (see the enclosed copy). Subsequently, the WTC Board of Directors approved a WTC Patent and Copyright Policy. Although a copy of this policy was distributed to you some months ago, it is included here to provide a self-contained information packet.

To provide further background, I am enclosing copies of the WTC Principles governing patent and copyright policies and procedures, and the Agreement between The Washington Technology Center and the Washington Research Foundation (WRF).

Finally, in accordance with the documents identified above, the enclosed technology disclosure policy is provided for your information and use in disclosing inventions related to WTC research projects. As noted in the instructions, the disclosure will generally be forwarded to the Washington Research Foundation (or other agent), at the discretion of the WTC, for evaluation of patents and commercial potential.

Please feel free to contact me if you have any questions concerning these policies or procedures.

EBS/bf

Enclosures

cc: John Rusin
Janell Douglas
John Piety

INVENTION DISCLOSURE

Washington Technology Center

Instructions

This Invention Disclosure Form is used to report inventions and to record the circumstances under which the invention was made. The Disclosure is a legally important document; care should be taken in its preparation since it provides both the basis for determining patentability and the data for drafting a patent application.

New and potentially useful technology developed by WTC employees with WTC and/or industry grant and contract support should be reported promptly consistent with the Center's Patent and Invention Policy.

The following instructions apply to the correspondingly numbered sections of the form.

1. Use a brief title, sufficiently descriptive to aid in identifying the invention.
2. Provide a brief description, pointing out novel features of the invention. Attach additional material which covers the following points:
 - a. General purpose
 - b. Technical description with references to drawings, schematics, sketches, flow diagrams, etc., as appropriate
 - c. Advantages and improvements over existing methods, devices or materials, and features believed to be new
 - d. Possible variations and modifications
 - e. State-of-the-art prior to invention, and similar or related patents (if known)
3. List all sources of support for the research which led to the conception or actual reduction to practice of the invention. Include WTC personnel, funds or materials as well as those of University or outside agencies, organizations and companies.
4. The invention history is legally important in determining the priority of invention and/or legal "bars" to patenting. The United States Patent law allows submission of a patent application up to one year after an enabling disclosure of the technology. Most foreign countries require a patent application prior to any enabling disclosure (an oral presentation or publication such as an article, abstract or theses, or other communication which would allow a knowledgeable person to duplicate the work).

University of Washington Correspondence

INTERDEPARTMENTAL

ELECTRICAL ENGINEERING, FT-10

DATE: February 26, 1987

TO: Professors L. Atlas, J. Ritcey and A. Somani

FROM: Bob Marks *BJM*

As you can see by the attached memo, I have sent our work to the right channels. If you have an additional paper, please send it to Fleming with a cover memo and a courtesy copy to Graham, Dziem and me.

Thanks!

University of Washington Correspondence

INTERDEPARTMENTAL

ELECTRICAL ENGINEERING, FT-10

DATE: February 26, 1987

TO: Lynn Fleming

FROM: Robert J. Marks II *RJM*

SUBJECT: Publications corresponding to "Analysis of Neural Nets" sponsored by the Boeing HTC.

Enclosed are four papers generated totally or in part under the support of the subject grant:

1. "Compact Neural Network: Implementation in Regular Structures", is scheduled to be presented at the IEEE First Annual Conferences on Neural Nets in San Diego, June 21-24, 1987.
2. Report on "One Step Convergence of Hopfield's Neural Net CAM", by Lawrence Wong, is a senior project final report prepared in the fall of 1986.
3. "A Class of Continuous Level Associative Memory Neural Nets", by R.J. Marks II, is scheduled to appear in Applied Optics.
4. "Optical Architectures for a Continuous Level Neural Net", contains material which is to be given at the IEEE Conference on Neural Information Processing Systems-Natural and Synthetic, this year at Boulder, Colorado.

Please contact me at 543-6990 if you have any questions.

cc: R. Graham
D. Nguyen

OPTICAL ARCHITECTURES FOR A CONTINUOUS LEVEL NEURAL NET

Robert J. Marks II
ISDL
2/20/87

INTRODUCTION

We propose optical processing architectures for implementing a recently proposed class of continuous level neural net (CLNN) associative memories [1]. As with other optical neural net architectures, the processors perform iteratively. They have the advantage, however, of requiring no electronic or phase conjugating optics in the feedback path. Thus, the neural net's stable states are iteratively generated at light speed. Furthermore, the processor components are all commercially available off-the-shelf items.

PRELIMINARIES

For purposes of continuity and establishing notation, we briefly review the CLNN. A more complete discussion can be found in Ref. [1].

In a system of L neurons, we store a total of N continuous level vectors $\{f_n | 0 \leq n \leq N\}$. Define the library matrix

$$E = [f_1 | f_2 | \dots | f_N]$$

and the interconnect matrix

$$I = E (E^T E)^{-1} E^T$$

Thus, t_{ij} is the interconnect value between the i^{th} and the j^{th} neuron.

Assume that P neural states are known for some library vector f . In general, any P of the neural states can be known. For notational

convenience and without loss of generality, assume that the first P states of f are known. Accordingly, we adopt the following partitioning notation:

$$f = [f_p \mid f_a]^T$$

where f_p is the vector of the first P elements of f and f_a is the remaining $Q = L - P$. The operation performed at the neural nodes can now be expressed as:

$$\underline{\eta} i = \underline{\eta} [i_p \mid i_a]^T = [f_p \mid i_a]^T$$

In synchronous form, the neural net iteratively performs the operation:

$$s_{M+1} = \underline{\eta} \underline{I} s_M \quad (1)$$

where s_M is the L vector of neural states at time M . Thus, if a state is known, the corresponding node clamps to the known value. The remaining nodes, responding to this stimuli, have floating states that are equal to the sum of their inputs. Convergence to f is guaranteed if $P < N$ and the first P rows of \underline{E} form a matrix of full rank. This is true independent of the choice of the initial state vector, s_0 .

Two seeming disadvantages of the CLNN with respect to Hopfield's are:

are:

- (1) the relative inexactness of analog processor results and
- (2) the generally infinite number of required iterations for convergence.

With regard to optical implementation, the responses to these objections, respectively, are

- (1) As a function of the accuracy and dynamic range of the input and processing, the input library vectors can be restricted to a given number of discrete levels. Then, corresponding to some performance level, the processor output can be quantized accordingly.
- (2) When iterations are being performed at light speed, the significance the convergence rate is reduced substantially.

A TABLE LOOK-UP NET

A table look-up net is one in which the same P nodes are always used as the net's stimulus and the remaining Q nodes iteratively converge to the desired response. Note that the iteration in (1) can be partitioned as:

$$\begin{bmatrix} f_P \\ \text{---} \\ S_{Q, M+1} \end{bmatrix} = \begin{bmatrix} I_P \\ \text{---} \\ I_Q \end{bmatrix} \begin{bmatrix} f_P \\ \text{---} \\ S_{Q, M} \end{bmatrix}$$

where I_P denotes the first P rows of I and I_Q is the remaining Q . Since the first P neural states are always clamped to the known values, there is no need to know I_P . Indeed, an equivalent expression is:

$$s_{\alpha, M+1} = I_{\alpha} \begin{bmatrix} f_{\alpha} \\ \text{---} \\ s_{\alpha, M} \end{bmatrix} \quad (2)$$

A basic methodology for optical implementation of this iteration is illustrated in Fig. 1. The known portion of the library vector, f_{α} , is input into the processor by an intensity modulated point source array (e.g. LED's). Multiplication by the I_{α} matrix is performed by a standard vector-matrix multiplication architecture [2]. (The astigmatic optics are not shown). The vector output, $s_{\alpha, M+1}$, is input into a the fiber bundle shown on the right. The bundle is then fed back into the input on the left hand side. This provides the $s_{\alpha, N}$ portion of the input vector required in (2). We are thus performing the iteration required by the table look-up net at light speed. Feedback could also be provided by mirrors.

The astute reader will have already noted three major problems with this processor:

- (1) There is no provision to detect the output.
- (2) There is no provision for compensating for absorbtive and other losses in the feedback loop.
- (3) The I_{α} matrix and the input generally contain both positive and negative numbers. Incoherent optics can only add and multiply positive numbers.

Each of these problems has a straightforward solution:

- (1) The output can be detected by placing a highly transmitting pellicle in the feedback path and using appropriate focusing optics. This clearly increases absorbtive losses and contributes further to problem number two:
- (2) If the matrix transmittance can be amplified, then we can compensate for absorbtive loss. One can easily show that if $N \ll L$, then $t_{1,2} \ll 1$. In such senerios, we can then "amplify" the matrix transmittance significantly and still not exceed the maximum passive transmittance value of unity.
- (3) The problem of performing bipolar operations with incoherent optics has a number of solutions. One straightforward technique is to rewrite each matrix and vector as the sum of a positive and negative matrix or vector:

$$f_p = f_p^+ + f_p^-$$

$$s_{a,m} = s_{a,m}^+ + s_{a,m}^-$$

$$I_a = I_a^+ + I_a^-$$

The matrix I_a^+ , for example, is formed by setting all of the negative elements in I_a to zero. Then (2) can be written as:

$$s_{a,m+1}^+ = I_a^+ \begin{bmatrix} f_p^+ \\ \dots \\ s_{a,m}^+ \end{bmatrix} + I_a^- \begin{bmatrix} f_p^- \\ \dots \\ s_{a,m}^- \end{bmatrix}$$

and

$$s_{\alpha, m+1}^- = I_{\rho}^+ \begin{bmatrix} f_{\rho}^- \\ \dots \\ s_{\alpha, m}^- \end{bmatrix} + I_{\rho}^- \begin{bmatrix} f_{\rho}^+ \\ \dots \\ s_{\alpha, m}^+ \end{bmatrix}$$

The corresponding optical implementation, although somewhat more involved, requires only positive multiplications and additions and is a straightforward generalization of the architecture in Fig.1. The positive and negative components are added electronically at the output.

AN OPTICAL IMPLEMENTATION OF THE CLNN

We now address optical implementation of the CLNN under the condition that any P neurons can act as the net stimulus. An architecture similar to that for the table look-up net is shown in Fig.2 for $L = 4$ neurons. In the figure, the middle two neural states are known and are input into the net structure by the middle two sources in the point source array. The bottom four by four transmittance represents the \underline{I} matrix. Multiplication is performed as before and the output is fed into the fibers on the right. The fiber bundle is positioned so that its other end provides the input to the net in the upper left hand corner. Since all of the input from the middle two neurons should come from only the corresponding sources, the light from the middle two fibers should not be reintroduced into the system. This can be done with either an electro-optic or optic-optic toggle switch that turns off the fibers corresponding to the locations of the neurons (sources) with known states. Such switches can operate in the gigahertz range with small attenuation

[3].

As is shown in Fig.2, the output of the switch is input into the net and multiplies the top four by four transmittance which also corresponds to the to the I matrix. This top transmittance, however, is adjusted for feedback losses as previously discussed. The contribution from the known states (sources) and the unknown states (switch outputs) are thus multiplied by their respective I matrices and the superposition of their contributions are collected by the fiber bundle on the right. The iteration therefore proceeds towards convergence at light speed. The processor can be straightforwardly augmented as before to allow for the required bipolar operations.

CONCLUSIONS

Using the **continuous level neural net (CLNN)** algorithm developed in [1], we have proposed two corresponding optical implementations that require no electronics or phase conjugation optics in the feedback path. After a more detailed feasibility study, we propose to prototype these architectures and investigate their ultimate performance.

REFERENCES

1. R.J. Marks II "A class of continuous level neural net associative memories" to appear in Applied Optics.
2. J.W. Goodman et.al. "Fully parallel, high-speed incoherent Optical method for performing discrete Fourier transforms" Optics Letters, vol.2, pp1-3 (1978).

3. For example, see H. Haga et.al. "LiNbO3 traveling-wave light modulator/switch with an etched groove" IEEE Journal of Quantum Electronics QE-22, 902-906 (1986).

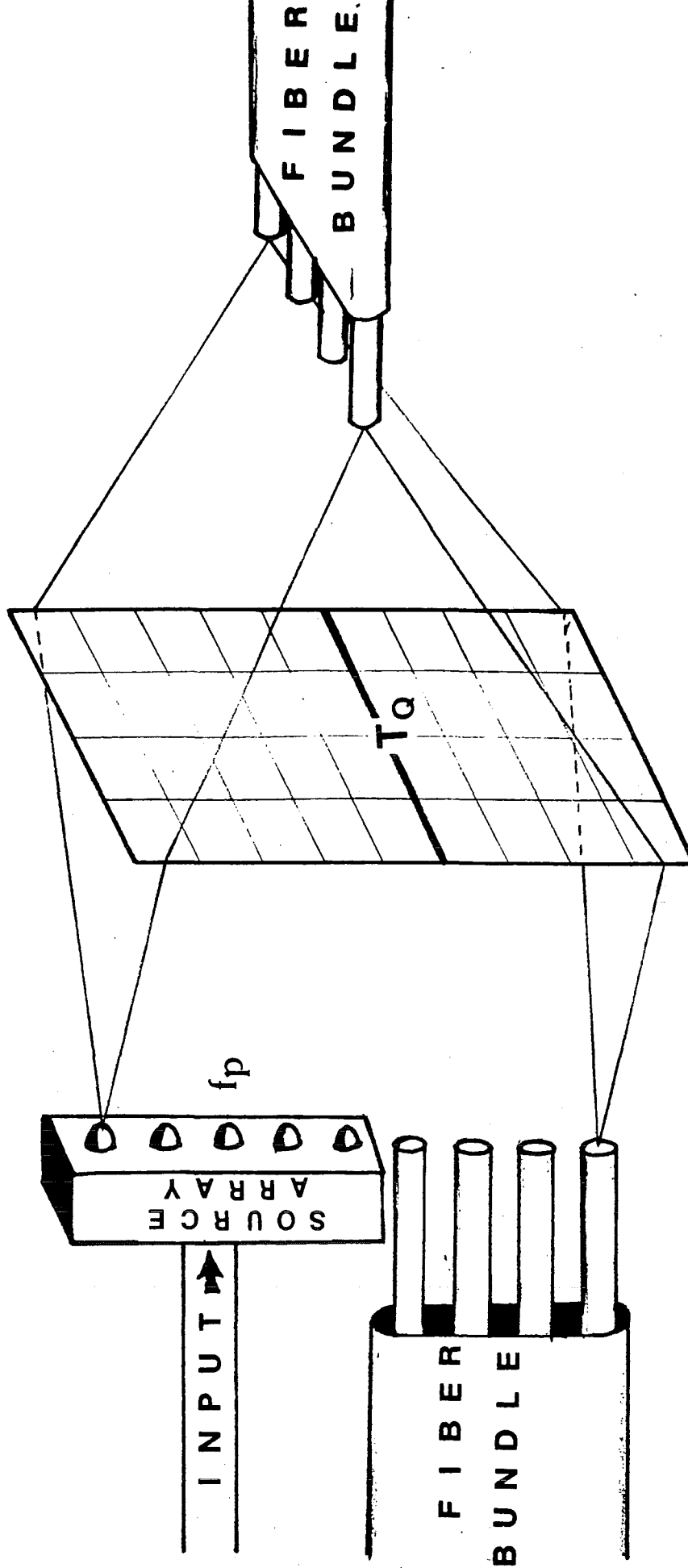


FIG 1

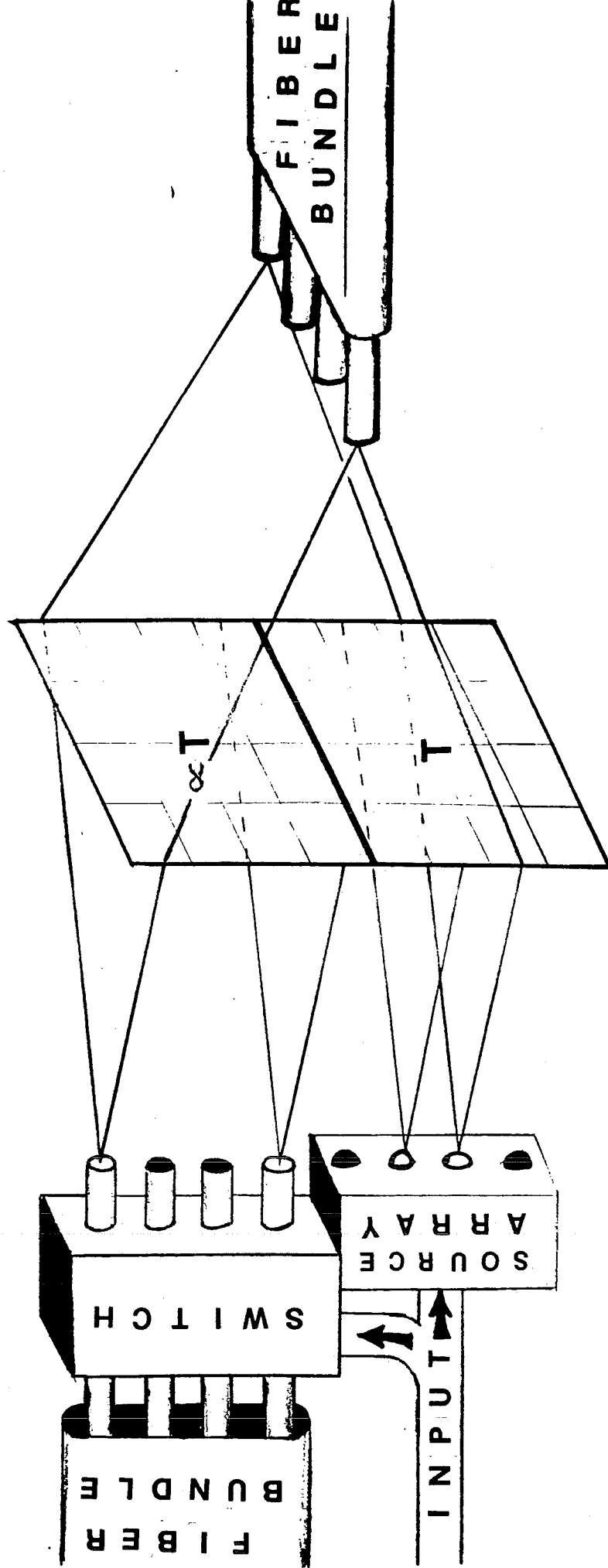


FIG 2

ISOL SEMINAR

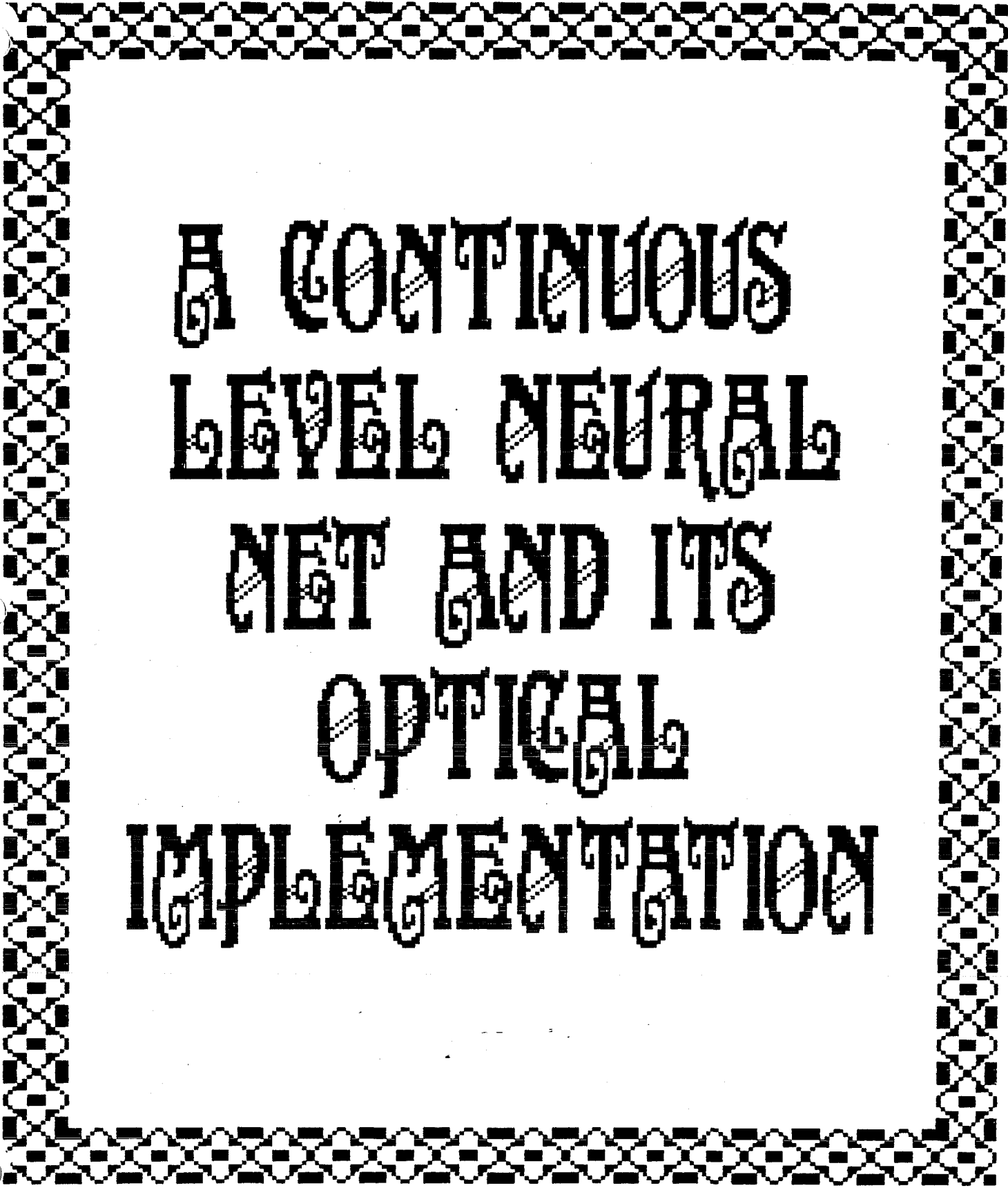
" A Continuous Level Neural Net and its Optical Implementation "

Prof. Robert J. Marks II

Tues, Feb. 10, 2:30 in Rm 108 EEB, U of W.

For more information on the seminar series, please contact :

Prof. Robert Marks
Interactive System Design Laboratory
Dept. of Electrical Engineering, FT-10
University of Washington
Seattle, WA 98195
(206) 543-6990



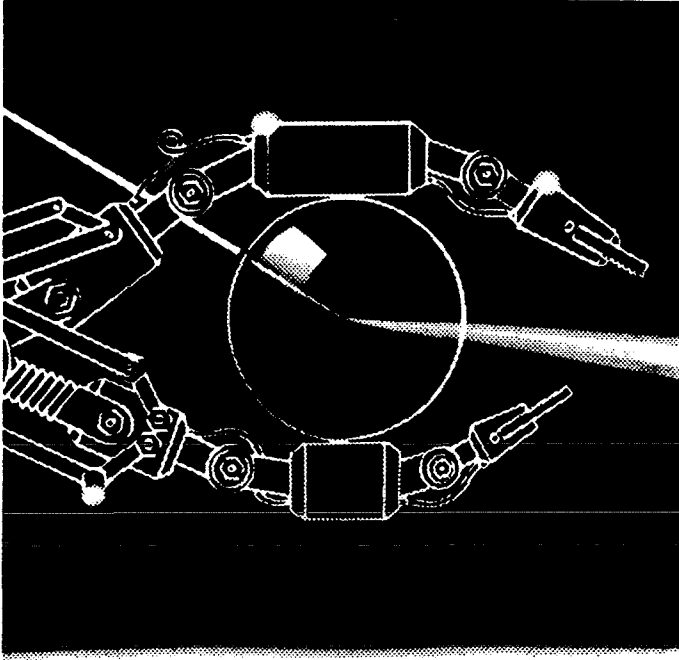
**A CONTINUOUS
LEVEL NEURAL
NET AND ITS
OPTICAL
IMPLEMENTATION**

Contents:

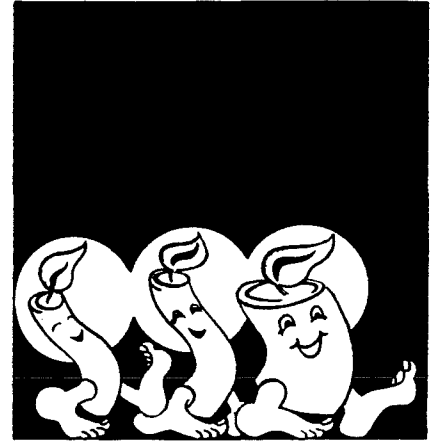
1. Introduction to CAM's
2. A Homogeneous Neural Net
3. The CLNN
4. Optical Implementation
5. Conclusions

Associative Memory

4 Memory Objects:



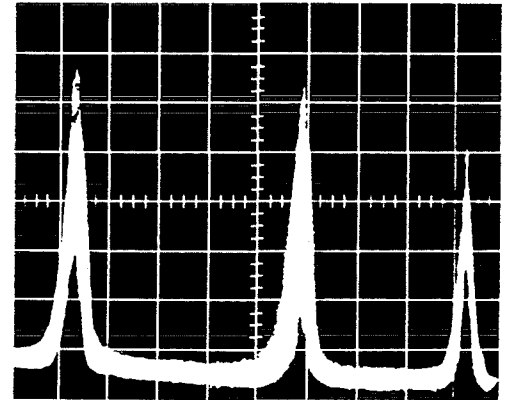
1. Robot Hand



2. Walking Candles

ISOL SEMINAR

3. Letters



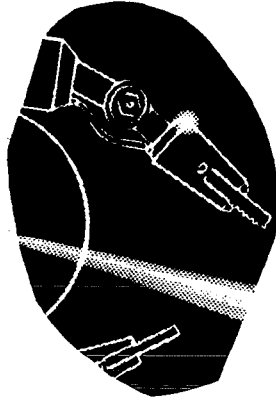
4. Scope Trace

Properties of CAM:

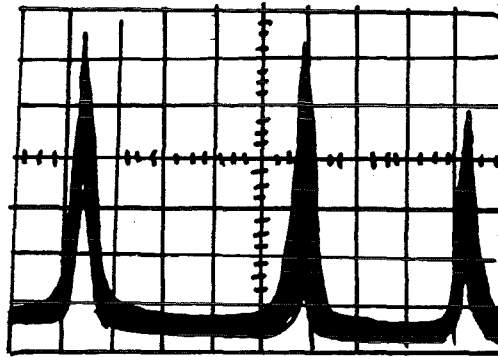
1. Recall from partial memory:



2. Works better with more information:



3. Recognize Perturbed objects

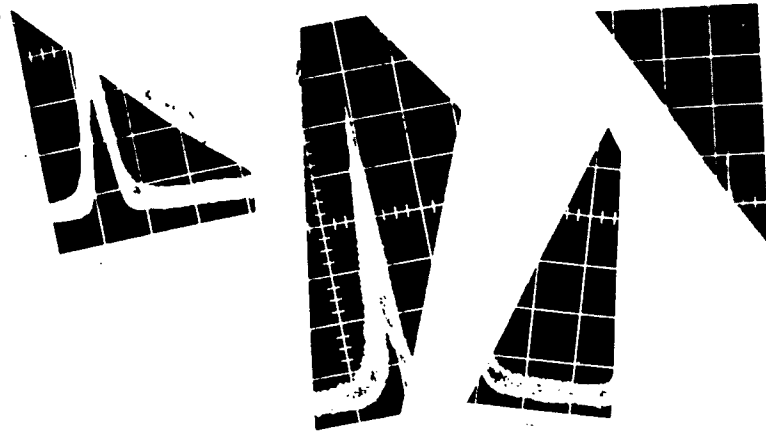


4. Error Correction

ISOL SENWAA

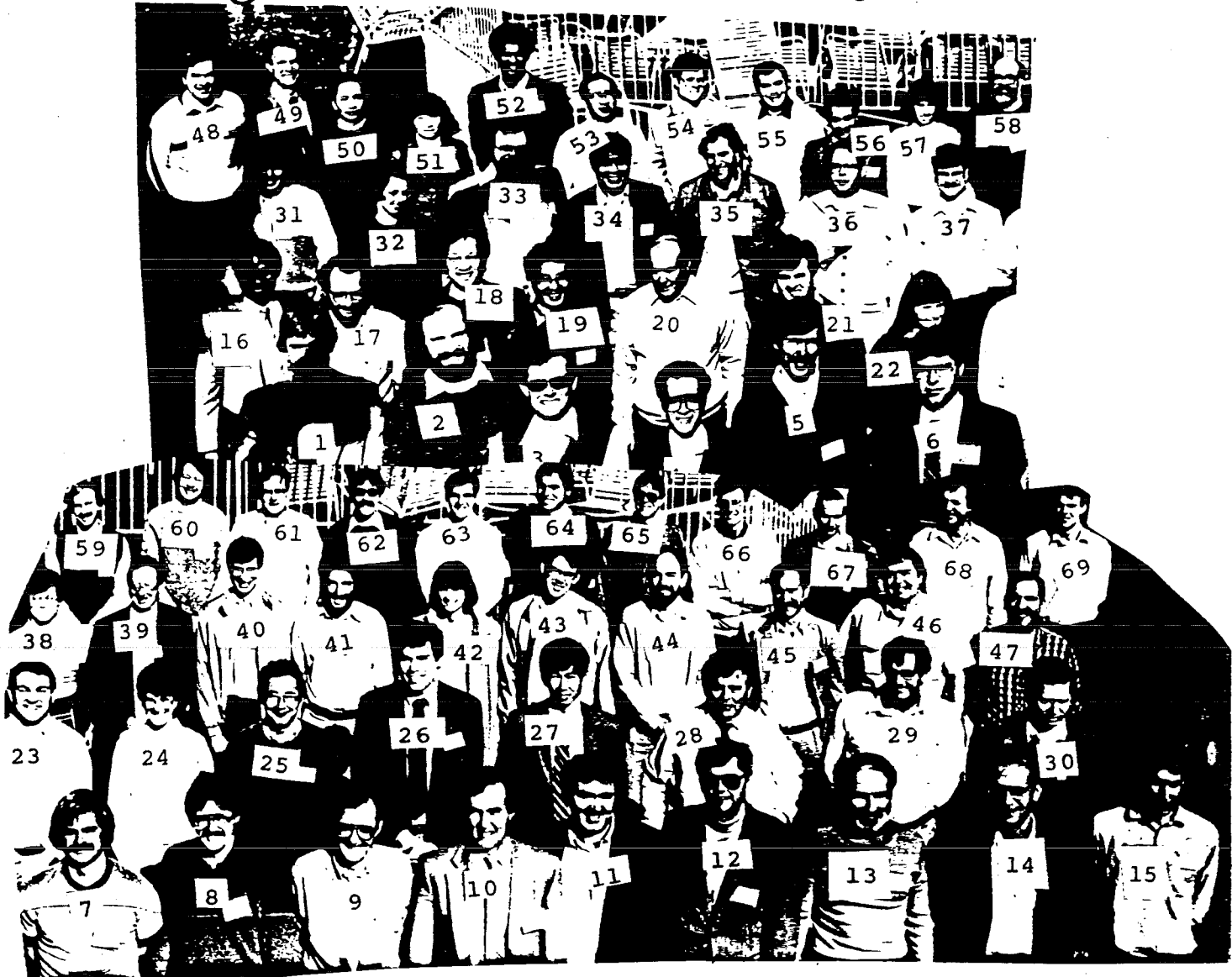


5. Fault Tolerance

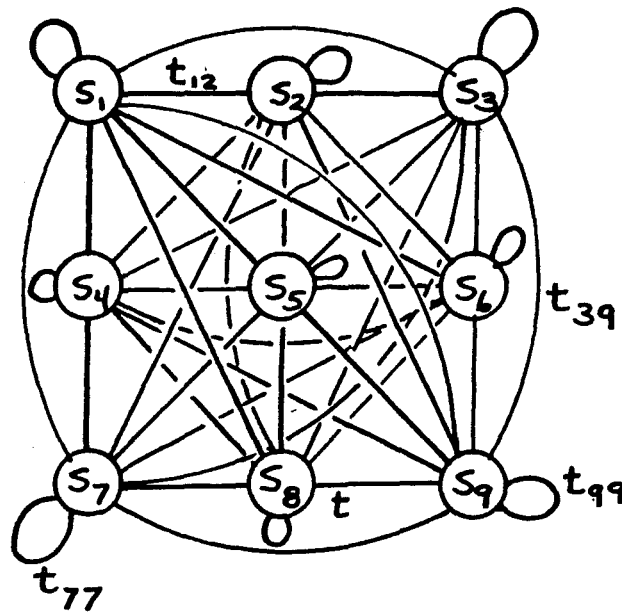


6-7. ● Works better for small libraries

● Recognize Uncorrelated Objects Better



A Homogeneous Neural Net



L neurons.

S_k = state of k^{th} neuron

\vec{S} = L vector of neural states

Interconnects: t_{ij}

i_k = sum of inputs into k^{th} neuron

$$= \sum_{i=1}^L t_{ik} S_i$$

\mathcal{N}_k = operator at k^{th} node

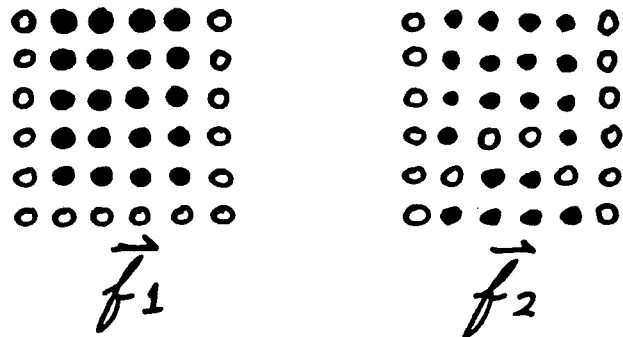
$$\begin{aligned} \text{Iteration: } S_{k+1} &= \mathcal{N}_k i_k \\ &= \mathcal{N}_k \sum_i t_{ik} S_i \end{aligned}$$

Synchronous form:

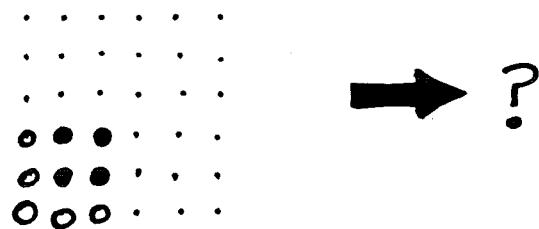
$$\vec{S}_{k+1} = \underline{\mathcal{N}} \underline{\mathbf{T}} \vec{S}_k$$

Application to CAM's

Idea: Memories



Program memory into interconnects



Two Methods:

1. Hopfield

2. CLNN (Continuous Level Neural Net)

The CLNN

Library matrix:

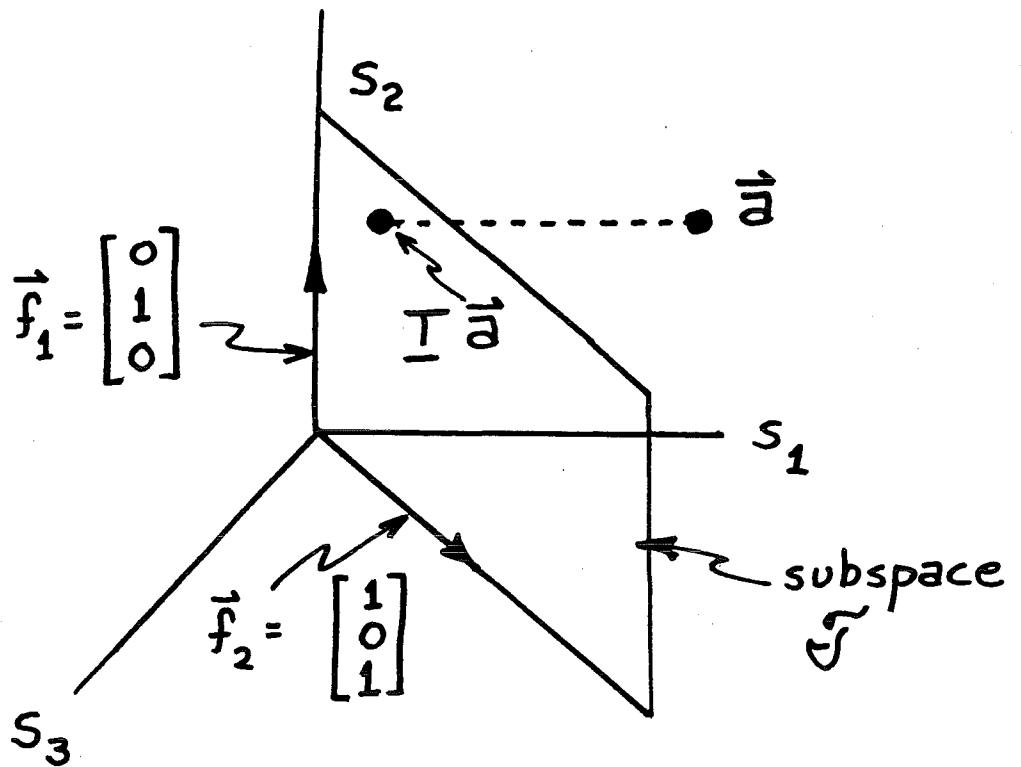
$$\underline{F} = [\vec{f}_1 : \vec{f}_2 : \dots : \vec{f}_N]^T$$

Interconnect Matrix:

$$\underline{I} = \underline{F} (\underline{F}^T \underline{F})^{-1} \underline{F}^T$$

Q: Why?

A: \underline{I} projects onto the column space of \underline{F} :



Neural Operator for CLNN

Let $\vec{f} \in \text{library}$

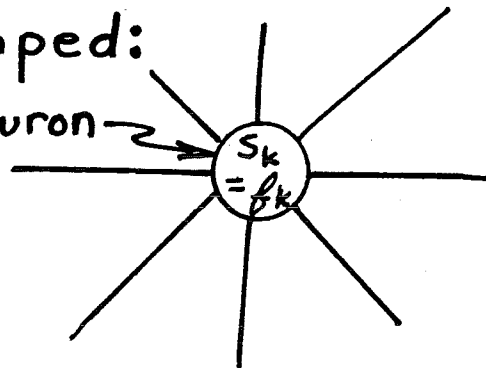
We know P of the elements of \vec{f} and wish to recall the remaining $Q = L - P$.
WLOG, let the first P state be known. Then, for any vector i :

$$\underline{n} \vec{i} = \underline{n} \begin{bmatrix} \vec{i}_P \\ \vdots \\ \vec{i}_Q \end{bmatrix} = \begin{bmatrix} \vec{f}_P \\ \vdots \\ \vec{i}_Q \end{bmatrix}$$

\therefore Two types of neural operators

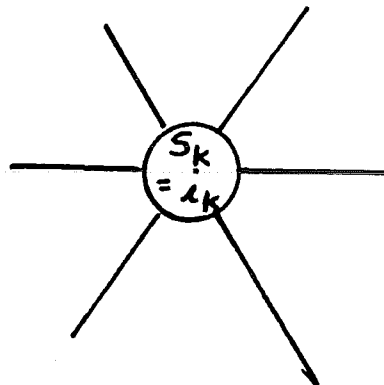
1. Clamped:

k^{th} neuron \rightarrow



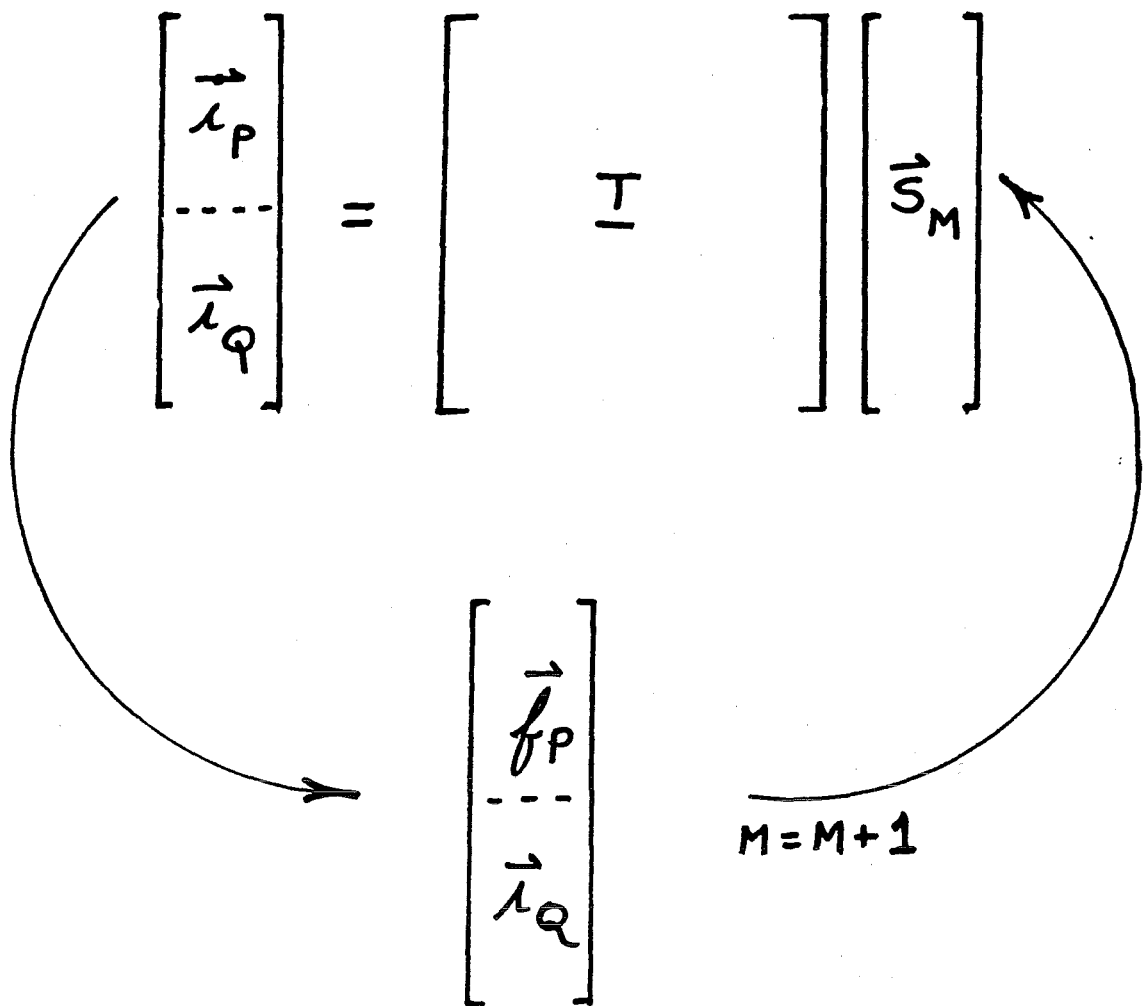
state is known
to be f_k

2. Floating:



state is unknown.
State = i_k
= sum of inputs

Synchronous Interpretation:



Q: Does $\vec{s}_M \xrightarrow{M \rightarrow \infty} \vec{f}$?

A: Usually

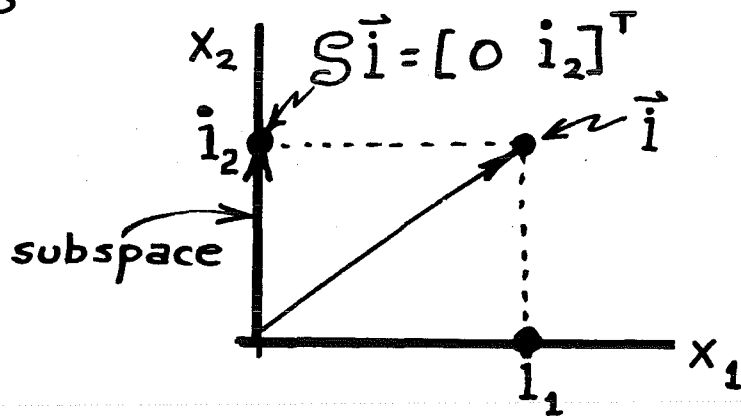
Signal Space Interpretation of the Neural Operator, \mathcal{N}

• Recall: $\mathcal{N} \vec{i} = \begin{bmatrix} \vec{f}_P \\ \vec{i}_Q \end{bmatrix}$

• Consider: $\mathcal{S} \vec{i} = \begin{bmatrix} \vec{0}_P \\ \vec{i}_Q \end{bmatrix}$

\mathcal{S} projects \vec{i} onto a Q dimensional subspace.

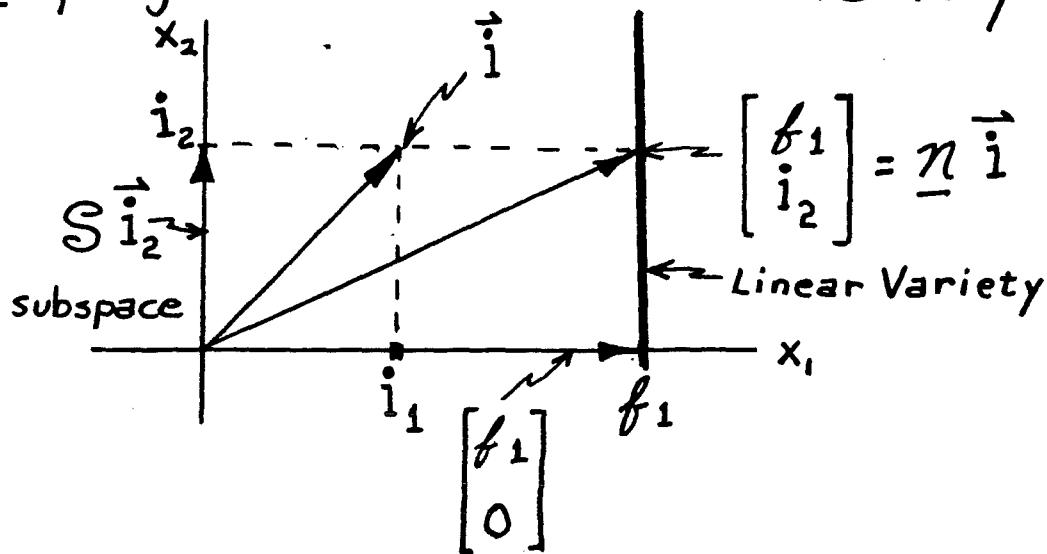
e.g. $\vec{i} = [i_1, i_2]^T$:



• Note:

$$\underline{n} \vec{i} = \begin{bmatrix} \vec{f}_P \\ \vec{0}_Q \end{bmatrix} + S \vec{i} = \begin{bmatrix} \vec{f}_P \\ \vec{0}_Q \end{bmatrix}$$

Thus, \underline{n} projects onto a linear variety:
e.g.



The linear variety is the subspace translated by the (orthogonal) vector $\begin{bmatrix} \vec{f}_P \\ \vec{0}_Q \end{bmatrix}$

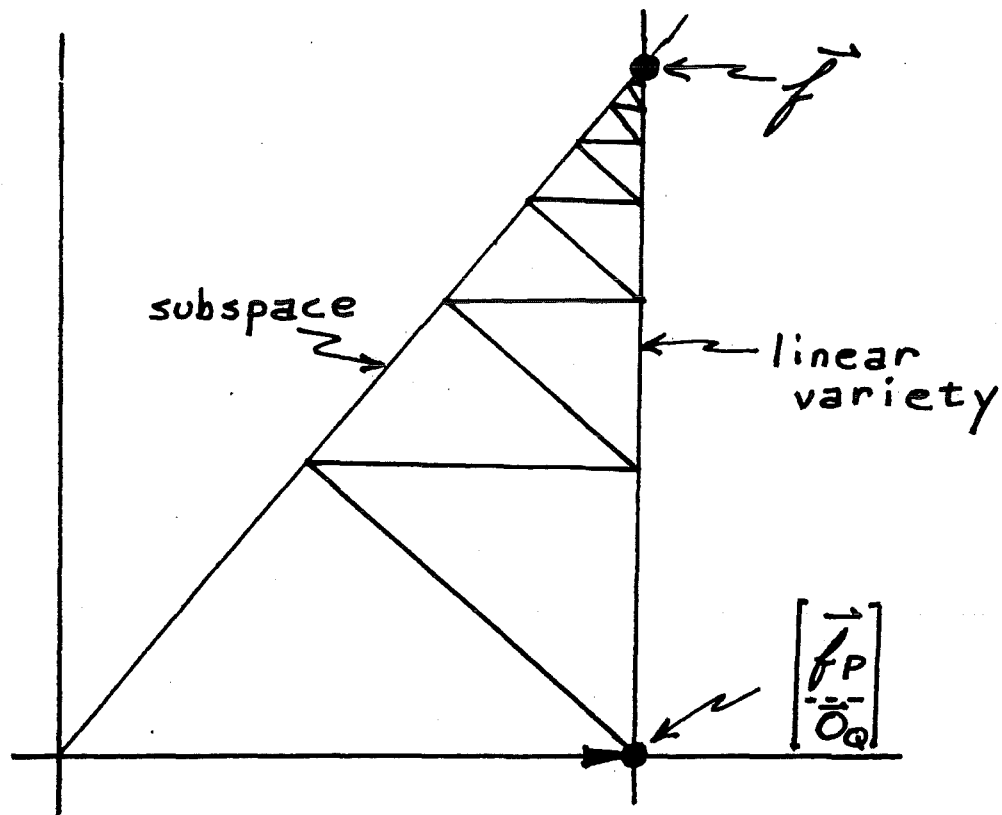
Our Story So Far:

- Library matrix: $\underline{F} = [\vec{f}_1 : \vec{f}_2 : \dots : \vec{f}_N]$
- Interconnect Matrix: $\underline{I} = \underline{F} (\underline{F}^T \underline{F})^{-1} \underline{F}^T$
(projects onto the space spanned by the library).
- Let $\vec{f} \in$ library. The neural operator
$$\underline{n} = \underline{n} \vec{i} = [\vec{f}_P : \vec{i}_Q]^T$$

projects onto the linear variety
- Our neural net performs the operation:

$$\vec{s}_{M+1} = \underline{n} \underline{I} \vec{s}_M$$

\vec{f} is in both the subspace & the linear variety

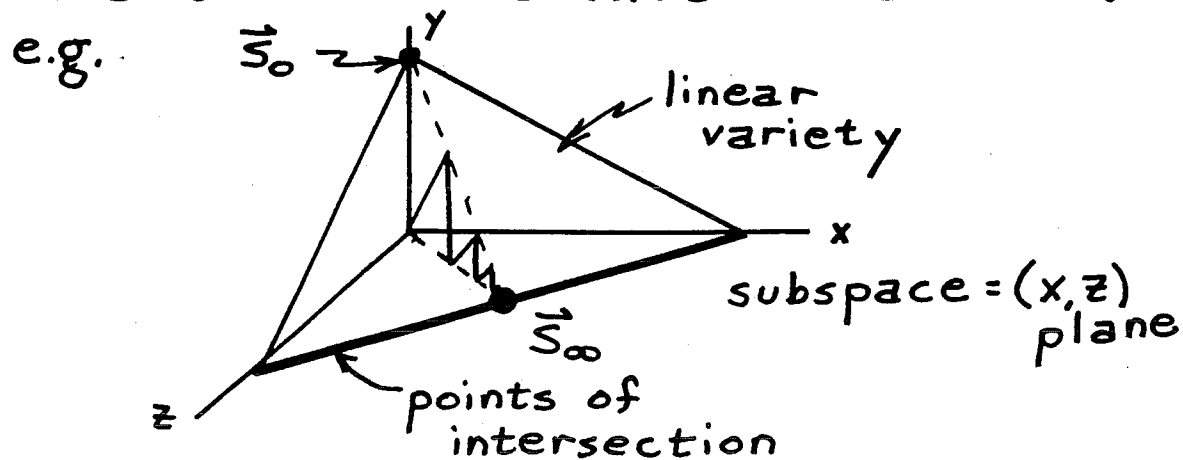


★Q: When is \vec{f} the only point of intersection?

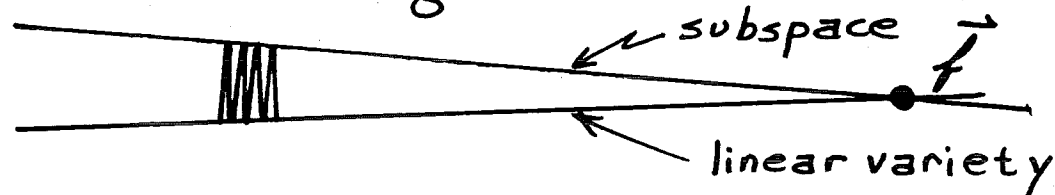
A: When $P \geq N$ and the first P rows of F form a full rank matrix.

★Q: Is this nec. for $\vec{s}_\infty = \vec{f}$?

A: No. If the varieties intersect in more than one point, convergence is to that intersection point closest to the initialization:



★Problem: Convergence can be slow:



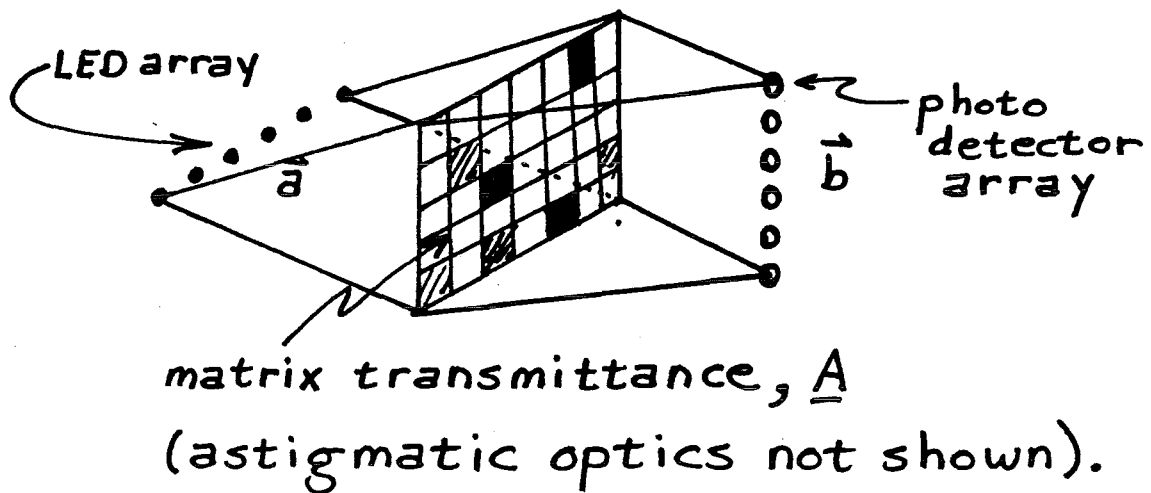
Solutions:

1. Use relaxation parameters.
2. Iterate at the speed of light.

OPTICAL IMPLEMENTATION

An optical matrix-vector multiplier:

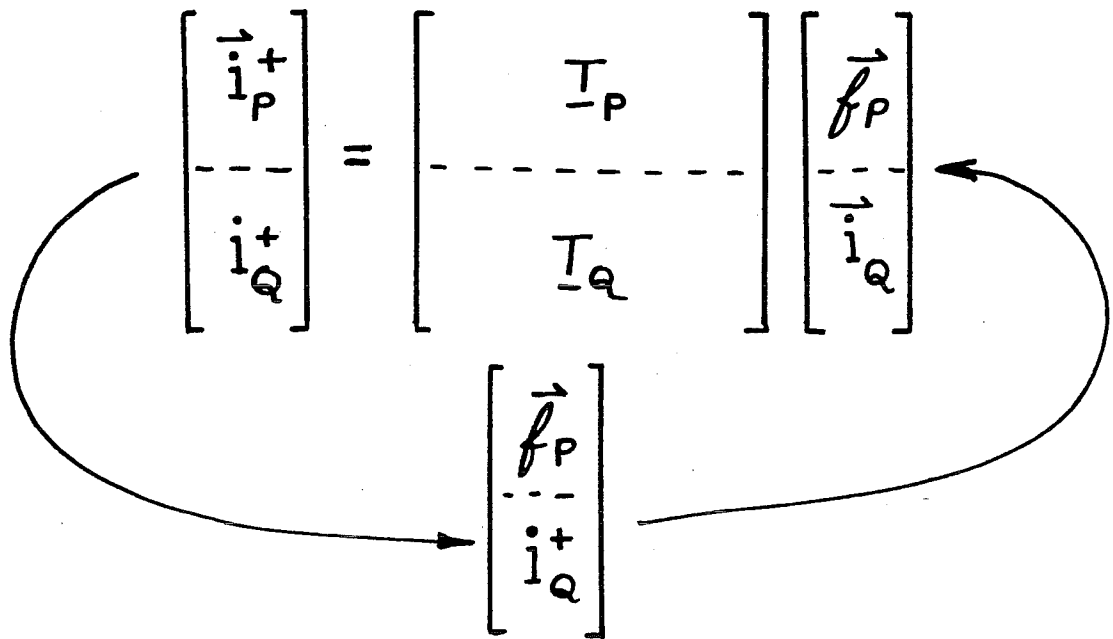
$$\vec{b} = \underline{A} \vec{a}$$



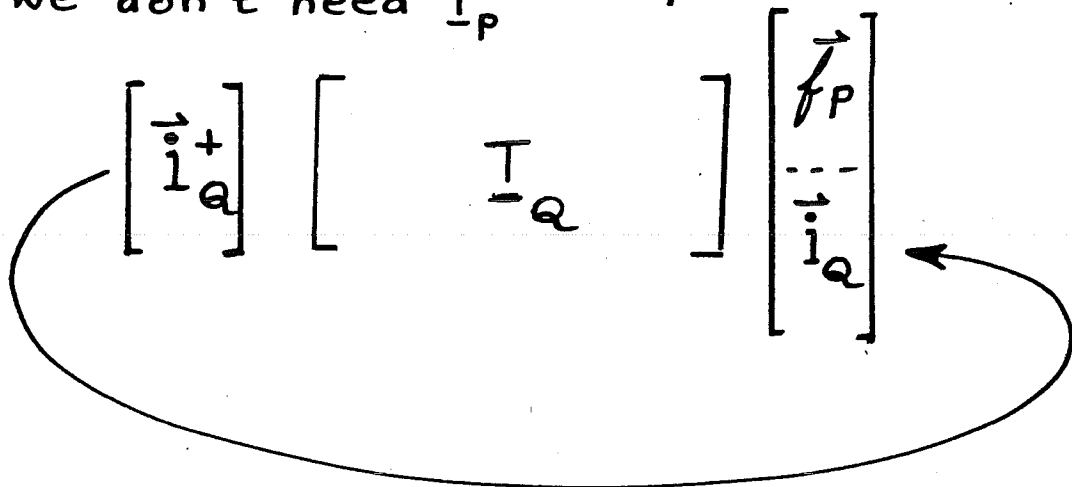
A Symptom-Diagnosis Neural Net (i.e. table look up)

Same P nodes always provide the input. The remaining Q are the response.

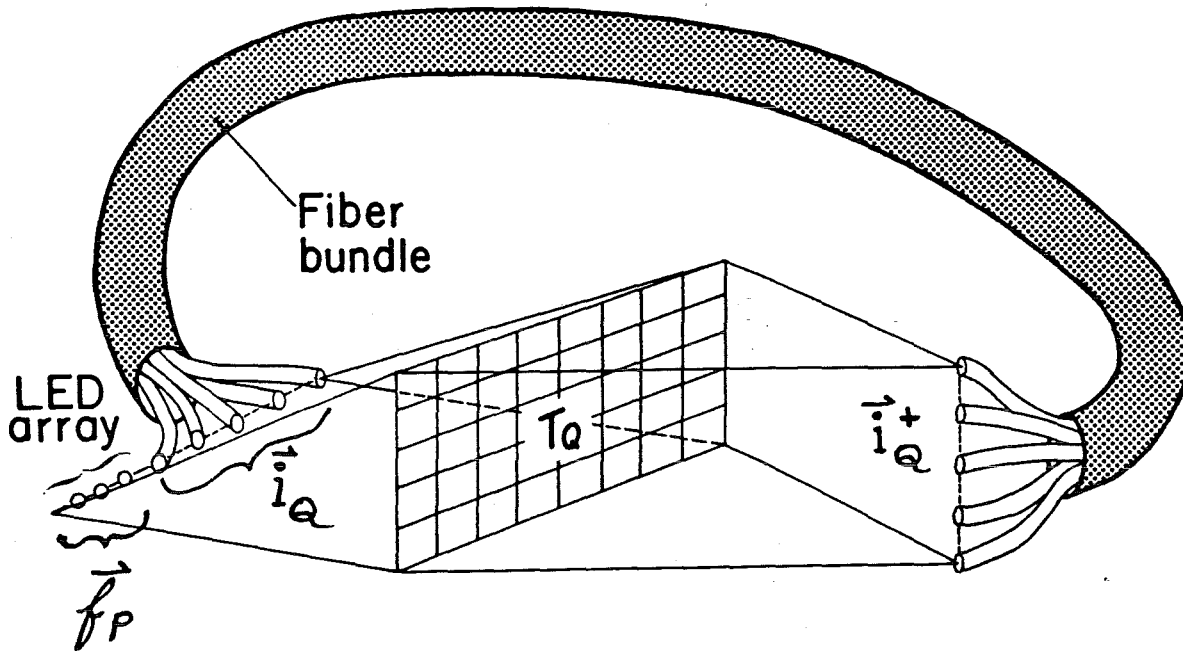
Algorithm:



There is no need to compute \vec{i}_P^+ .
 \therefore We don't need T_P



An Optical Implementation:



- Problems:

1. Detecting Output
2. Absorbtive Losses

- Solutions:

1. Place pellicle in feedback path.
2. "Amplify" \underline{I} mask transmittance.

Q: What is $(t_{ij})_{\max}$?

A: For orthogonal bipolar library:

$$(t_{ij})_{\max} = N/L$$

Here, we can tolerate an absorbtive loss of N/L per iteration.

Alternate Form:

$$\begin{aligned}\vec{i}_Q^+ &= \underline{T}_Q \begin{bmatrix} \vec{f}_P \\ \vec{i}_Q \end{bmatrix} \\ &= \begin{bmatrix} \underline{T}_3 & \vdots & \underline{T}_4 \end{bmatrix} \begin{bmatrix} \vec{f}_P \\ \vec{i}_Q \end{bmatrix} \\ &= \underline{T}_3 \vec{f}_P + \underline{T}_4 \vec{i}_Q \\ &= \vec{g} + \underline{T}_4 \vec{i}_Q\end{aligned}$$

\underline{T}_4 is $Q \times Q$.

Use Q neurons with interconnect matrix \underline{T}_4 . Neural operator:

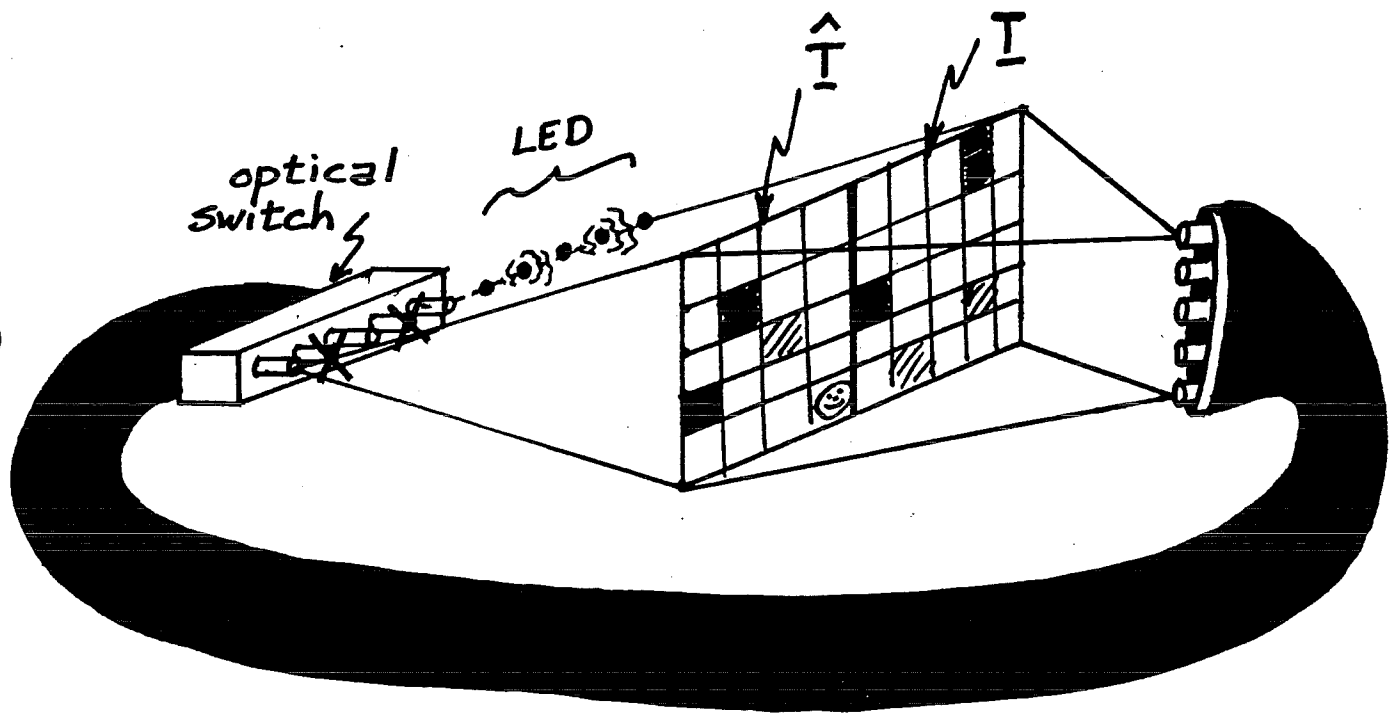
$$\underline{n} \vec{i} = \vec{i} + \vec{g}$$

Note that, in steady state, $\vec{i}_Q = \vec{i}_Q^+ = \vec{f}_Q$.
Thus:

$$\vec{f}_Q = \underline{T}_3 \vec{f}_P + \underline{T}_4 \vec{f}_Q$$

or
$$\vec{f}_Q = [\underline{I} - \underline{T}_4]^{-1} \underline{T}_3 \vec{f}_P$$

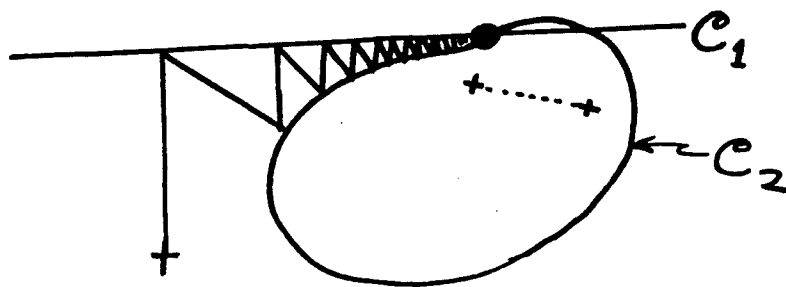
An Optical CLNN



$$\vec{S}_{M+1} = \underline{n} \underline{T} \vec{S}_M$$

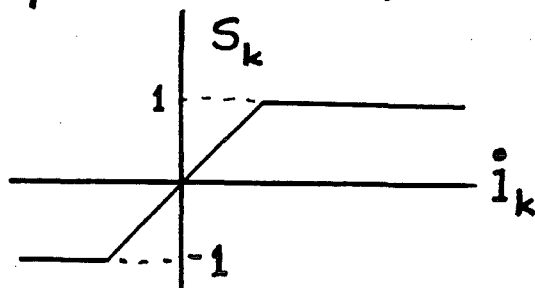
Future Work:

Extension to convex sets:



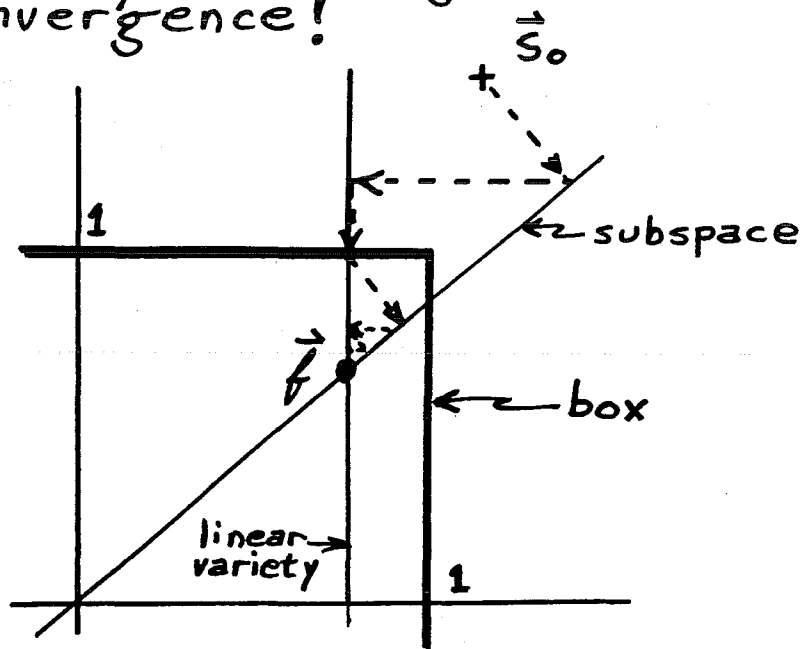
Application to CLNN:

For response nodes, instead of $S_k = i_k$,



Projects onto box.

Limited dynamic range accelerates convergence!





FINIS

**DR. ROBERT J. MARKS II
DEPT. OF ELECTRICAL ENGINEERING
UNIVERSITY OF WASHINGTON FT-10
SEATTLE, WASHINGTON 98195
U.S.A**

OPTOELECTRONICS BUILDS VIABLE NEURAL-NET MEMORY

COUPLED WITH HIGH-RESOLUTION RADAR, IT YIELDS NEW SMART SENSOR

Efforts to develop artificial neural networks modeled on the brain's highly fault-tolerant, massively parallel computing capability are quickly picking up speed. But researchers trying to build these networks with very large-scale integrated circuits are running into a spate of signal-distribution problems. Now a team at the University of Pennsylvania has taken a big step forward in this work by turning to optoelectronics instead of VLSI to build what chief researcher Nabil H. Farhat calls the first practical artificial neural net.

The Penn researchers were able to avoid the problems that cropped up in an artificial neural network's prodigious interconnections when implemented in silicon. They accomplished this by taking advantage of a simple principle of physics: light multiplexes and integrates through lenses without crosstalk. The team's work goes beyond such a neural network memory; by teaming it with a high-resolution imaging radar—which they developed—they can produce images showing details as small as 50 cm on full-sized aircraft—the highest resolution reported in the unclassified literature.

The Pennsylvania neural-net memory is an optical content-addressable associative memory (CAM), where the elements are searched in parallel by their content rather than by address. The radar and CAM work with a library of aircraft characterizers and need as little as 10% of the radar's full data set to find the closest match to a characterizer and thereby successfully identify a target model aircraft (Fig. 1).

Based on recent laboratory tests, the Penn researchers believe their system should be able to identify an incoming target aircraft at a range of a few hundred kilometers. "Its range is limited only by transmitter power and that will be extended considerably as equipment is developed," Farhat says. In commercial applications, imaging radar operating in the S or X bands with a 0.5-GHz bandwidth could prove useful for a variety of near-airport tasks, such as telling a pilot if his landing gear has been deployed.

The system will not be limited to interrogating large objects, however. When upgraded to operate in the 60- to 100-GHz bandwidth, it will be able to discern millimeter-sized detail at a range of several meters through many opaque materials. Such a capability makes the nondestructive evaluation of microwave-penetrable materials a natural application, Farhat says.

Over the past few years, theorists have taken giant strides in describing how a simple neural network might process information. But attempts to implement neural nets in VLSI circuitry have been mired in the maze of complex signal-distribution and interconnection problems among the many artificial neurons. For example, scientists at AT&T Bell Laboratories who are grappling with these problems in VLSI are making progress, according to a representative, but they have yet to engineer a solution they are willing to discuss publicly.

TWO DECADES OF RESEARCH

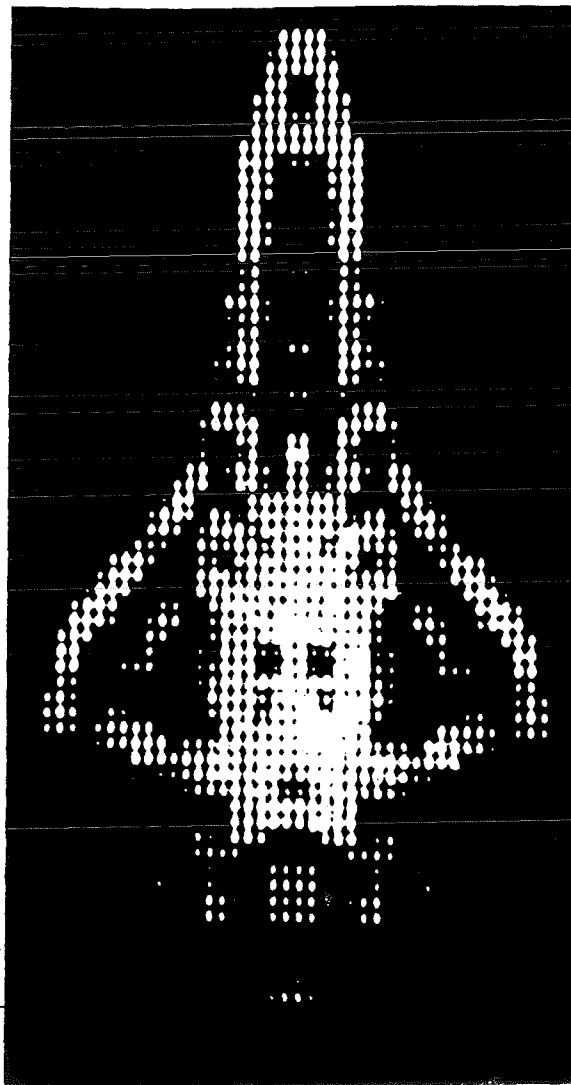
The need for a powerful parallel processor grew out of two decades of research at Penn into imaging radar. Farhat, zeroing in on his goal of near-visual-quality images, knew that real-time data generated for nearest-neighbor image searches would overwhelm all but the most powerful serial computers. A visit to the Jet Propulsion Laboratory, Pasadena, Calif., in 1983 introduced him to the neural-network concept, which resulted in lab versions of the CAM. The CAM can pare and interpret the flood of real-time data from the radar.

Dovetailing imaging radar and optical-memory technology and refining the CAM are the tasks at hand in Penn's lab.

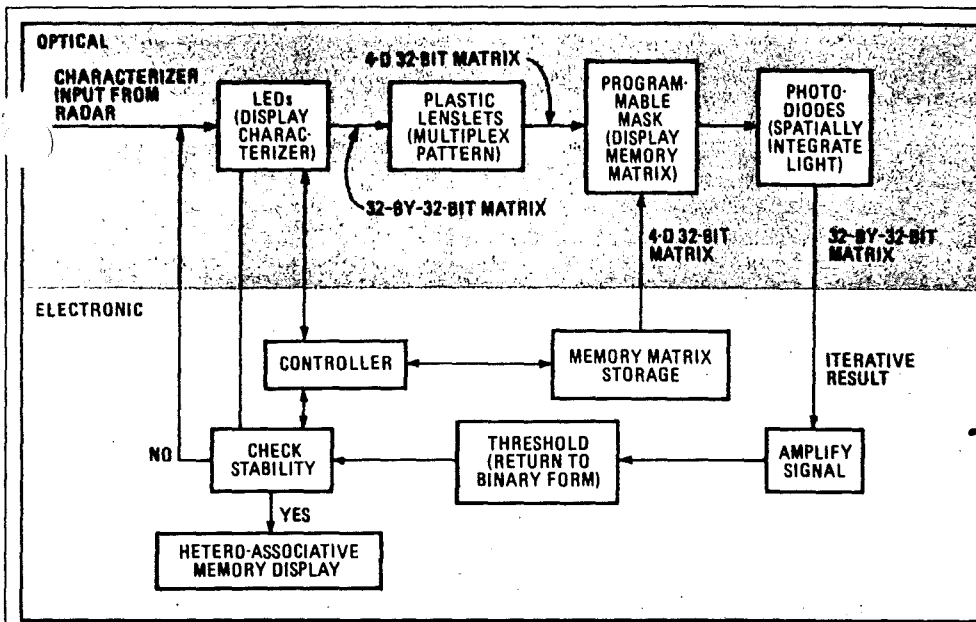
Although Farhat will not speculate on a commercialization time line, he is confident the CAM can be transferred from the lab to an optoelectronic circuit with present-day fabrication technologies. The CAM has uses in a wide spectrum of image-identification tasks, he says, especially those plagued by substantial amounts of missing or incorrect data.

Using off-the-shelf hardware such as light-emitting diodes, magneto-optic spatial light modulators, anamorphic lenses, photodiodes, and an electronic nonlinear feedback loop, Farhat's team of graduate students is probing the limits of the CAM's associative powers. In several tests, it has identified a scale-model aircraft from 10% of the full data set used to characterize the model aircraft with the high-resolution radar. Under real-world conditions, the memory will have to deal with spurious data from target vibration and wind buffeting, but Farhat expects its fault tolerance to be equal to the task.

The memory's phenomenal fault tolerance and robustness can be traced to a binary-coded memory



1. RADAR IMAGE. The optical neural-net memory creates an identifiable space shuttle image from limited data.



2. BIG FIVE. The five major subsystems in the optical neural memory are 32-by-32-bit arrays of LEDs, lenslets, photodiodes, a digital memory device, and a medium upon which to record the memory matrix mask.

matrix that attempts to mimic the synaptic connections in a simple biological neural network. The memory matrix is computed from the Hopfield algorithm, which is the basis for all neural-network-memory implementations. In its simplest form, the algorithm creates a two-dimensional memory matrix from a library of one-dimensional binary inputs known as characterizers. For practical image-identification problems, Farhat uses 2-d characterizers that expand to 4-d memory matrices.

Given a library of binary-coded characterizers each having n bits, the algorithm begins by taking the first characterizer and computing a simple numerical relationship between each bit and the remaining $n-1$ bits. This yields an expression of n^2 bits arranged in a matrix twice the dimensions of the input characterizer's matrix. The same operation is performed for each characterizer, with each result summed in the memory matrix's relevant position to create a decimal expression. After the algorithm has churned through the entire library of characterizers, it "clips" each matrix element—that is, it sets any positive numbers to one and any zeroes or negative numbers to zero—to return the matrix to binary form. As a finishing touch, the memory matrix's self-products are set to zero. This satisfies the algorithm's recipe—and the intuitive deduction that in this simplified model neurons do not communicate with themselves.

Once the memory matrix is created, the CAM can begin the real work of identifying objects through nearest-neighbor searches. These nearest-neighbor image searches are not conducted in the serial, bit-by-bit progressions used by a conventional recognition scheme. Edge enhancement plays a role in the identification process, but it is a derivative of the radar—not of data manipulation within the CAM. Edge-enhanced information about the target is formatted into a 2-d matrix, which forms the target's characterizer. The memory begins its search by multiplying the 4-d memory matrix by this 2-d characterizer. Then the 4-d result is reduced to a 2-d first approximation by summing array elements according to the Hopfield algorithm. This yields a decimal result that is clipped and fed back to the algorithm. The iteration ends when the result stabilizes—usually in two to four iterations—on a version very close to one of the library characterizers. When the CAM does not have enough information for a successful search, the result may oscillate between two characterizers.

Although minimum matrix dimensions for library characterizers will depend on identification tasks, Farhat says a 32-by-

ability to fill in missing data. Explaining the CAM's ability to compute an accurate approximation of one particular library characterizer is not as easy. Neural-network theorists generally return to the analogy of the neuron's state as firing or not firing (on or off). Each characterizer, then, is equal to a stable energy state for the memory matrix. When excited by partial information, the matrix comes to rest at or near the closest stable energy state of the input.

The precise commercial architecture for an optical CAM is still to be determined but will probably include two major functional systems—one to create the memory matrix, the other for nearest-neighbor searches.

FIVE MAJOR SUBSYSTEMS

To create and store a memory matrix from a series of library characterizers consisting of a 32-by-32-bit array of data points, an optoelectronic processor would consist of five major subsystems (Fig. 2):

- A single-chip, 32-by-32-bit array of GaAs LEDs for the display of each characterizer pattern.
- A 32-by-32-bit array of molded plastic anamorphic lenslets to multiplex the displayed pattern.
- A 32-by-32-bit photodiode array to record the output of the memory mask and integrate the result.
- A digital memory device to drive the LED display with characterizers. When the programmable mask is implemented, another memory unit will be used to store the memory matrix.
- In a primitive version, photographic transparency film on which the memory matrix mask is recorded. Eventually, programmable, nonvolatile, magneto-optic spatial light modulators will be used for real-time operation.

To create a memory matrix, a library characterizer is displayed on the LED array. Lenslets serially multiplex the image (all but one lenslet is covered at a time). The multiplexed image interacts with the mask programmed to represent the characterizer. The result recorded by the photodiode represents the first submatrix. The same procedure is followed for each lenslet. The results from each characterizer are summed in the matrix but finally clipped.

Nearest-neighbor searches require the addition of more electronic components. The two most notable are an array of masks, so the memory can be displayed in its entirety, and a nonlinear feedback loop to amplify multiplexed optical signals attenuated in the iterative cycles. Other circuits address the

32-bit matrix should be large enough for most radar applications. Nevertheless, calculations inherent in a library of 32-by-32-bit characterizers would challenge most serial computers.

It is the massively parallel computing capability of an optical search that makes Penn's CAM shine. Data throughputs will most likely be limited initially by the cycling capabilities of magneto-optic light modulators used as the programmable masks that display the memory matrix. Farhat expects to meet his near-term goal of reprogramming the masks at 1,000 frames per second. Using 32 masks in parallel, each displaying a 32-by-32-bit matrix, yields a throughput of about 3.2 million bits per second.

The extensive interrelationships created by the algorithm give a sense of the memory's

large applications by-32-chal-ers. arallel n opti-Penn's ghppts initial-ties of alators mable emory meet repro-1,000 ng 32 isplay-yields .2 mil-lation-orithm mory's ility to library eneral-or not stable partial closest CAM is majc ix, the ries of of data major he dis-enslets of the h char-ed, an-atrix. ilm on y, pro-odula-is dis-the im-plexed resent de rep-llowed er are : mor ray and a signals ass the

memory matrix and monitor the stability of the iterative result. During the search, the target's characterizer is multiplexed simultaneously by the entire lenslet array to interact with the complete memory matrix. Single photodiodes positioned behind the masks integrate the light, which, when clipped, yields a 32-by-32-bit iterative result.

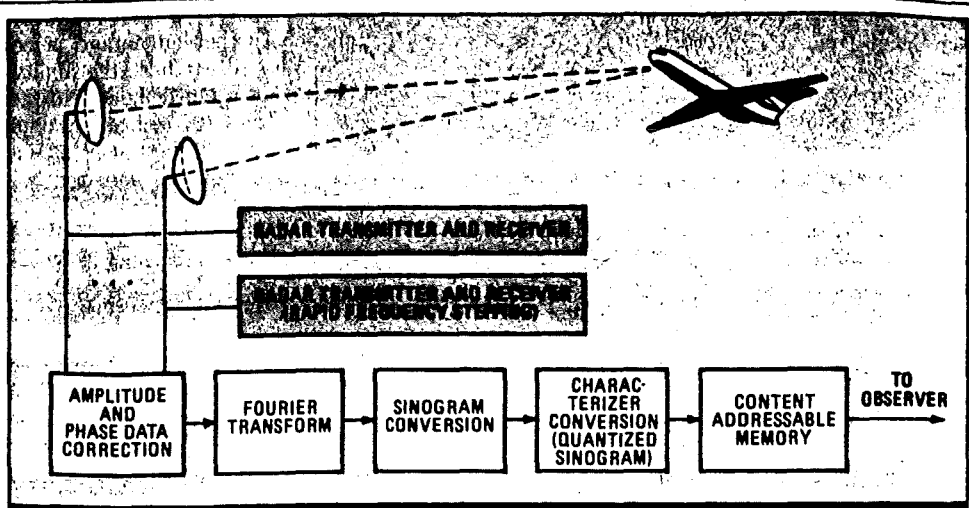
Once the CAM stabilizes, it still has to interpret the result in terms of one of the library characterizers. There will seldom be an exact match. In a remarkably clever solution, Farhat has utilized the human observer's ability to recognize a less-than-perfect image. In addition to creating homo-associative memories—relationships of the information

with itself—the CAM can form a hetero-associative memory. This means it can relate the same information to another image on the screen, such as an alphanumeric expression. In other words, the system output displayed on a cathode-ray tube is an imperfect four-character code for the object being identified—but something a technician can interpret nonetheless.

The CAM's ability to zero in on a target depends on the memory matrix's size and the number of characterizers that the matrix incorporates. For a 32-by-32-bit matrix, the CAM has a near-100% probability of stabilizing on a hit if the library consists of 30 or fewer characterizers, says Farhat. Though fewer than the hundreds needed for a library of characterizers to identify military and commercial aircraft, it is not a limitation. The CAM simply loads the first 30 into the memory matrix. If it doesn't succeed with that matrix, it loads additional sets.

As robust as the CAM is, to meet the needs of its intended applications in aircraft image identification, robotics, and a variety of other recognition tasks, it requires relatively precise images drawn by smart sensors that eliminate unimportant information. By harnessing an innovative combination of frequency diversity, holography, and Fourier analysis, Farhat hurdled three persistent problems besetting imaging radar: enormous aperture size (and its correspondingly high cost), noise, and image orientation.

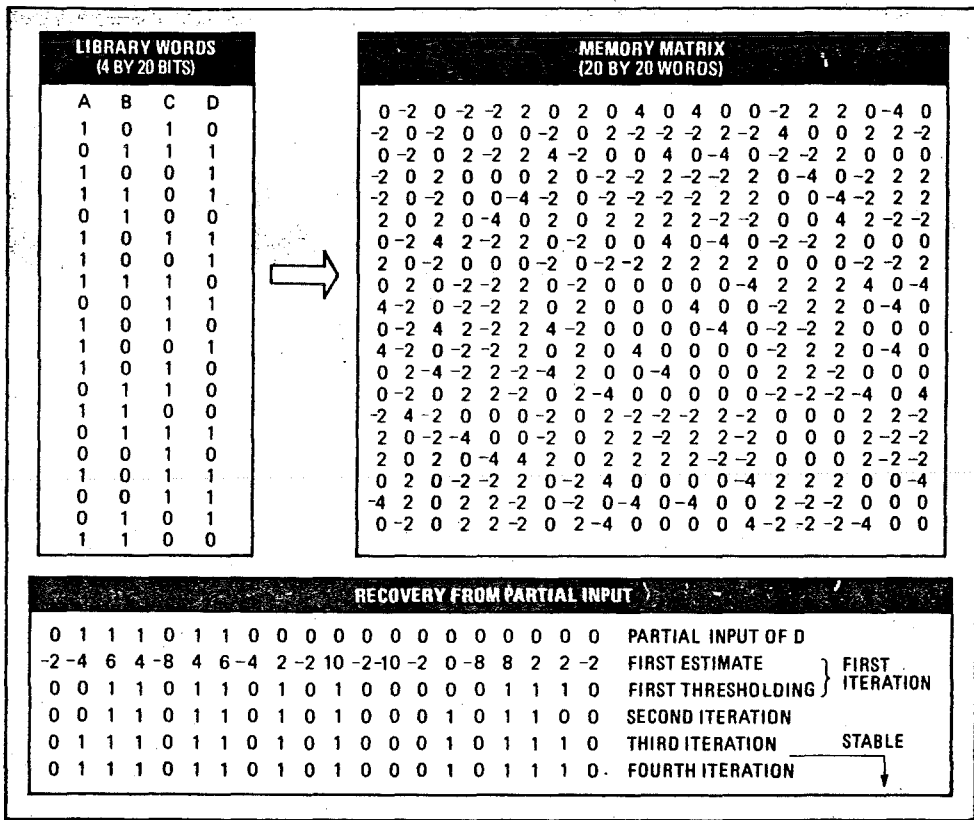
The high resolution of Farhat's imaging radar is itself a breakthrough, achieved by a combination of frequency diversity, polarization, and multiple views of the target. His chief innovation is wavelength diversity—stepping across a range of frequencies and using different polarizations (Fig. 3). Fourier analysis turns that data into the mathematical equivalent of the impulse response of the target.



3. GRABBING AN IMAGE. The University of Pennsylvania's highest-resolution imaging radar steps across a range of frequencies to provide data to the CAM for generating the final image.

Edge enhancement is a natural by-product of microwave frequencies and the scattering mechanism of the target. "The radar produces a primal sketch of the object," says Farhat. "Edges are enhanced and information about the object's flat parts is discarded because most of the radiation that hits a smooth surface scatters forward. In contrast, in the optical regime I would see the entire object because the surface is very rough compared to the wavelength."

To obtain a 2-d image, the radar must view its target from more than one aspect angle, or look. "Each look gives one frequency response," says Farhat. "When we perform a Fourier transform on that, we get the equivalent range profile. By repeating the measurement for many looks and putting the frequency responses one next to the other in polar format, we get a Fourier space slice." The system's tomographic capa-



4. ROBUST. The CAM's ability to fill in missing information is illustrated by setting the last 12 bits of library word D to zero. After only four iterations, the CAM found the correct word in a library of four.

bility makes it possible to produce a near-visual-quality projection image from any Fourier slice of the object, even though the radar's looks are from a variety of slant angles. "If you've looked at a target from head on to broadside, which is how you've probably characterized it totally because aircraft are symmetrical," says Farhat. "You might argue that you have to look at it from the rear, but in most cases you aren't interested on identifying something that left you. We store 90° characterizers in memory—60° would be enough, but we're looking at 90°."

Visual representation of the projection image does not provide the best CAM characterizers. Range profile data can also yield sinograms—sinusoidal traces that produce a more distinctive signature. The research team currently constructs characterizers from sinograms; but options such as polarization are worthy of consideration and are being studied, says Farhat.

The imaging radar is closer to commercial implementation than the CAM. Existing radar technology operating between 300 and 500 MHz in 3-MHz steps could be adapted to achieve 30- to 50-cm resolution, says Farhat. "What would be required would be precise, rapid frequency stepping to acquire the data. But using our laboratory equipment and the proper radio-frequency amplifiers, you could set up such a system." Several radar tracking stations would independently measure the target's frequency response from different directions. The data would be transferred to a central computer bank and corrected for phase differences to access a slice in the target's Fourier space.

About 200 frequency steps and 500 looks within a 90° azimuthal angle would generate a complete high-resolution image for a 50-m aircraft. The number of looks is determined by the target's size: the larger the target, the more looks are required. But this does not mandate a system of 500 discrete sensors distributed around the target. First, the aircraft's motion allows each sensor to address

the target at more than one angle. Second, the CAM's sinogram-derived associative memory can be counted upon to fill large blocks of missing information (Fig. 4). Using actual radar-retrieved data of a model airplane, the CAM (simulated on a computer because the optical version has not attained a 32-by-32-bit array size) has identified the target with as few as 12 looks when 128 looks and 128 frequency steps were used to create the library characterizers.

Having coaxed the imaging-radar system through research to the verge of development, Farhat has two remaining goals: achieving millimeter-level image resolution in the laboratory and refining the optical CAM to process the radar data in real-time.

A Department of Defense University Research Instrumentation Grant is funding a major upgrading of the radar laboratory to add equipment for millimeter-level image resolution, scheduled for completion this fall. It will facilitate frequency stepping as high as 60 GHz and enable

the study of such real-world problems as target vibration.

Another advantage is economic: millimeter resolution will give researchers the ability to characterize full-sized aircraft from detailed models in the laboratory, says Farhat.

The CAM technology must be upgraded in two key areas to mesh with the radar for real-time operation. Using a simple 5-by-5-bit neural network, researchers have finished ironing out the generic wrinkles of optoelectronic CAMs. Next they will implement a 16-by-16-bit neural network; within a year, Farhat expects to be using 32-by-32-bit sinogram characterizers derived from five target aircraft models available in the radar lab. "At that point, we will want to find out how well it recognizes the models on a statistical basis from any aspect angle," he says. Computer simulations indicate the 32-by-32-bit optical CAM might make do with 10% of the total characterizer data.

MOVING TO A MASK

In principle, moving from the present technique of storing interconnection data on transparent photographic film to a programmable mask should not pose serious difficulties, says Farhat. Litton Industries, Van Nuys, Calif., markets 48-by-48-bit magneto-optic spatial light modulators that can be used as the storage mechanism for 1-d neural networks, he says, and adapting the system to a 2-d neural network is a relatively simple matter of partitioning the 4-d memory matrix into 2-d components.

Over the long-term, the CAM could have an impact on such technologies as robotics, machine vision, artificial intelligence, and supercomputers. Recognition schemes could include ultrasound, colors, textures, infrared, and—perhaps the CAM's first commercial application—speech processing. Coupled with the imaging radar's smart sensing of primal images, the research will prove fruitful in gaining insights into imaging as a whole, including the eye-brain system, says Farhat. □

TAKING LESSONS FROM MOTHER NATURE

For more than 20 years, Nabil H. Farhat has been extracting images through opaque media with constantly improving results. Beginning with microwave holography in 1964, he and his ever-changing team of University of Pennsylvania graduate students managed by 1969 to derive fuzzy holographic views of concealed objects such as a handgun in a suitcase. Though the images were impressive, his research convinced him that single-wavelength holography was stuck with the inherent limitations of speckle noise, range, and cost.

Turning to nature, Farhat reasoned that if bats and dolphins can resolve their environment with great precision using multifrequency clicks and chirps, then spectral diversity might also provide a key for high-resolution radar imaging.

From the project's holographic beginnings, Farhat was constantly on the lookout for a hybrid optical and



FARHAT: The clue came from bats and dolphins.

electronic system for real-time data processing. While on a sabbatical trip to the California Institute of Technology in 1983, Farhat visited the Jet Propulsion Laboratory and became intrigued by the work in associative memory and neural networks that was being done by John Lamb and his colleagues. "They handed me a paper on the Hopfield model [an algorithm developed by John J.

Hopfield], and everything fit together," he recalls. "It was perfect, especially for an optical model."

A week later, he discovered that Caltech's Demetri Psaltis had similar interests. "We put our heads together and wrote a paper drawing the optical community's attention to how well neural networks dovetail with optics." In the human brain, individual neurons can become inoperative without damaging the neural network's integrity—a highly desirable trait for computer or imaging systems that must function for extended periods, as on future space missions that could last 50 to 100 years.

Together, imaging radar and associative memory have numerous applications from determining the condition of heat-resistant panels on the space shuttle to checking rush-hour traffic conditions around New York. Yes, Farhat says, a representative of New York's Port Authority has already contacted him.

onitor solution
xel of graphic
uter system
S color display
x your most
I/CA
ds mo ng

nic Convergence
ature Lenses, H
s, are examples
nology that can
icker-free image
ail, and color.

der in

world's most
hips, and we're
ucer of CRTs.
to support your
ation and service
ou get a partner
help your system
creen.

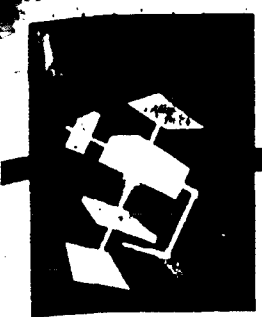
S monitor.
f 15", 19", or 25"
e the ideal moni
—with superior
e, brightness,
erfor ce.
s of g
matched prod-
-stock quality,
assistance

ty to thoroughly
tor and support
rite today.

r
ities

ed.
Bruno, CA 9400
902
dale, NJ 07401
00

CHI®
on solutions
10. 22



Optics and neural nets: trying to model the human brain

Attempts to build computers that work on the same principles as the brain will require radical rethinking of what we consider to be computing. The communications power of optics may play a vital role in this endeavor.

Tom Williams
Western Managing Editor

Humans are not logical. This familiar Vulcan proverb illuminates one of the issues frustrating computer scientists and users. While computers outperform the human brain in solving certain classes of mathematical and logical problems, they appear woefully inadequate for other tasks that humans can do instantly, such as pattern recognition and association in real time, using incomplete or distorted input. Why is there such a difference in ability? Is it possible to construct machines that can compete with the human brain in solving the problems that seem to come to it most naturally?

It appears that the quest for ever-faster switching speeds in digital circuits will not provide the solution. Even supercomputers using silicon and gallium arsenide circuits with subnanosecond gate delays bog down at true real-time pattern-recognition tasks, while our brains can perform such problems instantly. And the response time of a neuron is in the millisecond range—not even close to the speed of silicon ICs. An avenue of research is emerging that seeks to understand not only the structure of the brain but also the differences in the class of problems that it's best designed to solve.

Professor Demetri Psaltis of the California Institute of Technology (Pasadena, CA) believes that today's computers lend themselves to solving problems that, by nature, are structured in such a way that they use algorithms having many short steps. Computers break down, however, when confronted with problems that are inherently random, such as pattern recognition. Structured problems, even those that lend themselves to parallelization, are deemed difficult in terms of the time-complexity they involve—or, in other words, the number of steps they take. Humans, however, don't recognize scenes by executing sequential steps, but rather by a process of global associations—processing all the received data at



*John Caulfield
Director, Center for Applied Optics
University of Alabama in Huntsville*

Spatial light modulators—the critical component

Optical computers of many designs, sizes and functions are taking shape on blackboards and in laboratories around the world, as work goes on to improve the component used as input, output, scratchpad memory, interconnector and processor. This component is the spatial light modulator (SLM). A recent survey done by the Naval Research Laboratories listed 50 versions of the modulator with a wide array of properties. But what is an SLM, and why is it so critical?

An SLM is a transducer. It converts a two-dimensional pattern of light into a spatial pattern that can vary its brightness. Both continuous and binary outputs are available. There are many other forms the modulation can take. Many SLMs produce a spatial variation in polarization. Others produce patterns of relative phase. And still others produce coupled changes in two or more of these properties.

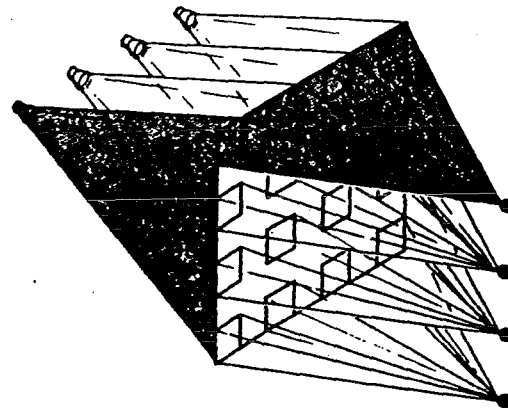
The SLMs can also do processing. The output spatial pattern doesn't need to be a faithful copy of the input pattern. Some of the input-output relationships include those shown below.

Input-Output Relationships	
Input Pattern	Output Pattern
Incoherent, wide band	Coherent, narrow band
Continuous in intensity	Binary in intensity or other property
Well-defined intensity pattern	Reversed-intensity pattern
Continuous image	Edge-enhanced, dynamic-range-compressed image

With a steeply nonlinear I/O pattern, the SLM can perform logic operations (AND, OR, NAND, NOR) on input light patterns. Also, such SLMs can do thresholding, level restoring, clocking, and so forth. One particular form of I/O nonlinearity is called bistability. In bistable SLMs, picture elements, or pixels, that are turned on may stay on until switched off. This provides for memory and for clocking.

SLMs can provide a type of scratchpad memory simply by time integration of multiple input pat-

terns. In the same way, they can convert a serially scanned input pattern into a parallel read output pattern. With appropriate input and output optics, we can cause each of N input points of light to be connected to each of N output points via an $N \times N$ SLM, as shown schematically below.



This simple arrangement is very powerful. By blocking $N-1$ of the N sources going to any output, we can affect any desired I/O interconnect pattern. By regarding the inputs and outputs as vector components, we can view the SLM as a matrix. This yields a parallel matrix-vector multiplier. The matrix may represent something as simple as an algebraic problem or something more complex such as a neural network.

SLMs can also be used to compare in parallel many "symptoms" with many indicated "disease patterns" for optical expert systems.

Using holograms to address the SLMs, we can access between 10^4 and 10^6 different patterns at a very high random-access rate. Unfortunately, current SLMs don't respond that fast. Using moderate values (10^5 patterns of 10^5 pixels accessed in 10^{-3}), we arrive at a phenomenal pixel usage for a vast store of 10^{10} pixels, accessible in 100 s. This capability is well beyond the current capability of electronics, and well below the ultimate limits of optics.

These are only some of the many diverse forms and uses of SLMs. Many experts throughout the world agree that the SLM is the most critical and versatile of the required components of all optical computers, whether they're neural networks or numerical processors.

once. To expect a digital computer to be able to approximate this process in step-by-step algorithms is unrealistic.

The goal of computer scientists is to design a machine that can approach these random-type problems on the basis of global association and with incomplete data to reach valid conclusions. The most successful approach to date appears to be the neural-net model based on the interconnection pattern of neurons observed in the brain.

While there are many differing opinions on how to best implement the neural-net model, and while any such neural model would be a greatly simplified version of the brain's organization, all agree that neural nets represent a radical departure from any previous digital computer architecture. The hallmark of the neural net is massive parallelism and high interconnectivity between a large number of relatively simple processors. The information in a neural processor is stored in the interconnection pattern rather than at specific spatial locations uniquely defined by a memory address.

The need for interconnects between each of a large number of neurons makes implementation in silicon of any system approaching the brain's complexity look intimidating to many researchers who see the inherent high bandwidth, high parallelism and global communications properties of optics as a possible solution. In addition, research is showing that there are certain types of processing tasks, such as matrix operations, that are highly parallel and could lend themselves to solutions by optical processors even though they're not modeled after the brain. Some feel there may be a natural overlap between the needs of neural networks and the capabilities of optics.

Both neural nets and highly parallel numeric problems—if implemented optically—would represent analog processes and would be a radical departure from the von Neumann sequential architectures that have characterized digital computers for the last 30 years. It must be noted, however, that there's a good deal of research into bistable optical devices—including some using gallium arsenide—that could offer a quantum leap in switching speeds beyond that of current machines. But given the nature of time-complexity, such machines would excel at the tasks at which today's architectures already perform well. They wouldn't come considerably closer to the category of random problems at which neural nets are taking aim.

Indeed, resolving such problems requires a fresh look not only at the unique requirements of the machines, but also at the capabilities offered by neural-net and optical technology. Consider the op-

tical technology most familiar to current computer users—the optical disk. Today's optical disks are used as if they're simply high-capacity magnetic media. Data is optically read bit-by-bit with a single laser. Light, however, has the ability to shine on an entire surface at once and has the potential to decode all the recorded bits in parallel at once. If there were some way to utilize that capability, optical storage would be faster, and more important, would represent an entirely different way of using data in a computer system.

The job of building a functional model of the human brain has only just begun. The work is in its most tentative and fundamental stages, and functioning intelligent systems won't be built on neural networks for decades, if then. Nonetheless, there's widespread recognition that it's possible to build hardware models of certain brain-like structures that act like analogous neural circuits observed by neurophysiologists. We know that a working model exists in nature—our own brains—and it's now possible to see the direction we must take to emulate that system.

In addition, work has emerged from academia in the form of startup companies. Such companies are looking for certain applications that would lend themselves to solutions on the basis of what has already been learned. "This is an important step," says Lauren Yasolino, president of Synaptics (San Jose, CA). "Applying the research will provide a feedback loop and aid development of the technology." Synaptics' vice-president of research, Federico Faggin, cautions that neural-net computers will probably not operate like human brains. "If we modeled airplanes after nature, they would have feathers," he says. "The goal is to understand and grasp some fundamental principles and translate those into silicon."

Since research in neural nets is at such a fundamental stage, no agreement has been reached on how to best implement an actual neural circuit. In fact, even an attempt to pin down a definition of neural network leads to lively debates among neurophysiologists, engineers and information scientists. Still, a general consensus is emerging that such circuits will behave like neurons in that their stored information will be distributed among the various nodes and their connections rather than being located at discrete memory addresses. The circuits exhibit high parallelism and interconnectivity and are basically analog in nature. Qualifying "basically analog" is important here because neurons in the brain exhibit analog gradients of stimulation in the area of the dendrites, but send information along

serially
output
optics,
ht to be
in $N \times N$

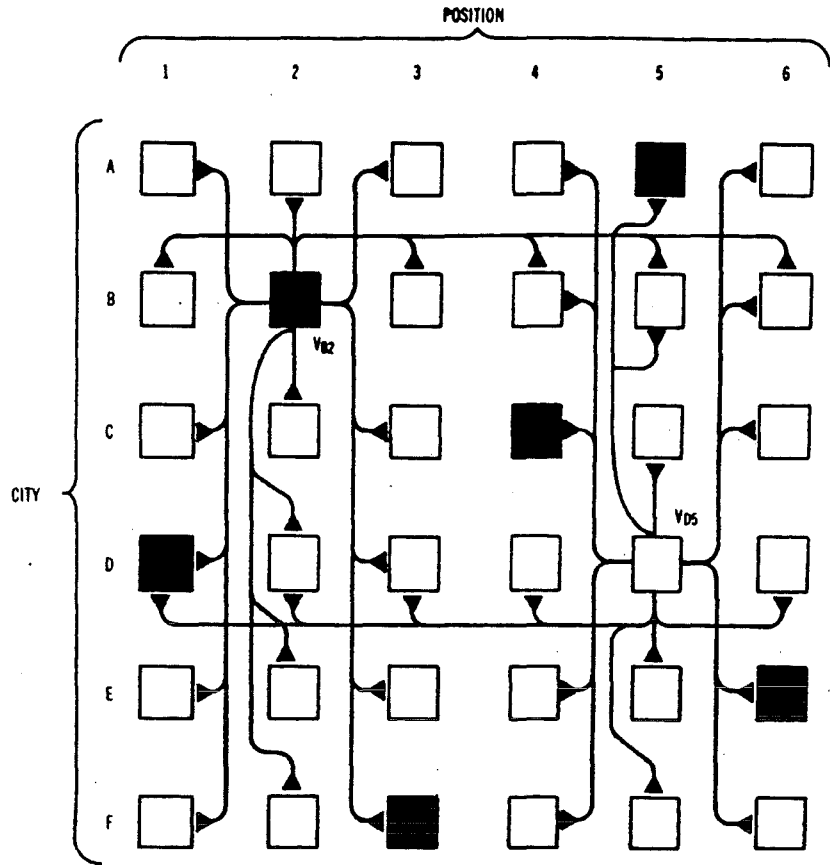
erful. By
output,
pattern.
tor com-
rix. This
The ma-
an alge-
such as

parallel
disease

we can
erns at a
tely, cur-
noderate
d in 10^{-3}),
pr a vast
his capa-
y of elec-
imits of

se forms
hout the
tical and
ll optical
ks

This stylized schematic of connective syntax for a problem, in which a salesperson must visit a given number of cities in such a sequence that the total distance traveled is minimized, shows the partial connectivity for two neurons (B2 and D5). Stimulating connections, shown only for two adjoining columns each, must extend to all other neurons. Inhibitory connections, shown in red, ensure that if a neuron is "on," all other neurons in the same row and column are suppressed. Resistance values representing distances between cities are built into the connections.



their axons in pulses, which is what the term "firing rate" applies to when speaking of nerves. The fact that these pulses carry information is clear, but they vary in pulse width, amplitude and frequency, and the mechanism of information encoding is not yet understood.

Researchers' attention has recently started shifting from these nerve pulses, or action potentials, to the synapses of nerves where the real processing occurs and the real information exists. Nerve pulses take place along axons, and many brain cells don't even have axons. "Concentrating on action potentials is like walking into Electronics 101 and hearing the professor say 'OK, the really important thing about electronics is touch-tone dialing,'" says Carver Mead, an electrical engineering professor at California Institute of Technology. Mead is also affiliated with Synaptics.

Given the high connectivity and distributed nature of information in neural nets, it's clear that for a neural network to store and process useful information, the number of nodes has to reach a certain level of complexity. As John Neff, project manager for the Defense Advanced Research Proj-

ects Administration (Arlington, VA), says, "For these [neural nets] to be practical, they're going to have to have a million or more nodes." Such a system would be inherently robust because removing a given node wouldn't destroy vital data, although at some point, the number of mistakes caused by disabling nodes would cease to be acceptable. Synaptics' Faggin is among those who see a real possibility in implementing neural nets in silicon. "We can now really think of doing waferscale integration," he says. "Flaws that inevitably occur on a whole wafer would only represent a statistical factor in the wafer's quality, they wouldn't render the whole wafer useless."

Professor John Hopfield, professor of chemistry and biology at the California Institute of Technology has proposed a theoretical model that illustrates how a neural network might lend itself to problems that entail combinatorial complexity. Hopfield's model involves a network of interconnected neurons that's set up to reveal an optimization in terms of a global minimal energy state for the system. In other words, when the circuit is started, it will reach a stable state that represents the minimum sum of energy for the whole circuit. Certain nodes, how-

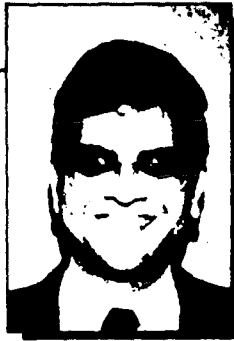
System Work
Deter
Intr
recu
Data
Map
asse

PRIN
CIRC
DESI

INF
CIR
DE

C

BRI



*Ravindra Athale
Optical Computing Manager
BDM Corp*

Neural net models for computations

Creating an artificial intelligence system that has the flexibility and the creativity of the human mind is one of the oldest goals of computer science. Ironically, it has remained the most elusive, despite the tremendous strides made in microelectronics and in general-purpose digital computing. The AI field has made substantial progress during recent years, performing difficult but well-defined tasks. These tasks include playing chess, diagnosing diseases and inferring chemical structures of complex molecules.

Yet the staggering computational capabilities that can be achieved by the supercomputers and complex, powerful organizations of the AI systems of today seem far from accomplishing some elementary but poorly defined tasks, such as understanding and producing continuous speech, moving in a complex and dynamic three-dimensional space with only the aid of somewhat noisy two-dimensional detectors, and making inferences using common sense. Most human beings find these tasks easy to perform, whereas the more well-defined tasks that AI can solve require long and arduous training for humans.

This observation has led scientists in the neuroscience, psychology, mathematics, physics, computer science and electrical engineering fields to conclude that intelligent biological systems, such as the human brain, are organized along fundamentally different lines than most AI systems. In spite of the differences in their backgrounds and in the approaches that they follow, the researchers share a common goal of trying to gain a basic understanding of how intelligent biological systems solve incompletely defined problems, and applying these principles to the design and construction of AI systems so that they can solve those problems.

The unique features of biological systems can be seen from two different perspectives—the hardware used and the control strategies used. The basic elements of the biological hardware are neurons, which are simple processing elements, and a system of synapses, dendrites and axons that interconnect different neurons in a complex network. In addition, there are a host of different biochemical reactions that control the behavior of these units. It's postulated that the human brain may have as high as 10^{10} neurons, and each neuron may be connected with up to 10,000 other neurons via modifiable synapses. The system, therefore, has up to 10^{14} free variables. In a conventional computing system, the number of free variables, including RAM and mass storage capacity, may be up to 10^{10} . But since the biological components are a million times slower than their electronic counterparts, the large number of free variables alone doesn't explain the mismatch in capabilities.

It appears that the disposition of the computational resources is the key to the puzzle. In a biological system, the computational load is basically evenly distributed between communications and decision making, so at any given time a substantial fraction of the decision-making units are performing meaningful computation. Electronic systems, on the other hand, can use only a minute fraction of the total hardware at any given instant in time. Thus, biological systems are more efficient in their use of available computational resources.

Because of their distributive and redundant decision making, biological systems possess a large degree of fault tolerance to partial failure of the hardware—a feature that isn't shared by their electronic cousins. The communications in biological systems not only occurs simultaneously in parallel between decision units, but also with the out-

ever, will be more "on" than others and from them, one can read the solution to the problem.

One of the characteristics of synaptic interfaces is that they include connections that stimulate a neighboring neuron as well as connections that inhibit stimulation of that neuron. Some even have inhibitory feedback loops onto themselves. It's clear that information is stored not only in the state of a node, or in the existence of a connection, but also in the strength of that connection.

One of Hopfield's examples involves the problem of a salesperson having to visit a given number

of cities, visiting each city once, in such a sequence that the total distance traveled is minimized. Hopfield has arranged a set of connections with neurons in rows and columns. The rows, which are labeled alphabetically, will correspond to the cities on the tour. The columns correspond to the locations of the cities in the tour. Thus, if the node in row A, column 5 is most strongly energized when the system reaches stability, city A will be the fifth city in the sequence of the tour.

Setting up such a problem requires what Hopfield calls "a complex topology of syntax-rein-

systems can
actives—the
egies used,
hardware are
g elements,
s and axons
n a complex
of different
behavior of
human brain
nd each neu-
00 other neu-
ystem, there-
n convention-
ee variables,
acity, may be
components
ir electronic
ree variables
capabilities.
he r outa-
zzle. bio-
d is t ally
ications and
a substantial
are perform-
nic systems.
nute fraction
stant in time.
icient in their
rces.
edundant des-
ssess a large
failure of the
by their elec-
in biological
ously in par-
with the out-

side world. Electronic supercomputing has turned a computational problem into an input/output problem by speeding up the individual decision-making units and increasing their connectivity. It's entirely plausible that the unusual structure of a biological computational system determines the unique functions that it may perform.

Another critical feature of biological systems is the emphasis that is placed on self-organization and learning. Digital electronic systems rely on software for guiding the flow of control and data signals. This feature may prove to be crucial to useful applications of neural nets because the network can form a complete internal representation from a partial description of the problem. It may no longer be necessary to completely understand the problem in order to solve it. In addition, the same established learning principles may be applied successfully to all problems within a given class, changing the goal from understanding the differences between problems to discovering the similarities.

In traditional computing, solving a problem usually involves several distinct stages including defining the problem, selecting the methodology and algorithm, coding the algorithm, and performing the computation. Computation usually receives the most attention, followed by coding. The development of parallel and high-speed machines can help alleviate the computational load, but that in itself can't translate into high system performance without corresponding work on coding and problem analysis and definition.

Biological systems, which we recognize to be true self-organizing systems, tend to integrate the different stages of problem solving in a single system. Thus, the time-consuming and poorly understood stages of precise problem definition and

methodology/algorithm selection may be replaced by the establishment of general boundary conditions on the data and the interactions between decision units, followed by a loose description of the learning, communication and decision-making dynamics. Complex robotic movements may be achieved by "learning by doing" instead of by detailed numerical modeling and algorithms for simultaneous constraint satisfaction. Similarly, the problem of knowledge acquisition, representation, updating, and retrieval may be accomplished as a natural outgrowth of the rules of interaction governing the basic units of neurons and synapses.

Neural net research is a risky venture. The human brain presents a formidable challenge and, while an understanding of an isolated process may be discovered, its significance to overall brain function remains a mystery. In this respect, neurophysiology is like psychology. If you're looking to justify why a particular mechanism occurs, one can probably be constructed using a little imagination and a large data base. Unfortunately, no known methods exist to reliably test theories of this type, and barring a revolutionary advance in experimental neuroscience, they will remain in our imaginations. Similarly, neural modeling, done by writing down nonlinear, coupled, dynamical equations to solve a specific problem without thinking about how they would be implemented, postpones and compounds the severity of an unavoidable collision with reality.

Considering the slow progress made by neuroscience and computer science in comprehending the depth of poorly defined problems, the most viable approach to an advance in neural net-based computational systems is judicious modeling, combined with an awareness of the potential capabilities of biological and AI systems.

h a sequence
mized. Hop-
with neurons
n are labeled
cities on the
locations of
de in row A,
hen the sys-
fif" city in
s wh Hop-
syntax-rein-

forcing connections." Each node must first be capable of directly stimulating every other node. Also, each node must also have connections to inhibit other nodes. In this example, if city A, position 5 node is going to be "on" at the end of the problem, all other nodes in row A and column 5 must be suppressed somehow, requiring another set of connections. Another aspect of the connectivity syntax requires a way to represent the distances between the cities. This is done by adding resistances into the connection pattern that correspond to these distances.

It's important to note that the circuit is operated in an analog range. At the beginning, all neurons are in a nonzero, low-energy state. For a small number of cities, the circuit rapidly computes the right answer. When the number of cities is increased, the processing time remains about the same, and the circuit settles on a set of the best answers. A 30-city tour, for example, requires 900 neurons and has 10^{30} possible tours. The neural circuit is able to find the 107 best solutions in a few time constants or in about $1 \mu s$. This represents a selection factor of 10^{23} , according to Hopfield.



*Howard C. Anderson
Senior Software Engineer
Motorola*

Pattern recognition by filtered Fourier transforms

It's no surprise that our computers don't understand us yet—we've been trying to communicate with them via an inferior medium language. Most of the input to a normal human brain is through the visual system. Far less data is received through the auditory channel. Since most biological systems are designed for utmost efficiency, expecting auditory-based lingual information to be the most important component of human thought processes seems inconsistent with evolution.

The brain's ability to perform pattern-recognition tasks sets it apart from machines of von Neumann architecture. The filtered Fourier transform pattern-recognition technique is representative of early attempts to understand and simulate these abilities on von Neumann machines.

In 1966, Matthew Kabrisky of the Air Force Institute of Technology (W-Patterson AFB, OH) published the book, "A Proposed Model for Visual Information Processing in the Human Brain." In subsequent works, Kabrisky and several of his students investigated the possibility that two-dimensional filtered Fourier transforms are involved in the computational processes that occur in the human brain.

In 1967, Radoy, one of Kabrisky's students, demonstrated a pattern-recognition system that can recognize alphabetic characters. In essence, such a system overlays a pattern with a grid, extracts the brightness of the grid squares, enters those brightness values in a complex-pattern matrix and calculates a discrete two-dimensional Fourier

transform, which is also a complex matrix of the same order as the pattern matrix. It then stores the transform matrix. Pattern recognition is performed by saving the transform matrices of various patterns and then comparing the transform of a new pattern with the transforms of the stored patterns. A new pattern is recognized as the stored pattern whose transform is most closely matched with the transform of the new pattern. This operation is done by calculating the Euclidean distance between the transforms.

Radoy found that ignoring the terms in the transform matrix that were associated with high-frequency components only minimally affected recognition of alphabetic characters. Using a technique known as low-pass spatial filtering, he reduced storage requirements of pattern transforms by a factor of 100 without seriously degrading the machine's ability to recognize patterns.

In 1969, Tallman, another of Kabrisky's students, experimented with hand-printed samples of all 26 alphabetic characters from 25 different people. By using the filtered Fourier transform technique, Tallman was able to achieve a 95 percent recognition rate for the set of 650 characters.

Kabrisky has pointed out that written characters, whether arabic numerals or Chinese Kanji characters, evolved so that they are distinguishable by people. The filtered Fourier transform technique seems to identify the essence of a character—that which distinguishes it from other characters.

For an even better set of solutions, a technique known as annealing has been proposed. Annealing lets the system reach one stable state and then energizes it to find an even lower global energy level. This is similar to heating a crystalline structure to some temperature and then cooling to produce a more perfect crystalline pattern.

The fact that the neural net doesn't pick out one single best answer (often there are several) also fits the analog or "fuzzy" nature of deciding among conflicting solutions. Nevertheless, the network is able to consider the solutions simultaneously. By comparison, a typical microcomputer can find a comparably good solution in about 0.1 s, according to Hopfield. But the microcomputer has about 10^4 times as many devices as neural net.

This reveals several facts, the most significant being the role the connectivity pattern plays in representing the data, the problem and the program for the solution. The pattern of the interconnections programmed the system, and the use of the network to find a global minimum depends on the pattern of interconnection. In addition, the solution in this example is a static state, whereas the brain operates in real time with ever-changing input. Nevertheless, the neural net arrives at an acceptably correct solution and demonstrates the feasibility of a neural-net approach for solving highly complex problems.

Brain researchers know that if a synapse isn't used, it eventually disappears, and that synaptic connections that are repeatedly stimulated are strengthened. As a result, the brain not only repre-

matrix of the
n stores the
is performed
various pat-
orm of a new
red patterns,
ored pattern
hed with the
operation is
distance be-

s in the trans-
l with high-
ally affected
Using a tech-
filtering, he
pattern trans-
ously degrad-
e patterns.

briskv's stud-
ed sr'es of
diffe' peo-
nsfo' tech-
a 95 percent
characters.

ritten charac-
Chinese Kanji
are distin-
Fourier trans-
essence of a
it from other

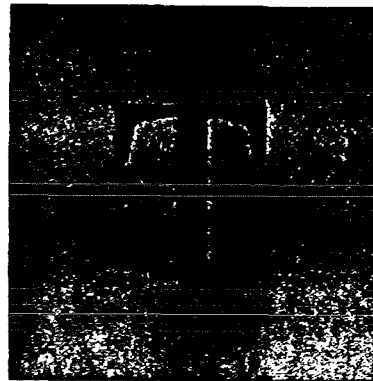
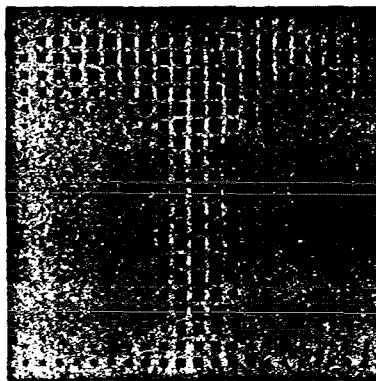
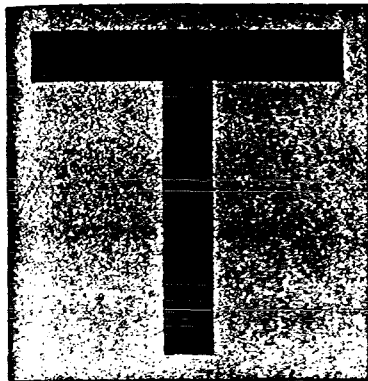
significant be-
lays in repre-
program for
rconnections
f the network
he pattern of
ion in this ex-
in operates in
Nevertheless,
correct solu-
f a neural-net
problems.
synapse isn't
tha' naptic
imul d are
ot only repre-

For example, note the T pattern in the figures below. Intrinsic brightness of the elements of this pattern is indicated by the size of the dark squares that make up the image. Negative values (those that appear in the inverse transforms) are shown by dashes of various lengths. Longer dashes indicate more negative values. If you take the Fourier transform of the pattern of the first image, and then take the Fourier transform of the transform, you will get the original pattern—that is, Fourier transforms are invertible. If you filter (eliminate high-frequency terms) the Fourier transform of the T before inverting it, you will produce the middle image: the 5x5 filtered inverse transform. It's interesting that a pedestal forms at the base of the T in the filtered inverse transform and that serifs form at the ends of the horizontal bar. Compare

this with the Triplex Roman T of the Hershey font set shown in the third figure.

Is it possible that the serifs and pedestal came into vogue in printer font sets because that's the form of the most distinguishable T? Some think that it's the most aesthetically pleasing form. Does the concept "aesthetically pleasing" derive from peculiarities of our internal image processors?

Whether actual Fourier transform processes are occurring in the brain remains a matter of speculation. In any case, in terms of speed, the von Neumann architecture does not seem to be the appropriate architecture for simulating the brain's pattern-recognition processes. Neural-net machines demonstrating self-organization of memory seem to be on the right track. These new machines will help to revolutionize the computer industry.



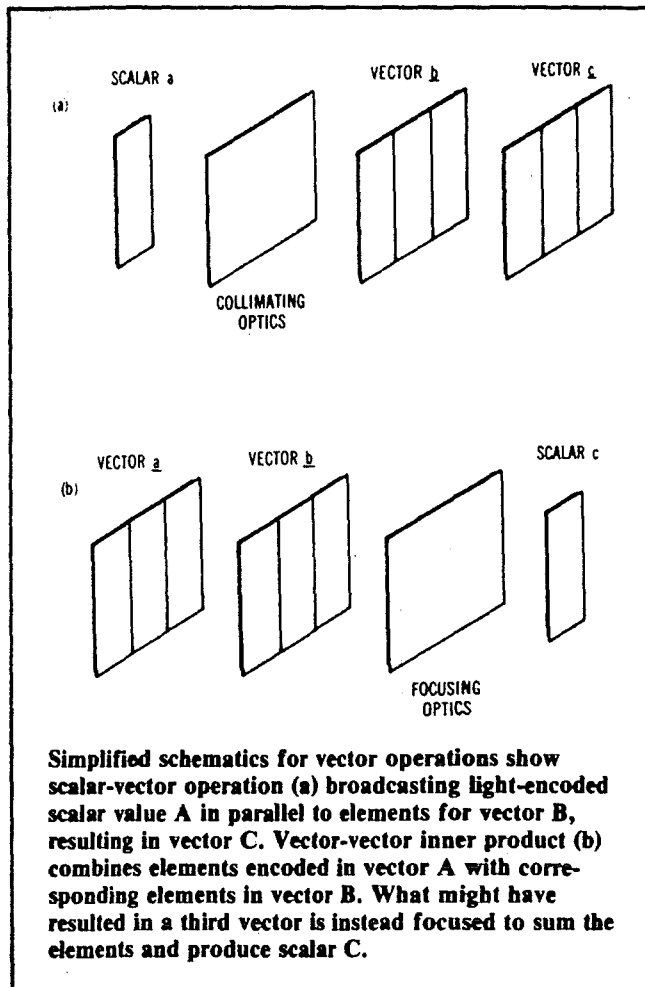
sents information by the strength of stimulation and neuron firing in existing nodes and connections, but it's also constantly reconfiguring its connections. In addition to implementing the incredibly dense system of interconnections required in a practical neural-net system, a method will have to be developed to dynamically reconfigure it.

This requirement has led some researchers to look to the inherent communications capabilities of light and optics. As Darpa's Neff points out, "The inherent advantage of photons over electrons in communications is that you can have a high number of channels very close together, and they won't interfere with each other the way that electrons do. In addition, a single light source can broadcast to millions of points simultaneously." Optics also lend

themselves to other types of computation tasks—such as those involving a high degree of parallelism.

There is still a good deal of discussion among neural-net researchers as to how much can be accomplished using existing silicon technology. In fact, in the near future, the first practical neuron-like circuits will appear in silicon. Synaptics' Mead has demonstrated a silicon model of part of the human retina which he developed at the California Institute of Technology. "The retina is one of the best studied parts of the brain because it's out there," he says. Because it's isolated from the rest of the brain, researchers have the advantage of knowing exactly what the input to the retina is, and they can positively isolate its outputs, such as optic nerves.

(continued on page 58)



Optics...

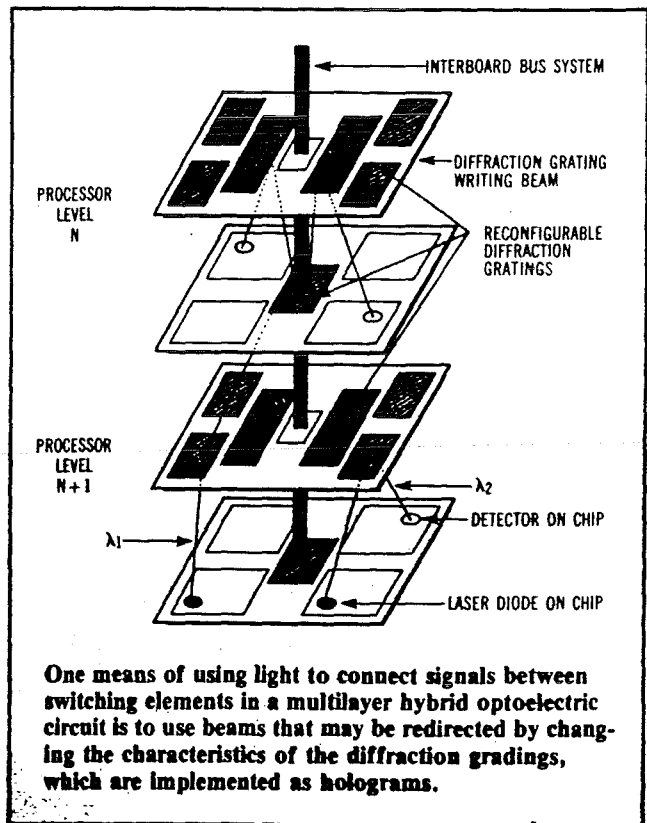
(continued from page 55)

Mead's model takes advantage of the fact that the cones in the eye are stimulated by light, and that they have outputs that feed back onto them and inhibit stimulus in somewhat the same manner as in the Hopfield model. The eye responds to the changes in light value rather than to absolute light intensity. The cones and Mead's CMOS photodetectors output a time derivative of intensity on a logarithmic scale rather than a linear one. Another layer, the amacrine layer, computes a spatial derivative of the time derivative provided by the sensors. The amacrine cells provide a passive resistive network that modifies the output of neighboring cells/detectors.

The silicon retina, like the natural one, relies on rates of change to detect moving objects, and it can do so in real time, unlike digital computers. The human eye is constantly undergoing minute motions to create the images we see. If that motion were to stop, the time-dependent rate-of-change computations would cease, and the image would fade. The silicon retina responds in the same way.

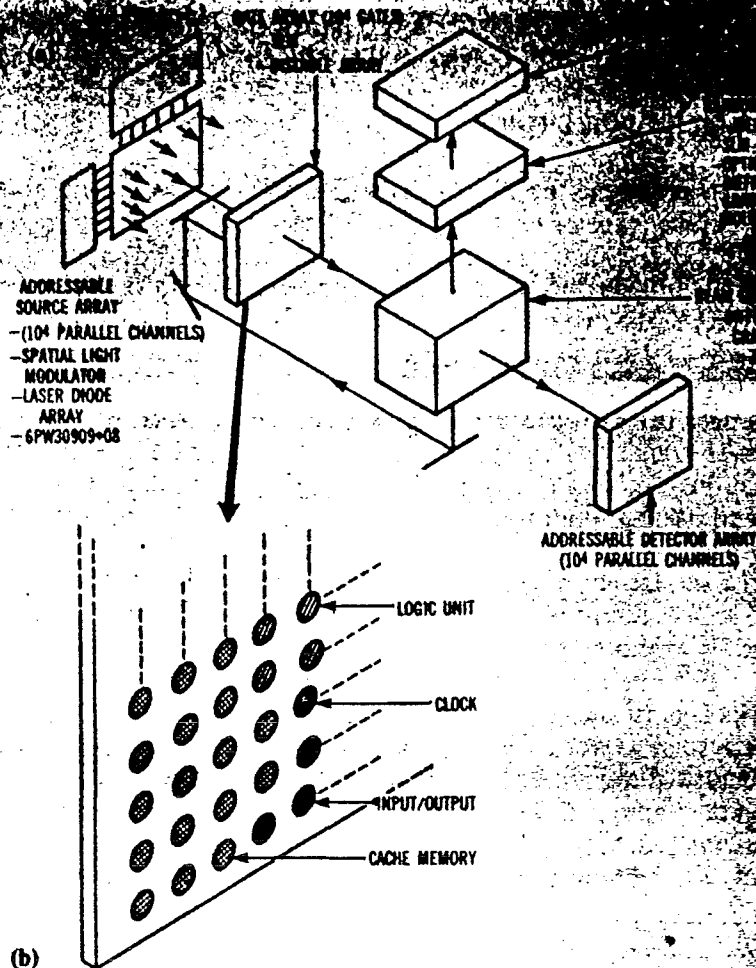
As mentioned earlier, work on optical bistable devices is making progress. Hitachi (Tokyo, Japan) has developed an optical switch that can switch between two channels at 833 MHz, and spatial light modulators exist that operate in a nonlinear, or binary, mode. It is also true that optical bistable devices will probably find more immediate applications in real-world designs—especially in telecommunications—than some of the proposed optical systems discussed here. In computer systems, there's a branch of research that's looking into using optics to communicate among circuit boards as well as among VLSI components on the same board. One problem that's badly in need of a solution is clock skew between high-speed components. Optics are also being considered for this field. Still, the possibilities of using the high bandwidth and global communications abilities of light to aid in the neural-network class of future computers is also a lively area of research, although it's still very much confined to the university and the laboratory level. In addition, the use of light for all-optical processors is receiving serious attention.

The potential computational power of optical processors can be shown by the use of two-dimensional spatial light modulators (SLMs) in matrix processing. Matrix operations with light take advantage of the inherent global communications and the ability to easily integrate intensities to



al bistable
yo, Japan)
sw. be-
patia ht
nlinear, or
bistable de-
te applica-
ly in tele-
oposed op-
puter sys-
t's looking
ong circuit
ents on the
n need of a
eed compo-
ed for this
high band-
ties of light
uture com-
lthough it's
ity and the
light for all-
attention.
of optical
of two-di-
(SLMs) in
with light
om ica-
tens s to

A proposal for a digital all-optic computer shows beams from a source array directed at a bistable switching array whose output is directed to both memory and to an output array (a). A feedback loop with mirrors allows input to the bistable array to be altered. Bistable elements of the array (b) can be logically grouped to form functional processor components. Altering the configuration via the beam controller would alter the system architecture.



D BUS SYSTEM

CTION GRATING
VG BEAM

RECONFIGURABLE
FRACITION
RATINGS

ECTOR ON CHIP

DIODE ON CHIP

between
toelectric
l by chang-
dir

arrive at a sum—an operation usually tedious for von Neumann machines. The optical implementation of elementary matrix operations has been described by Ravindra Athale, an optical computing manager at BDM Corp (McLean, VA).

The simplest such operation is the scalar-vector multiplication in which a single scalar element A multiplies each element of a vector of N elements. To perform this operation optically, the scalar would be represented by the light output of some source, such as a laser diode or light-emitting diode. The vector to be multiplied would be encoded on a one-dimensional SLM of the type that outputs the product of the encoded and the input light values. The output of the scalar passes through optics that broadcast it in parallel to all elements of the vector at once. The output of the SLM representing the elements of the resulting vector is then imaged in parallel onto an N -element detector array. In this manner, arrays of any complexity that can be supported by the optical hardware can be multiplied by a scalar in one operation.

Another operation involves multiplying two vectors element-by-element to arrive at an inner product, which is the output scalar. Here the output light from each element of the first vector is sent only to the corresponding element of the second vector encoded on the SLM. The outputs of each element of the SLM are then summed by focusing them onto the single detector, which represents the resulting scalar value.

These two simple examples can be increased in complexity to produce vector-matrix and matrix-matrix multipliers and even higher order functions. Just as different types of optics are needed for different types of operations, different types of SLMs and detectors may be used for different purposes. For instance, a time-integrating detector array can be used to sum results. Such an array holds the input of one element and adds the weight of the next as indicated by the intensity of light. The result is the sum of successive elements. Another type of detector receives the sum of several elements simultaneously as a focused beam. Obviously, the kind of

matrix processors described here show the potential of the technology rather than describing practical systems. One challenge is how to design hardware that can handle the demands placed on it by the different computation tasks.

Also important is the fact that the examples were analog designs. The value of each element as vector or matrix was coded in the SLM as an analog intensity value via the transmittance function of one of the SLM's cells. Since the output result—the transmitted light—has an amplitude proportional to the product of the two numbers encoded as light, the accuracy of such values depends on the accuracy of the SLM. The SLM must have uniformly linear characteristics to accurately represent numerical values as gradations of light. It must be recognized from the start that this type of optical system doesn't lend itself to high numerical precision.

It's possible to construct SLMs with nonlinear characteristics that can represent digital numbers. In such a device, each cell would be a 1 or a 0, and multiple cells would be grouped to represent bytes or words. Encoding data as binary numbers, however, works against the efforts toward parallelism, forces the system to work with digital logic and introduces many of the repetitive operations that the parallel optical approach hopes to avoid.

Considering that there are devices such as SLMs that allow combinatorial operations using light,

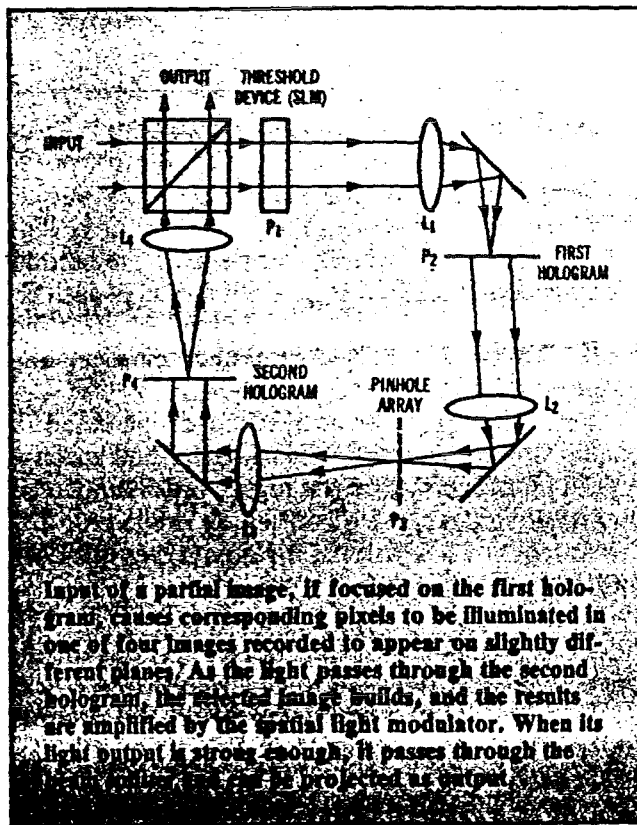
and that the analog nature of these devices fits well with the distributed and "fuzzy" character of neural processing, how might optics be used in the service of the neural model? According to Darpa's Neff, the communications abilities of optics could be used in hybrid optoelectronic systems that emphasize the need for reconfigurable connectivity between switching elements, such as between laser diodes and detectors. "As the emphasis on switching decreases, the emphasis on connectivity rises, and optics becomes more of an option," he notes.

One proposed hybrid scheme involves layers of hybrid optoelectronic chips containing laser diodes and detectors and a system of reconfigurable diffraction gratings that act as frequency-selectable filters to pass and/or direct the various beams containing data to appropriate places on different layers of circuit boards. In Neff's proposal, the diffraction gratings are holograms created by mixing waves of four different frequencies. Diffraction-grating writing beams would also be used to change the characteristics of the hologram to redirect the switching beams.

The next step, suggests Neff, might be an all-optical computer in which an optical source array (such as an SLM or a laser diode array) acts upon a processing array, which could also be a type of SLM or an array of optical bistable devices. In the latter case, the bistable devices would give it a more digital character, since they would act as logic gates. If bistable gates were used, they could be grouped to make up processing elements such as ALUs, shift registers, clock signals and so forth.

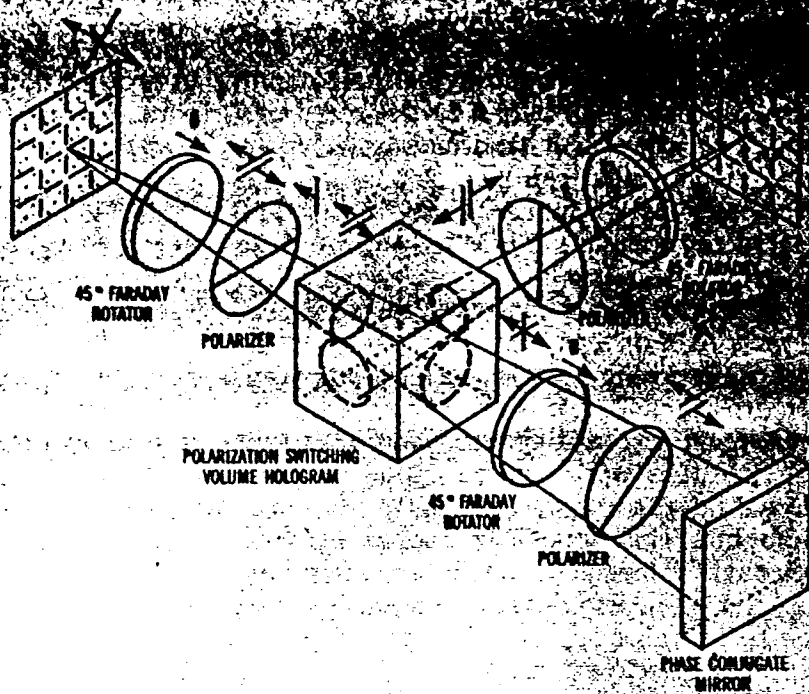
Even if such a machine were implemented in a bistable mode, there would still be great emphasis on the configuration of connectivity, since rearranging the connections would redefine not only the functional logic elements but also the entire architecture of the system. The critical element used to control the interconnection would be some kind of beam controller, such as a large diffraction grating, that could be programmed for the desired interconnects. This controller would also interact with the processing unit via a feedback loop to the input side of the array. Neff stresses that no one has built such a computer, but, he says, "It's technically believable to achieve such a system consisting of 1 million parallel channels."

If this scale of parallelism is possible in a bistable system, what about an analog machine with global connectivity? Numbers vary, but California Institute of Technology's Psaltis envisions arbitrarily connecting 10^4 neurons, which would translate to 10^8 connections in which each neuron could connect directly to every other neuron. Making that



fits well
 acted of
 ed
 Dar
 ics could
 that em-
 tivity be-
 een laser
 n switch-
 ity rises,
 he notes.
 layers of
 er diodes
 able dif-
 eelectable
 ams con-
 different
 l, the dif-
 y mixing
 fraction-
 to change
 direct the

A proposal for a backward error propagation learning network shows that a hologram recorded in the photorefractive crystal directs the input beams to selected points in the output array. If this doesn't correspond to the desired pattern, processing in the output array sends an error signal to the phase conjugate mirror assembly, altering the hologram. As the output pattern converges with the desired pattern, the error signal diminishes.



an all-op-
 erce array
 acts upon
 a type of
 es. he
 it a re
 as logic
 could be
 s such as
 so forth.
 ed in a bi-
 phasis on
 arranging
 the func-
 chitecture
 to control
 l of beam
 ating, that
 intercon-
 t with the
 input side
 built such
 ally believ-
 f 1 million
 a bistable
 with global
 rnia Insti-
 arbitrarily
 can to
 could on-
 aking that

connectivity programmable on such a scale requires techniques that aren't yet understood.

It's possible to build simple associative memories using SLM and detector arrays, but the high connectivity envisioned by researchers such as Neff and Psaltis requires some kind of medium with many more resolvable spots in a given area or volume than today's SLMs. Such resolvable spots would be used to refract and redirect individual beams of light to make or break connections between neurons. Two candidates that have been suggested are magneto-optic surfaces and photorefractive crystals.

Those searching for erasable optical media in optical disks are looking closely at magneto-optics, and when a solution is found, the very dense bit pattern on optical disks can be reconfigured. Such disks or surfaces implemented with the same technology used in magneto-optical disks could theoretically be used to specify the connections between several thousand lasers and detectors, according to Psaltis. Thus, the optical disk, which is currently used as if it were merely a denser form of magnetic media, might be used to its fuller potential in optical/neural systems.

But to truly achieve massive connectivity and dynamic reconfigurability, Psaltis suggests holography using photorefractive crystals. This is especially interesting because the global distributed manner in which the brain stores and processes in-

formation has often been compared to a hologram. One of the most striking characteristics of holograms is that the stored image can be reconstructed from only part of the hologram. This has led researchers to look into implementing associative memories using holograms—a field that looks promising. Further, recording connections as a hologram in a photorefractive crystal increases the number of possible connections by virtue of being in three dimensions, and makes the connections programmable.

Light in a photorefractive crystal generates free charges that are eventually trapped in a pattern similar to the intensity pattern of the incoming light. The spatially varying charge density that results creates internal fields that change the index of refraction within the crystal and produces in the hologram. When light shines into the crystal, it's refracted in directions determined by these varying refraction indices, giving the image of the hologram. In the case of our hypothetical computer, the image represents the pattern of interconnects. And that pattern—the hologram—can be modified by light from a feedback loop, giving the desired dynamic reconfigurability.

In the Hopfield model, the program as well as the information is stored in the communications network in a neural computer, and to be at all flexible, the system must be quickly reconfigurable. As a result, some form of intelligence is needed to process the information received and to determine how

to adapt its configuration pattern to the ongoing process of real-time computation. The system needs the ability to learn.

Some experiments have shown that holography can correlate matrices of output devices and detectors. One such experiment has demonstrated an associative memory that can pick out one of four recorded pictures of human faces, given only a partial picture as input. Fourier transforms of the images are stored in two holograms. Each image is recorded at different spatial frequencies so that they appear to be on separate planes. Partial-image data for the desired image is shined through a beam splitter into a loop formed with the holograms. The first hologram acts as a detector, and its output causes the holographic representation of only the selected image to begin to appear from the second hologram. This output is fed back into the system through a threshold device, an amplifying SLM. After a few iterations, the output is strong enough to pass through the beam splitter and be projected as the selected image.

An interesting extension of implementing connectivity via holograms has been proposed by Psaltis and a graduate student, Kelvin Wagner. The system, which is called a backward error propagation (BEP) learning network, would use holograms in photorefractive crystals to make connections between an input array and an output array. But processing in the output array—or in subsequent neural layers associated with it—would generate error signals based on a desired connection pattern.

The output array would also need the ability to send signals back to the photorefractive crystal and then to a system of polarizers and a phase conjugate mirror, which would adjust the phase of the error signal to alter the hologram in the direction of the image, producing the desired connectivity. Such an error signal could be continuous or pulsed, but would die away as the forward signal approached the desired connection pattern.

The models described here are representative of a wide range of research going on in both neural networks and in optical computing. None of them represents a practical working computer system and even those that have been implemented are experiments to prove principles and test hypotheses.

There is a realization that the need for a so-called "new paradigm" for neural-like computing systems carries with it the need for a new approach to the information science describing such machines. Neff, Psaltis, Hopfield, Mead and Faggin all caution that the idea of neural networks and learning systems doesn't imply a heterogeneous "mush" of

infinitely replicated and interconnected neurons. Just as the brain is highly structured, these new systems will need a structure and hierarchy as well as an organizational basis to determine how they will learn, how they will preprocess and select input information, and how different parts of such intelligent systems will perform specific functions.

This is a science still in its most rudimentary stages. It will build in a kind of feedback loop as people try to solve relatively specialized problems using neural models and learn from their experiences. "The nervous system is based on a set of organizing principles different from any computational paradigm we know," says Mead. The process of understanding that paradigm, he argues, must start from the bottom up. Neural "primitives" are computationally powerful in their own right and include such things as exponential functions and integration with respect to time. "At the bottom level, the power of neural networks comes from the fact that they don't insist on taking a beautiful thing that creates an exponential and turning it into a 1 or a 0," Mead says. "They take what is there and use it."

Although building a working model of the brain is still a distant dream, using neural network models to perform certain special tasks is within reach. As applications are found, techniques will be developed and neurobiologists will take advantage of neural models just as computer scientists learn from neurobiology in an ongoing cooperative research effort. As for a working brain, it may be far off, but it's not impossible. As Lee Giles, program manager for the Air Force Office of Scientific Research at Bolling Air Force Base (Washington, DC) notes, "We have existing proof—us!" **CD**

Please rate the value of this article to you by circling the appropriate number in the "Editorial Score Box" on the Inquiry Card.

High 261

Average 262

Low 263

Optical information processing based on an associative-memory model of neural nets with thresholding and feedback

Demetri Psaltis and Nabil Farhat*

Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125

Received July 9, 1984; accepted November 15, 1984

The remarkable collective computational properties of the Hopfield model for neural networks [Proc. Nat. Acad. Sci. USA 79, 2554 (1982)] are reviewed. These include recognition from partial input, robustness, and error-correction capability. Features of the model that make its optical implementation attractive are discussed, and specific optical implementation schemes are given.

Optical information-processing systems can have high processing power because of the large degree of parallelism as well as the interconnection capability that is achievable. Typically, more than 10^6 parallel processing channels are available in the optical system, and furthermore each of these channels can be optically interconnected (broadcasted) to 10^6 other channels. The majority of optical processors are analog systems, designed to perform linear operations. The accuracy of an analog processor is limited by the linear dynamic range of the devices used (detectors, light modulators). In principle, the accuracy and the repertoire of achievable operations can be improved with systems that perform nonlinear operations on binary encoded data using bistable optical devices. Optical bistability is a subject that has received considerable attention recently as a means of achieving efficient high-speed logic, and it has been demonstrated with several nonlinear optical materials and devices. If we are to use such bistable devices to realize powerful, nonlinear optical computers, it is important to find algorithms that are well matched to the characteristics of the optical processor and utilize effectively its parallelism and interconnection capability. In this Letter we examine a method for synthesizing optical processing systems, based on optical associative memory and threshold logic, that appears to meet these requirements well.

Associative (or content-addressable) memories are of interest in computer science, and it is theorized that information is stored in the human brain in this manner. Holographic associative memories have been described by Gabor,¹ who also commented on the similarity of the holographic memory to the way information may be stored in the human brain. More recently, Hopfield² introduced an associative-memory model to describe the collective behavior of neural networks. Hopfield's model consists basically of an associative memory similar to the holographic, with the addition of threshold and feedback. The incorporation of nonlinear feedback enhances dramatically the error-correcting capability of the holographic memory.

Let $v_i^{(m)}$ be a binary word that is N bits long. M such words are stored in a matrix T_{ij} according to

$$T_{ij} = \begin{cases} \sum_m [2v_i^{(m)} - 1][2v_j^{(m)} - 1] & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad (1)$$

If T_{ij} is multiplied by one of the stored binary vectors $v_i^{(m)}$, the product $\hat{v}_i^{(m)}$ is an estimate of the stored vector $[2v_i^{(m)} - 1]$:

$$\hat{v}_i^{(m)} = \sum_j T_{ij} v_j^{(m)} = N_0 [2v_i^{(m)} - 1] + \sum_{m \neq m_0} \left\{ \sum_j [2v_j^{(m)} - 1] v_j^{(m)} \right\} \times [2v_i^{(m)} - 1] - M v_i^{(m)}, \quad (2)$$

where the last term accounts for $T_{ij} = 0$ and N_0 is the number of 1's in $v_i^{(m)}$. We assume that for $m \neq m_0$ the binary words $v_i^{(m)}$ are statistically described in the following simple manner:

$$P[v_i^{(m)} = 1] = 1/2, \quad P[v_i^{(m)} = 0] = 1/2, \quad (3)$$

where $v_i^{(m)}$ are independent for all i and m . Then $E[\hat{v}_i^{(m)}] = (N/2)[2v_i^{(m)} - 1]$ and $\text{var}[\hat{v}_i^{(m)}] = N(M - 1)/2$. We define the signal-to-noise ratio (SNR) of the estimate $\hat{v}_i^{(m)}$ as the ratio of the magnitude of the expected value of $\hat{v}_i^{(m)}$ to the standard deviation of the estimate:

$$\text{SNR} = \frac{|E[\hat{v}_i^{(m)}]|}{\{\text{var}[\hat{v}_i^{(m)}]\}^{1/2}} = [N/2(M - 1)]^{1/2}. \quad (4)$$

If N is sufficiently larger than M , then with high probability $TH[\hat{v}_i^{(m)}] = v_i^{(m)}$, where $TH[\hat{v}_i^{(m)}] = 1$ if $\hat{v}_i^{(m)} > 0$ and zero otherwise. Thus the vector-matrix product in Eq. (2) combined with the thresholding operation results in a pseudoeigensystem in that the output vector equals the input. Now suppose that the full vector $v_i^{(m)}$ is in fact such a pseudoeigenvector of the system but that only N_1 of N bits ($N_1 \leq N$) of $v_i^{(m)}$ are known. In this case we define an input vector consisting of the N_1 known bits, and the rest are set equal to zero. When this vector is multiplied by the matrix T_{ij} , an estimate of the complete vector $[2v_i^{(m)} - 1]$ is obtained. The SNR of the estimate is now $\text{SNR} = [N_1/2(M - 1)]^{1/2}$. If N_1 becomes sufficiently small, then, with high probability, $TH[\hat{v}_i^{(m)}] \neq v_i^{(m)}$ for some of the values of i . Let N_2 be the number of correct bits in $TH[\hat{v}_i^{(m)}]$. If $N_2 > N_1$, we can multiply this thresholded estimate by T_{ij} and obtain a new estimate with a higher SNR. This procedure can be continued until the number of correct bits in the thresholded vector is equal to N . The crucial issue is under what conditions N_2 will be higher than N_1 . If the SNR of the initial

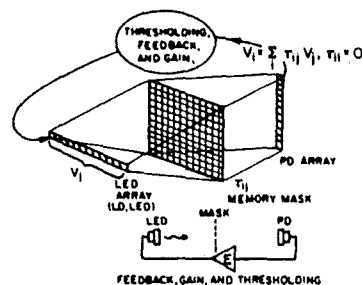
estimate $\{|N_1/(2M - 1)|^{1/2}\}$ is sufficiently large, then the probability of N_2 being bigger than N_1 will be high; in this case the nonlinear, iterative procedure described here will be likely to converge to the correct vector v . Ideally, each of the M stored binary words is a pseudoeigenvector of the nonlinear system. Notice that each pseudoeigenstate is a stable state of the system, whereas any other input vector (state) will cause a change to occur in the next cycle. In general, the system converges to the stable state that is at the shortest Hamming distance away from the initial state.

This model has been studied computationally by Hopfield.² In simulations, correct convergence was obtained reliably for $M \leq 0.15N$ and $N_1 \approx 0.75N$, taking $N = 30$. At present there is no (adequate) theoretical prediction of the maximum number of words that can be stored or the maximum Hamming distance between the input vector and one of the stored words that is required for convergence. Several interesting properties were observed. The model does not require synchronism. Convergence can be obtained if the output vector is fed back to the input as a whole or, randomly, one element at a time. There is some evidence that asynchronous operation is actually preferable. The system is quite insensitive to imperfections such as nonuniformities, the exact form of the threshold operation, and errors in the T_{ij} matrix. Convergence to the correct vector was obtained even when the T_{ij} matrix was thresholded. Such properties are most desirable when an optical implementation is considered.

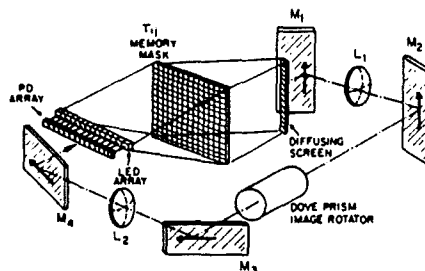
One possible optical implementation of the Hopfield model is through the arrangements shown in Fig. 1, in which the array of light-emitting diodes (LED's) represents N logic elements with binary states $v_j = 0, 1, j = 0, 1, \dots, N$ (LED on or off), which are to be interconnected in accordance with the model. This is achieved by the addition of nonlinear feedback (feedback, thresholding, and gain) to the well-known optical vector-matrix multiplier.³ Gain is included in the feedback loop to compensate for losses. Two possible feedback schemes are shown. One uses electronic wiring and the other is optical, with the thresholding (point nonlinearity) and the gain concentrated between the photodiode (PD) array and the LED array, which can be fabricated monolithically on GaAs. Furthermore, with the accelerating pace of research in thin-film nonlinear light amplifiers⁴ and optical bistable devices,⁵ it can be possible to substitute a single distributed bistable light-amplifier device for the PD/LED arrays and the intervening thresholding and amplifying electronics.

Multiplication of the vector v_j by the T_{ij} matrix in these schemes is accomplished by horizontal imaging and vertical smearing of v_j using anamorphic optics (omitted from Fig. 1 for simplicity). A bipolar T_{ij} can be realized optoelectronically with incoherent light by assigning its negative and positive values to adjacent rows. Light passing through each row is focused onto adjacent pairs of photodiodes of the PD array that are electronically connected in opposition, as shown in Fig.

Here the positive and negative elements of each row of the T_{ij} matrix are separated into two subrows, one for positive values and one for negative. The light transmitted through the two subrows is integrated horizon-



(a) ELECTRONIC FEEDBACK SCHEME



(b) OPTICAL FEEDBACK SCHEME

Fig. 1. Two schemes for adding nonlinear feedback to an optical vector-matrix multiplier utilizing (a) electronic feedback and (b) optical feedback.

tally with the aid of another set of anamorphic lenses (omitted from Figs. 1 and 2) and brought to focus on two adjacent photodiodes of the PD array connected in opposition. The output of the first diode-pair circuit will be proportional, $v_1 = \sum_j T_{1j}v_j$. This output is applied through an electronic thresholding circuit to the first element of the LED array, as shown in Fig. 1. Similar connections are made between other detector pairs of the photodetector array and corresponding elements in the LED array. Thus each LED assesses the state of its input $v_i = \sum_j T_{ij}v_j$ and fires according to whether v_i exceeds the threshold or not.

We now consider the possibility of optically storing two-dimensional (2-D) functions (images). Let $v^{(m)}(i, i')$ be the bipolar binary (1, -1) images to be stored. If we directly extend the Hopfield model to two dimensions, then these images must be stored in a four-dimensional function in the following general form:

$$T(i, i', j, j') = \sum_m^M v^{(m)}(i, i')v^{(m)}(j, j'). \quad (5)$$

In order to implement a 2-D Hopfield memory optically, we need to realize a 2-D, linear optical system whose spatial impulse response is the four-dimensional function defined in Eq. (5). Since we have only two spatial coordinates to work with in an optical system, it is difficult to implement such a system directly for the nonseparable, shift-variant kernel defined in Eq. (5). One possible solution is the use of wavelength multiplexing and/or time-domain processing to obtain additional independent variables. Another solution is based on holographic associative memories, as we have discussed earlier.⁶ Here we present an implementation based on spatial-frequency multiplexing.

The entire optical system, including nonlinear feedback, is shown in Fig. 3. The system accepts a 2-D

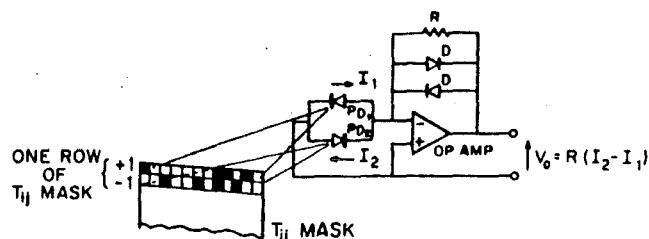


Fig. 2. Scheme for realizing bipolar mask transmittance with incoherent light.

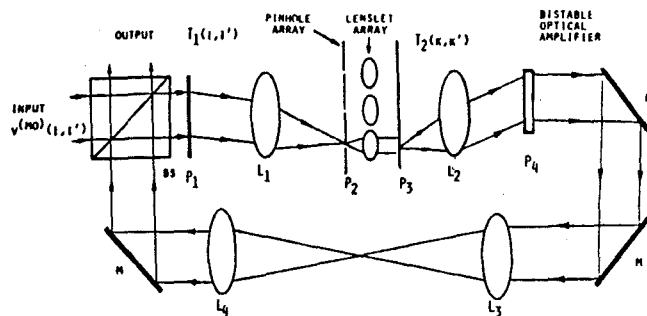


Fig. 3. Coherent optical implementation for 2-D inputs.

input $v^{(m0)}(i, i')$, which illuminates the system from the left in Fig. 3. The 2-D interconnection pattern between the planes P_1 and P_4 that is prescribed by Eq. (5) is stored on two separate optical transparencies, denoted T_1 and T_2 in Fig. 3. Each image $v^{(m)}(i, i')$ is placed on a separate spatial-frequency carrier, and the images are added to form the first transparency T_1 . When T_1 is illuminated with an input image, the products between the input and all the stored images are formed. The lens L_1 produces the Fourier transforms of all the products and displays them spatially separated at the back focal plane of L_1 (plane P_2). An array of pinholes is placed at P_2 , the position of each pinhole being at the spatial frequency of each carrier used in the recording of transparency T_1 . Therefore the amplitude of the light transmitted through each individual pinhole is proportional to the integral of the product of the input image and the corresponding image stored in T_1 . In the discrete notation used in this Letter, the amplitude of the light transmitted through the m th pinhole is proportional to $\sum \sum v^{(m0)}(i, j)v^{(m)}(i, j)$. The second transparency, $T_2(k, k')$, consists of a 2-D array of Fourier-transform holograms of all the stored images. Each hologram is formed at a separate position on the hologram, the transform of the m th image being centered at the location of the corresponding m th pinhole. Light emerging from each pinhole illuminates only the corresponding hologram. The lens L_2 takes the Fourier transform between planes P_3 and P_4 , thereby reconstructing all the images stored in T_2 . The light amplitude at P_4 is a weighted sum of all the stored images, the weights being proportional to the inner product between the input and stored images. This is precisely the desired output that is produced by the interconnection prescription given in Eq. (5). The modulation of the light at plane P_4 will in general be bipolar, and it is interferometrically detected by the nonlinear optical amplifier at P_4 , which performs the thresholding op-

eration. If interferometric detection proves to be too cumbersome, it is possible to modify the interconnection pattern such that the output is always positive. This results in a loss of storage capacity by a factor of 2, but it may be a welcome trade-off. The thresholded image is fed back to the input through mirrors and imaging optics (lenses L_3 and L_4). The output of the system is taken at the beam splitter. Optical gain must be included in the cavity (preferably through the bistable optical element) to compensate for losses that are due to the passive components. The requirement that $T(i, i', j, j') = 0$ for $i = i', j = j'$ must also be satisfied because otherwise the diagonal elements always become equal to M , whereas the off-diagonal elements have an average value of \sqrt{M} . The result is that for large M the system becomes an imaging system; any input is replicated at the output. This is avoided by forming each of the Fourier-transform holograms in P_3 with a randomly chosen, uniform phase.

We have described several specific optical implementations of the Hopfield model; undoubtedly others are also possible. The most important feature of all such implementations is the robustness of a system that utilizes nonlinear feedback. The systems that we have described behave basically as associative memories (the whole is retrieved from a partial input), even with open-loop operation. However, the nonlinear feedback can correct errors of the open-loop system since it forces the state of the system to change continuously until a stable condition is reached. The nonlinearity plays a crucial role; if linear feedback were used, the system would either be unstable or converge to the eigenstate of the open-loop system with the highest eigenvalue, independently of the initial condition.

This error-correcting capability can provide the accuracy that is lacking from analog optical processors without, however, sacrificing the processing power that can be derived from the global processing capability of optics; the class of processors that we described are fully interconnected optical systems and hence utilize fully the parallelism and the interconnectivity capability of optics. In general, there is an excellent match between the global, linear operations and local, point nonlinearities that are required for the implementation of the Hopfield model, and the capabilities and limitations of optical techniques.

The authors thank John Hong and Yaser Abu-Mostafa for many helpful discussions on this subject.

* On scholarly leave from the University of Pennsylvania, Philadelphia, Pennsylvania 19104.

References

1. D. Gabor, IBM J. Res. Dev. **13**, 156 (1969).
2. J. J. Hopfield, Proc. Nat. Acad. Sci. USA **79**, 2554 (1982).
3. J. Goodman, A. R. Dias, and I. M. Woody, Opt. Lett. **2**, 1 (1978).
4. Z. Porada, Thin Solid Films **109**, 213 (1983).
5. H. M. Gibbs, S. L. McCall, and T. N. C. Venkatesan, Opt. News **5**(3), 6 (1979).
6. D. Psaltis and N. Farhat, presented at the Meeting of the International Commission for Optics, ICO-13, Sapporo, Japan, August 1984.

Optical implementation of the Hopfield model

Nabil H. Farhat, Demetri Psaltis, Aluizio Prata, and Eung Paek

Optical implementation of content addressable associative memory based on the Hopfield model for neural networks and on the addition of nonlinear iterative feedback to a vector-matrix multiplier is described. Numerical and experimental results presented show that the approach is capable of introducing accuracy and robustness to optical processing while maintaining the traditional advantages of optics, namely, parallelism and massive interconnection capability. Moreover a potentially useful link between neural processing and optics that can be of interest in pattern recognition and machine vision is established.

I. Introduction

It is well known that neural networks in the eye-brain system process information in parallel with the aid of large numbers of simple interconnected processing elements, the neurons. It is also known that the system is very adept at recognition and recall from partial information and has remarkable error correction capabilities.

Recently Hopfield described a simple model¹ for the operation of neural networks. The action of individual neurons is modeled as a thresholding operation and information is stored in the interconnections among the neurons. Computation is performed by setting the state (on or off) of some of the neurons according to an external stimulus and, with the interconnections set according to the recipe that Hopfield prescribed, the state of all neurons that are interconnected to those that are externally stimulated spontaneously converges to the stored pattern that is most similar to the external input. The basic operation performed is a nearest-neighbor search, a fundamental operation for pattern recognition, associative memory, and error correction. A remarkable property of the model is that powerful global computation is performed with very simple, identical logic elements (the neurons). The interconnections provide the computation power to these simple logic elements and also enhance dramatically the stor-

age capacity; approximately $N/4 \ln N$ bits/neuron can be stored in a network in which each neuron is connected to N others.² Another important feature is that synchronization among the parallel computing elements is not required, making concurrent, distributed processing feasible in a massively parallel structure. Finally, the model is insensitive to local imperfections such as variations in the threshold level of individual neurons or the weights of the interconnections.

Given these characteristics we were motivated to investigate the feasibility of implementing optical information processing and storage systems that are based on this and other similar models of associative memory.^{3,4} Optical techniques offer an effective means for the implementation of programmable global interconnections of very large numbers of identical parallel logic elements. In addition, emerging optical technologies such as 2-D spatial light modulators, optical bistability, and thin-film optical amplifiers appear to be very well suited for performing the thresholding operation that is necessary for the implementation of the model.

The principle of the Hopfield model and its implications in optical information processing have been discussed earlier.^{5,6} Here we review briefly the main features of the model, give as an example the results of a numerical simulation, describe schemes for its optical implementation, then present experimental results obtained with one of the schemes and discuss their implications as a content addressable associative memory (CAM).

II. Hopfield Model

Given a set of M bipolar, binary (1, -1) vectors $\mathbf{v}_i^{(m)}$, $i = 1, 2, 3 \dots N$, $m = 1, 2, 3 \dots M$, these are stored in a synaptic matrix in accordance with the recipe

$$T_{ij} = \sum_{m=1}^M v_i^{(m)} v_j^{(m)}, \quad i, j = 1, 2, 3 \dots N, \quad T_{ii} = 0, \quad (1)$$

$\mathbf{v}_i^{(m)}$ are referred to as the nominal state vectors of the

Nabil Farhat is with University of Pennsylvania, Moore School of Electrical Engineering, Philadelphia, Pennsylvania 19104; the other authors are with California Institute of Technology, Electrical Engineering Department, Pasadena, California 91125.

Received 24 December 1984.

0003-6935/85/101469-07\$02.00/0.

© 1985 Optical Society of America.

memory. If the memory is addressed by multiplying the matrix T_{ij} with one of the state vectors, say $v_i^{(m)}$, it yields the estimate

$$\begin{aligned} \hat{v}_i^{(m)} &= \sum_j T_{ij} v_j^{(m)} \\ &= \sum_{j \neq i} \sum_m v_i^{(m)} v_j^{(m)} v_j^{(m)} \\ &= (N-1)v_i^{(m)} + \sum_{m \neq m_0} \alpha_{m,m_0} v_i^{(m)}, \end{aligned} \quad (2)$$

where

$$\alpha_{m,m_0} = \sum_j v_j^{(m)} v_j^{(m_0)}.$$

$\hat{v}_i^{(m)}$ consists of the sum of two terms: the first is the input vector amplified by $(N-1)$; the second is a linear combination of the remaining stored vectors and it represents an unwanted cross-talk term. The value of the coefficients α_{m,m_0} is equal to $\sqrt{N-1}$ on the average (the standard deviation of the sum of $N-1$ random bits), and since $(M-1)$ such coefficients are randomly added, the value of the second term will on the average be equal to $\sqrt{(M-1)(N-1)}$. If N is sufficiently larger than M , with high probability the elements of the vector $\hat{v}_i^{(m)}$ will be positive if the corresponding elements of $v_i^{(m)}$ are equal to $+1$ and negative otherwise. Thresholding of $\hat{v}_i^{(m)}$ will therefore yield $v_i^{(m)}$:

$$v_i^{(m)} = \text{sgn}[\hat{v}_i^{(m)}] = \begin{cases} +1 & \text{if } \hat{v}_i^{(m)} > 0 \\ -1 & \text{otherwise.} \end{cases} \quad (4)$$

When the memory is addressed with a binary valued vector that is not one of the stored words, the vector-matrix multiplication and thresholding operation yield an output binary valued vector which, in general, is an approximation of the stored word that is at the shortest Hamming distance from the input vector. If this output vector is fed back and used as the input to the memory, the new output is generally a more accurate version of the stored word and continued iteration converges to the correct vector.

The insertion and readout of memories described above are depicted schematically in Fig. 1. Note that in Fig. 1(b) the estimate $\hat{v}_i^{(m)}$ can be viewed as the weighted projection of T_{ij} . Recognition of an input vector that corresponds to one of the state vectors of the memory or is close to it (in the Hamming sense) is manifested by a stable state of the system. In practice unipolar binary (0,1) vectors or words $b_i^{(m)}$ of bit length N may be of interest. The above equations are then applicable with $[2b_i^{(m)} - 1]$ replacing $v_i^{(m)}$ in Eq. (1) and $b_i^{(m)}$ replacing $v_i^{(m)}$ in Eq. (2). For such vectors the SNR of the estimate $\hat{v}_i^{(m)}$ can be shown to be lower by a factor of $\sqrt{2}$.¹

An example of the T_{ij} matrix formed from four binary unipolar vectors, each being $N = 20$ bits long, is given in Fig. 2 along with the result of a numerical simulation of the process of initializing the memory matrix with a partial version of $b_i^{(4)}$ in which the first eight digits of $b_i^{(4)}$ are retained and the remainder set to zero. The Hamming distance between the initializing vector and $b_i^{(4)}$ is 6 bits and it is 9 or more bits for the other three

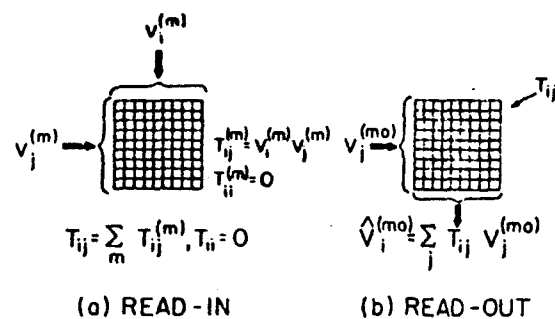
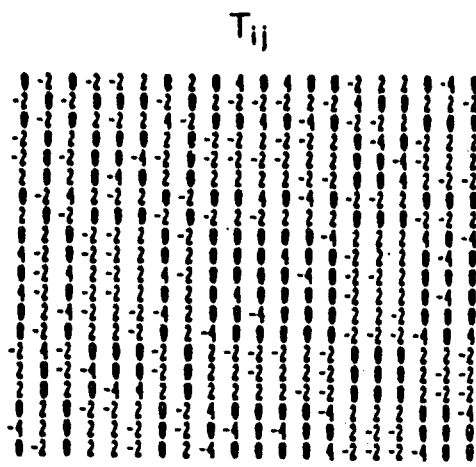
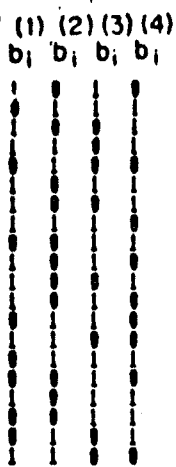


Fig. 1. (a) Insertion and (b) readout of memories.

stored vectors. It is seen that the partial input is recognized as $b_i^{(4)}$ in the third iteration and the output remains stable as $b_i^{(4)}$ thereafter. This convergence to a stable state generally persists even when the T_{ij} matrix is binarized or clipped by replacing negative elements by minus ones and positive elements by plus ones evidencing the robustness of the CAM. A binary synaptic matrix has the practical advantage of being more readily implementable with fast programmable spatial light modulators (SLM) with storage capability such as the Litton Lightmod.⁷ Such a binary matrix, implemented photographically, is utilized in the optical implementation described in Sec. III and evaluated in Sec. IV of this paper.

Several schemes for optical implementation of a CAM based on the Hopfield model have been described earlier.⁵ In one of the implementations an array of light emitting diodes (LEDs) is used to represent the logic elements or neurons of the network. Their state (on or off) can represent unipolar binary vectors such as the state vectors $b_i^{(m)}$ that are stored in the memory matrix T_{ij} . Global interconnection of the elements is realized as shown in Fig. 3(a) through the addition of nonlinear feedback (thresholding, gain, and feedback) to a conventional optical vector-matrix multiplier⁸ in which the array of LEDs represents the input vector and an array of photodiodes (PDs) is used to detect the output vector. The output is thresholded and fed back in parallel to drive the corresponding elements of the LED array. Multiplication of the input vector by the T_{ij} matrix is achieved by horizontal imaging and vertical smearing of the input vector that is displayed by the LEDs on the plane of the T_{ij} mask [by means of an anamorphic lens system omitted from Fig. 3(a) for simplicity]. A second anamorphic lens system (also not shown) is used to collect the light emerging from each row of the T_{ij} mask on individual photosites of the PD array. A bipolar T_{ij} matrix is realized in incoherent light by dividing each row of the T_{ij} matrix into two subrows, one for positive and one for negative values and bringing the light emerging from each subrow to focus on two adjacent photosites of the PD array that are electrically connected in opposition as depicted in Fig. 3(b). In the system shown in Fig. 3(a), feedback is achieved by electronic wiring. It is possible and preferable to dispose of electronic wiring altogether and replace it by optical feedback. This can be achieved by combining the PD and LED arrays in a single compact hybrid or



(a)

(b)

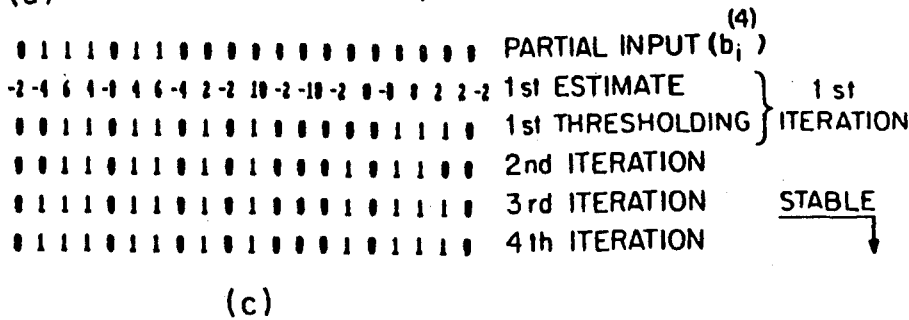


Fig. 2. Numerical example of recovery from partial input; $N = 20$, $M = 4$. (a) Stored vectors, (b) memory or (synaptic) matrix, (c) results of initializing with a partial version of $b_i^{(4)}$.

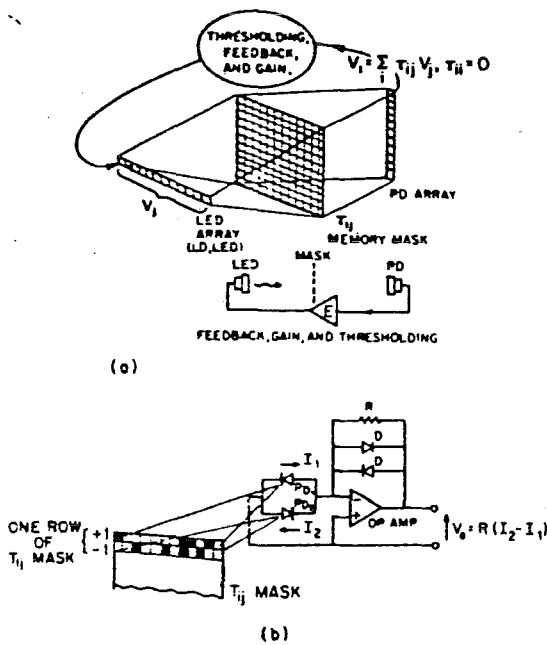


Fig. 3. Concept for optical implementation of a content addressable memory based on the Hopfield model. (a) Matrix-vector multiplier incorporating nonlinear electronic feedback. (b) Scheme for realizing a binary bipolar memory mask transmittance in incoherent light.

monolithic structure that can also be made to contain all ICs for thresholding, amplification, and driving of LEDs. Optical feedback becomes even more attractive when we consider that arrays of nonlinear optical light amplifiers with internal feedback⁹ or optical bistability

devices (OBDs)¹⁰ can be used to replace the PD/LED arrays. This can lead to simple compact CAM structures that may be interconnected to perform higher-order computations than the nearest-neighbor search performed by a single CAM.

We have assembled a simple optical system that is a variation of the scheme presented in Fig. 3(a) to simulate a network of $N = 32$ neurons. The system, details of which are given in Figs. 5-8, was constructed with an array of thirty-two LEDs and two multichannel silicon PD arrays, each consisting of thirty-two elements. Twice as many PD elements as LEDs are needed in order to implement a bipolar memory mask transmittance in incoherent light in accordance with the scheme of Fig. 3(b). A bipolar binary T_{ij} mask was prepared for $M = 3$ binary state vectors. The three vectors or words chosen, their Hamming distances from each other, and the resulting T_{ij} memory matrix are shown in Fig. 4. The mean Hamming distance between the three vectors is 16. A binary photographic transparency of 32×64 square pixels was computer generated from the T_{ij} matrix by assigning the positive values in any given row of T_{ij} to transparent pixels in one subrow of the mask and the negative values to transparent pixels in the adjacent subrow. To insure that the image of the input LED array is uniformly smeared over the memory mask it was found convenient to split the mask in two halves, as shown in Fig. 5, and to use the resulting submasks in two identical optical arms as shown in Fig. 6. The size of the subrows of the memory submasks was made exactly equal to the element size of the PD arrays in the vertical direction which were placed in register

Stored words:

Word 1 : 1 1 0 0 0 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 0 0 0 0 0 1 0
 Word 2 : 0 1 1 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 1 1 1 1 0 1 0 1 1 0 : 0
 Word 3 : 1 0 1 1 0 0 1 1 1 1 1 1 1 1 0 0 0 1 0 1 1 0 0 0 0 1 1 0 0 0 0

Hamming distance from word to word:

WORD	1	2	3
1	0	15	14
2	15	0	19
3	14	19	0

Clipped memory matrix:

0 -1 1 1 -1 -1 1 1 1 1 -1 1 1 1 1 -1 1 -1 1 1 1 1 -1 -1 1 1 1 -1 -1 1 -1 1 -1 -1 -1 -1 -1
 -1 0 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 1 1 1 1 -1 1 1 1 1 1 -1 -1 1 -1 1 -1 1 -1 1 -1
 1 1 0 -1 -1 -1 -1 -1 1 1 1 1 1 -1 1 -1 -1 -1 1 1 -1 1 -1 -1 -1 -1 1 -1 -1 1 -1 1 -1
 1 -1 0 1 1 1 1 1 1 1 1 1 -1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 1 1 -1 1 -1 1
 -1 -1 -1 1 0 1 1 -1 1 -1 -1 -1 -1 -1 1 1 -1 1 -1 -1 1 -1 1 1 1 -1 1 -1 1 1 -1 1
 -1 -1 1 1 0 1 1 1 1 1 1 -1 1 -1 -1 -1 -1 1 -1 -1 -1 -1 1 1 1 -1 1 -1 1 1 -1 1
 1 -1 1 1 1 0 1 1 1 1 1 1 -1 1 -1 -1 -1 -1 1 1 1 -1 -1 1 -1 1 -1 -1 -1 -1 -1
 1 -1 1 1 1 1 0 1 1 1 1 1 -1 1 -1 -1 -1 -1 1 1 1 -1 -1 1 -1 1 -1 -1 -1 -1 -1
 -1 -1 1 1 -1 -1 1 -1 -1 -1 0 -1 -1 1 1 -1 -1 -1 -1 1 1 1 -1 -1 1 1 1 1 -1 -1 -1
 1 -1 1 1 -1 -1 1 1 1 1 -1 0 1 1 1 -1 1 -1 1 1 1 -1 -1 1 -1 1 -1 -1 -1 -1 -1
 1 -1 1 1 -1 -1 1 1 1 1 -1 0 1 1 1 -1 1 -1 1 1 1 -1 -1 1 -1 1 -1 -1 -1 -1 -1
 1 1 1 -1 -1 -1 1 1 1 1 1 1 0 -1 1 -1 -1 1 -1 1 1 -1 -1 -1 1 -1 -1 -1 -1 -1
 1 -1 -1 1 1 1 1 1 1 1 1 -1 0 -1 -1 -1 -1 -1 -1 -1 -1 1 1 -1 1 -1 1 -1 1
 -1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 0 1 1 -1 1 1 1 1 1 1 -1 -1 1 -1 1 -1 1
 1 -1 1 1 -1 -1 1 1 1 1 -1 1 1 1 -1 1 -1 0 1 1 -1 -1 1 -1 1 -1 -1 -1 -1
 1 1 -1 1 1 -1 1 -1 1 -1 1 1 -1 -1 1 1 -1 1 1 -1 -1 -1 1 1 1 1 1
 1 1 1 -1 -1 -1 -1 -1 1 1 1 1 1 -1 -1 -1 -1 0 1 -1 1 -1 -1 -1 1 -1 -1 1
 1 1 1 -1 -1 -1 -1 -1 1 1 1 1 -1 -1 -1 -1 0 -1 1 -1 -1 -1 1 1 1 1 1
 -1 1 -1 -1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 0 1 -1 1 -1 1 1 1 1 1
 -1 1 -1 -1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 0 1 -1 1 -1 1 1 1 1
 -1 -1 1 1 1 1 -1 1 -1 -1 -1 -1 -1 -1 -1 -1 1 1 -1 1 1 1 1 1 1 1
 -1 -1 1 1 1 1 -1 1 -1 -1 -1 -1 -1 -1 -1 -1 1 1 -1 1 1 1 1 1 1 1
 -1 1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 1 1 1 1 1 1 1 1 1 1 1
 -1 -1 -1 1 1 1 -1 1 -1 -1 -1 -1 -1 -1 -1 1 1 1 1 1 1 1 1 1 1 1
 -1 -1 -1 1 1 1 -1 1 -1 -1 -1 -1 -1 -1 -1 1 1 1 1 1 1 1 1 1 1 1

Fig. 4. Stored words, their Hamming distances, and their clipped T_{ij} memory matrix.



Fig. 5. Two halves of T_{ij} memory mask.

against the masks. Light emerging from each subrow of a memory submask was collected (spatially integrated) by one of the vertically oriented elements of the multichannel PD array. In this fashion the anamorphic optics required in the output part of Fig. 3(a) are disposed of, resulting in a more simple and compact system. Pictorial views of the input LED array and the

two submask/PD array assemblies are shown in Figs. 7(a) and (b), respectively. In Fig. 7(b) the left memory submask/PD array assembly is shown with the submask removed to reveal the silicon PD array situated behind it. All electronic circuits (amplifiers, thresholding comparators, LED drivers, etc.) in the thirty-two parallel feedback channels are contained in the electronic amplification and thresholding box shown in Fig. 6(a) and in the boxes on which the LED array and the two submask/PD array assemblies are mounted (see Fig. 7). A pictorial view of a composing and display box is shown in Fig. 8. This contains an arrangement of thirty-two switches and a thirty-two element LED display panel whose elements are connected in parallel to the input LED array. The function of this box is to compose and

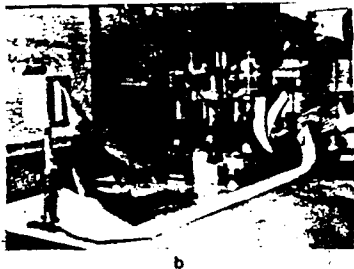
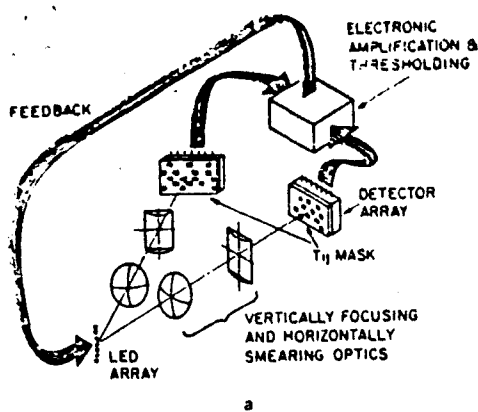


Fig. 6. Arrangement for optical implementation of the Hopfield model: (a) optoelectronic circuit diagram, (b) pictorial view.

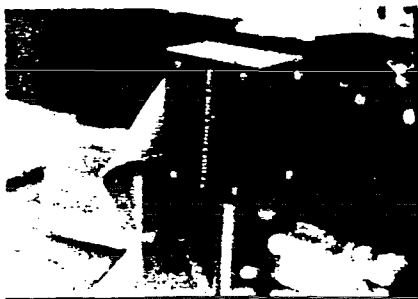


Fig. 7. Views of (a) input LED array and (b) memory submask/PD array assemblies.

display the binary input word or vector that appears on the input LED array of the system shown in Fig. 7(a). Once an input vector is selected it appears displayed on the composing box and on the input LED box simultaneously. A single switch is then thrown to release the system into operation with the composed vector as the

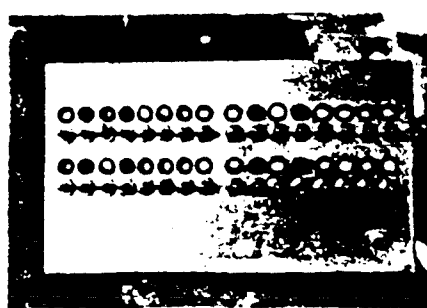


Fig. 8. Word composer and display box.

initializing vector. The final state of the system, the output, appears after a few iterations displayed on the input LED array and the display box simultaneously. The above procedure provides for convenient exercising of the system in order to study its response vs stimulus behavior. An input vector is composed and its Hamming distance from each of the nominal state vectors stored in the memory is noted. The vector is then used to initialize the CAM as described above and the output vector representing the final state of the CAM appearing, almost immediately, on the display box is noted. The response time of the electronic feedback channels as determined by the 3-dB roll-off of the amplifiers was ~ 60 msec. Speed of operation was not an issue in this study, and thus low response time was chosen to facilitate the experiment.

IV. Results

The results of exercising and evaluating the performance of the system we described in the preceding section are tabulated in Table I. The first run of initializing vectors used in exercising the system were error laden versions of the first word $b_i^{(1)}$. These were obtained from $b_i^{(1)}$ by successively altering (switching) the states of 1, 2, 3 . . . up to N of its digits starting from the N th digit. In doing so the Hamming distance between the initializing vector and $b_i^{(1)}$ is increased linearly in unit steps as shown in the first column of Table I whereas, on the average, the Hamming distance between all these initializing vectors and the other two state vectors remained approximately the same, about $N/2 = 16$. The final states of the memory, i.e., the steady-state vectors displayed at the output of the system (the composing and display box) when the memory is prompted by the initializing vectors, are listed in column 2 of Table I. When the Hamming distance of the initializing vector from $b_i^{(1)}$ is < 11 , the input is always recognized correctly as $b_i^{(1)}$. The CAM is able therefore to recognize the input vector as $b_i^{(1)}$ even when up to 11 of its digits (37.5%) are wrong. This performance is identical to the results obtained with a digital simulation shown in parenthesis in column 2 for comparison. When the Hamming distance is increased further to values lying between 12 and 22, the CAM is confused and identifies erroneously other state vectors, mostly $b_i^{(3)}$, as the input. In this range, the Hamming distance of the initializing vectors from any of the stored vectors is approximately equal making it more difficult for the CAM to decide. Note that the performance of

Table I. Optical CAM Performance

Hamming distance of initializing vector from $b_i^{(m)}$	Recognized vector ($m = 1$)	Recognized vector ($m = 2$)	Recognized vector ($m = 3$)
0	1 (1)	2 (2)	3 (3)
1	1 (1)	2 (2)	3 (3)
2	1 (1)	2 (2)	3 (3)
3	1 (1)	2 (2)	3 (3)
4	1 (1)	2 (2)	3 (3)
5	1 (1)	2 (2)	3 (3)
6	1 (1)	2 (2)	3 (3)
7	1 (1)	2 (2)	3 (3)
8	1 (1)	2 (2)	3 (3)
9	1 (1)	2 (2)	3 (3)
10	1 (1)	1 (1)	3 (3)
11	1 (1)	2 (2)	3 (3)
12	3 (3)	3,2 (3)	3 (3)
13	3 (3)	3 (3)	3 (2)
14	3 (3)	1,3 (1)	3 (2)
15	1 (OSC)	1 (1)	2,3 (2)
16	3 (OSC)	1 (1)	2 (2)
17	3 (OSC)	1 (OSC)	2 (2)
18	3 (3)	1 (2)	3 (OSC)
19	3 (2)	2 (2)	2 (2)
20	3 (1)	2 (2)	2 (OSC)
21	1,2 (1)	2 (2)	3 (OSC)
22	3 (1)	2 (2)	3 (OSC)
23	1 (1)	2 (2)	3 (OSC)
24	1 (1)	2 (2)	3 (3)
25	1 (1)	2 (2)	3 (3)
26	1 (1)	2 (2)	3 (3)
27	1 (1)	2 (2)	3 (3)
28	1 (1)	2 (2)	3 (3)
29	1 (1)	2 (2)	3 (3)
30	1 (1)	2 (2)	3 (3)
31	1 (1)	2 (2)	3 (3)
32	1 (1)	2 (2)	3 (3)

the CAM and results of digital simulation in this range of Hamming distance are comparable except for the appearance of oscillations (designated by OSC) in the digital simulation when the outcome oscillated between several vectors that were not the nominal state vectors of the CAM. Beyond a Hamming distance of 22 both the optical system and the digital simulation identified the initializing vectors as the complement $\bar{b}_i^{(1)}$ of $b_i^{(1)}$. This is expected because it can be shown using Eq. (1) that the T_{ij} matrix formed from a set of vectors $b_i^{(m)}$ is identical to that formed by the complementary set $\bar{b}_i^{(m)}$. The complementary vector can be viewed as a contrast reversed version of the original vector in which zeros and ones are interchanged. Recognition of a complementary state vector by the CAM is analogous to our recognizing a photographic image from the negative.

Similar results of initializing the CAM with error laden versions of $b_i^{(2)}$ and $b_i^{(3)}$ were also obtained. These are presented in columns 2 and 3 of Table I. Here again we see when the Hamming distance of the initializing vector from $b_i^{(3)}$, for example, ranged between 1 and 14, the CAM recognized the input correctly as $b_i^{(3)}$ as shown in column 3 of the table and as such it did slightly better than the results of digital simulation. Oscillatory behavior is also observed here in the digital simulation when the range of Hamming distance between the ini-

tializing vector from all stored vectors approached the mean Hamming distance between the stored vectors. Beyond this range the memory recognizes the input as the complementary of $b_i^{(3)}$.

In studying the results presented in Table I several observations can be made: The optically implemented CAM is working as accurately as the digital simulations and perhaps better if we consider the absence of oscillations. These are believed to be suppressed in the system because of the nonsharp thresholding performed by the smoothly varying nonlinear transfer function of electronic circuits compared with the sharp thresholding in digital computations. The smooth nonlinear transfer function and the finite time constant of the optical system provide a relaxation mechanism that substitutes for the role of asynchronous switching required by the Hopfield model. Generally the system was able to conduct successful nearest-neighbor search when the inputs to the system are versions of the nominal state vectors containing up to ~30% error in their digits. It is worth noting that this performance is achieved in a system built from off-the-shelf electronic and optical components and with relatively little effort in optimizing and fine tuning the system for improved accuracy, thereby confirming the fact that accurate global computation can be performed with relatively inaccurate individual components.

V. Discussion

The number M of state vectors of length N that can be stored at any time in the interconnection matrix T_{ij} is limited to a fraction of N . An estimate of $M \approx 0.1N$ is indicated in simulations involving a hundred neurons or less¹ and a theoretical estimate of $M \approx N/4 \ln N$ has recently been obtained.² It is worthwhile to consider the number of bits that can be stored per interconnection or per neuron. The number of pixels required to form the interconnection matrix is N^2 . Since such a T_{ij} memory matrix can store up to $M \approx N/4 \ln N$ (N -tuples), the number of bits stored is $MN = N^2/4 \ln N$. The number of bits stored per memory matrix element or interconnection is $MN/N^2 = (4 \ln N)^{-1}$, while the number of bits stored per neuron is $MN/N = M$.

The number of stored memories that can be searched for a given initializing input can be increased by using a dynamic memory mask that is rapidly addressed with different T_{ij} matrices each corresponding to different sets of M vectors. The advantage of programmable SLMs for realizing this goal are evident. For example, the Litton Lightmod (magneto-optic light modulator), which has nonvolatile storage capability and can provide high frame rates, could be used. A frame rate of 60 Hz is presently specified for commercially available units of 128×128 pixels which are serially addressed.⁷ Units with 256×256 pixels are also likely to be available in the near future with the same frame rate capability. Assuming a memory mask is realized with a Litton Lightmod of 256×256 pixels we have $N = 256$, $M \approx 0.1N \approx 26$ and a total of $26 \times 60 = 1560$ vectors can be searched or compared per second against an initializing input vector. Speeding up the frame rate of the Litton

Lightmod to increase memory throughput beyond the above value by implementing parallel addressing schemes is also possible. Calculations show that the maximum frame rate possible for the device operating in reflection mode with its drive lines heat sunk is 10 kHz.⁷ This means the memory throughput estimated above can be increased to search 2.6×10^5 vectors/sec, each being 256 bits long, or a total of 6.7×10^8 bits/sec. This is certainly a respectable figure, specially when we consider the error correcting capability and the associative addressing mode of the Hopfield model; i.e., useful computation is performed in addition to memory addressing.

The findings presented here show that the Hopfield model for neural networks and other similar models for content addressable and associative memory fit well the attributes of optics, namely, parallel processing and massive interconnection capabilities. These capabilities allow optical implementation of large neural networks based on the model. The availability of nonlinear or bistable optical light amplifiers with internal feedback, optical bistability devices, and nonvolatile high speed spatial light modulators could greatly simplify the construction of optical CAMs and result in compact modules that can be readily interconnected to perform more general computation than nearest-neighbor search. Such systems can find use in future generation computers, artificial intelligence, and machine vision.

The work described in this paper was performed while one of the authors, N.F., was on scholarly leave at the California Institute of Technology. This author wishes to express his appreciation to CIT and the University of Pennsylvania for facilitating his sabbatical leave. The work was supported in part by the Army Research Office and in part by the Air Force Office of Scientific Research.

The subject matter of this paper is based on a paper presented at the OSA Annual Meeting, San Diego, Oct. 1984.

References

1. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proc. Natl. Acad. Sci. USA* 79, 2554 (1982).
2. R. J. McEliece, E. C. Posner, and S. Venkatesh, California Institute of Technology, Electrical Engineering Department; private communication.
3. G. E. Hinton and J. A. Anderson, *Parallel Models of Associative Memory* (LEA Publishers, Hillsdale, N.J., 1981).
4. T. Kohonen, *Content Addressable Memories* (Springer, New York, 1980).
5. D. Psaltis and N. Farhat, "A New Approach to Optical Information Processing Based On the Hopfield Model," in *Technical Digest, ICO-13 Conference, Sapporo* (1984), p. 24.
6. D. Psaltis and N. Farhat, "Optical Information Processing Based on an Associative-Memory Model of Neural Nets with Thresholding and Feedback," *Opt. Lett.* 10, 98 (1985).
7. W. Ross, D. Psaltis, and R. Anderson, "Two-Dimensional Magneto-Optic Spatial Light Modulator For Signal Processing," *Opt. Eng.* 22, 485 (1983).
8. J. W. Goodman, A. R. Dias, and L. M. Woody, "Fully Parallel, High-Speed Incoherent Optical Method for Performing Discrete Fourier Transforms," *Opt. Lett.* 2, 1 (1978).
9. Z. Porada, "Thin Film Light Amplifier with Optical Feedback," *Thin Solid Films* 109, 213 (1983).
10. H. M. Gibbs *et al.*, "Optical Bistable Devices: The Basic Components of All-Optical Circuits," *Proc. Soc. Photo-Opt. Instrum. Eng.* 269, 75 (1981).

Engineering Summer Conferences

300

Chrysler Center/North Campus
Ann Arbor, Michigan 48109-2092
313/764-8490

College of Engineering
The University of Michigan

8521—COMPUTER VISION AND IMAGE PROCESSING

July 29-August 2, 1985

Fee: \$750

Chairman: R.C. Jain

With the advent of high speed computers, processing and extracting information from images has become an important technology. This course presents techniques for processing images and recovering useful information with emphasis on solving problems having a variety of applications.

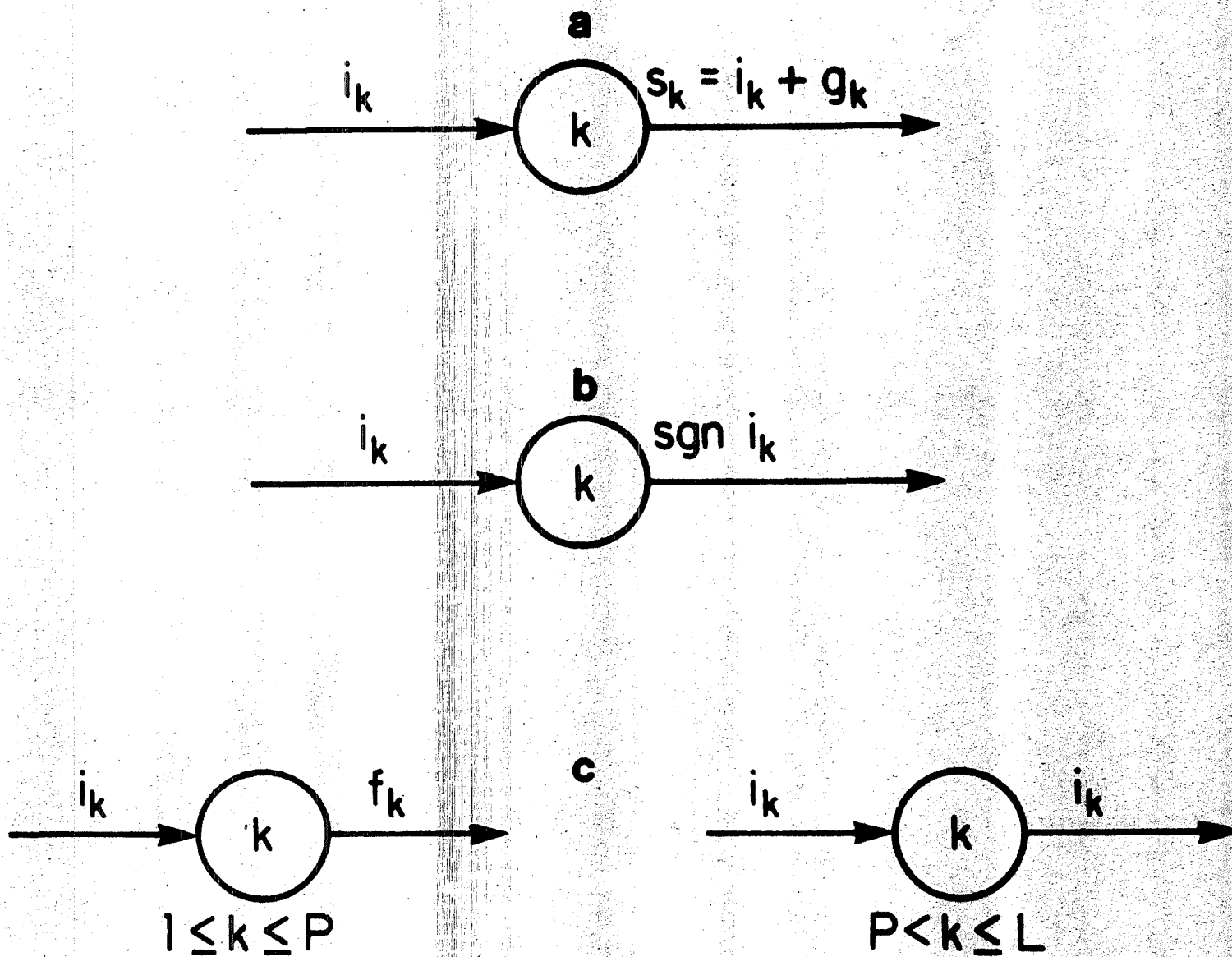


FIG 1

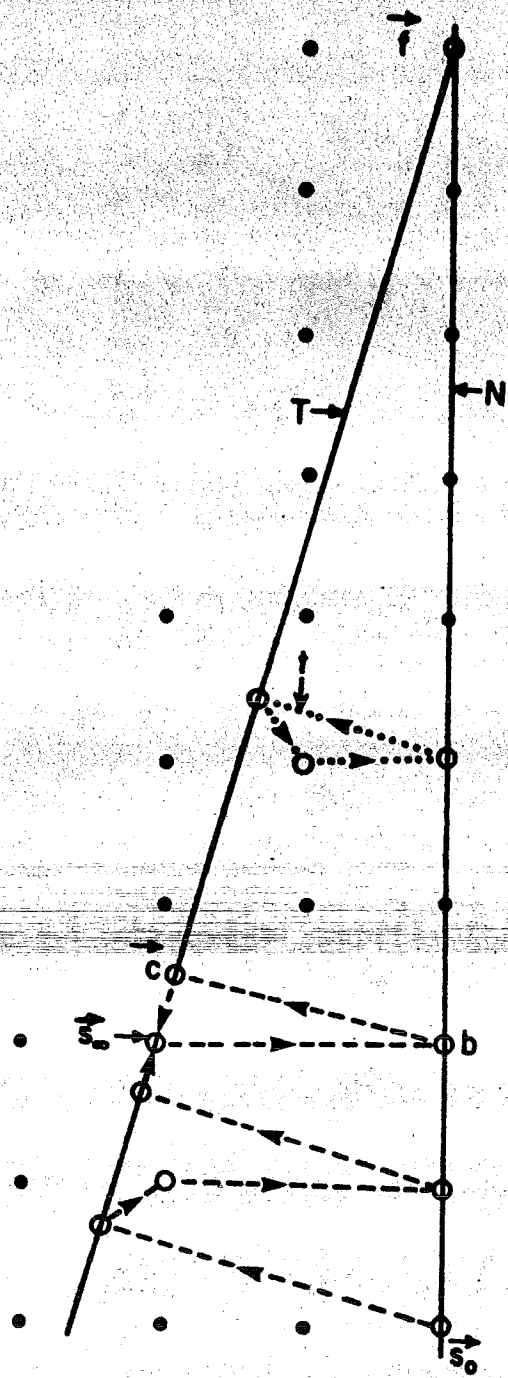


FIG 3

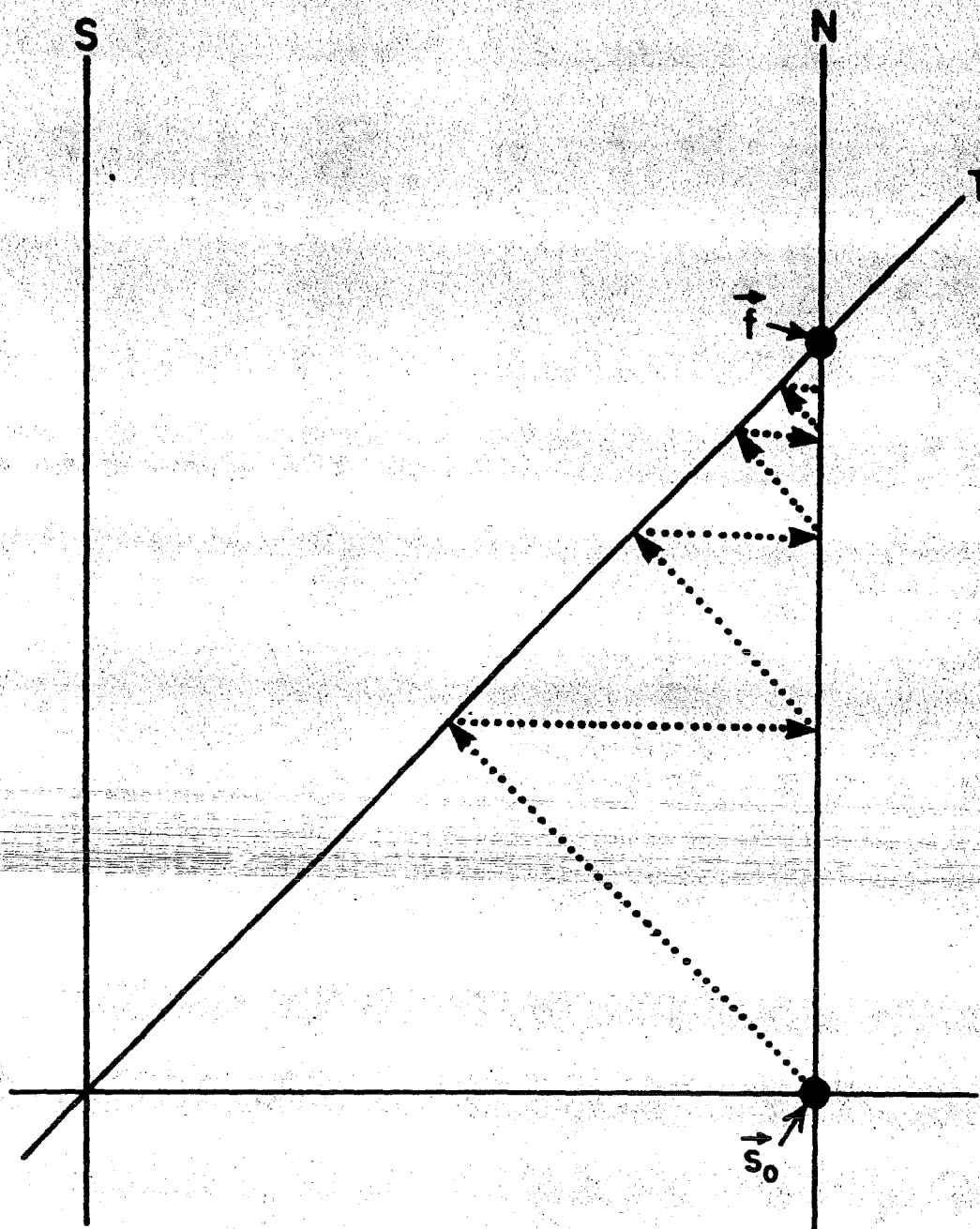


FIG 2

RECALL TECHNIQUE	MULTIPLIES/ITERATION
Extrapolation Net	L^2
...Outer Product Technique	$2NL$
Table Look-Up Net	Q^2
...Outer Product Technique	$2NQ$

Table 1.

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

Department of Electrical Engineering, FT-10
Telephone: (206) 543-2150

November 5, 1986

Dr. Robert Graham
Boeing High Technology Center
Boeing Electronics Co.
P.O. Box 3707, MS 7J-05
Seattle, WA 98124-2207

Dear Rob,

Attached are two papers. The first is "A Continuous Level Associative Memory Neural Net" submitted to the Second Topical Meeting on Optical Computers at Lake Tahoe in March 1987. Although your support is gratefully acknowledged, the paper's contents are subsumed in an archival journal paper "A Continuous level Memory Extrapolation Neural O Net" submitted to Applied Optics in August of this year. This was prior to Boeing's support of our work.

The second paper, entitled "An All Optical Iterative Neural Net Recall Memory" outlines an architecture for a neural net based processor that operates at light speed. I wish to present this result at the Tahoe conference and submit the paper to an archival journal. This, of course, will be done in accordance to the "Analysis and Application of Neural Net" contract. I will assume, unless informed otherwise, that the time period for your review begins today.

Best regards,

Robert J. Marks II
Associate Professor

cc: J.A. Ritcey
L.E. Atlas
A. Somani
E. Stear, Washington High Tech Center

AN ALL OPTICAL ITERATIVE

NEURAL NET RECALL MEMORY

Robert J. Marks II

Interactive Systems Design Lab

Department of Electrical Engineering

University of Washington

Seattle, WA 98195

11-3-86

ABSTRACT

We propose an architecture for a continuous level discrete valued table look-up memory. Unlike other iterative memory recall optical processors, the processor performs at optical speeds, e.g. there are no electronics or slow optics (such as phase conjugators) in either the forward or feedback paths. Techniques to compensate for processor losses are presented.

INTRODUCTION

Hopfield's neural net content addressable memory¹ has stirred a flurry of interest in the signal processing community. Optical implementations of such nets have been proposed and implemented²⁻³. Unlike planar VLSI, optical implementations are not restricted to nearest neighbor interconnects.

In a previous paper⁴ the author has described a class of neural net associative and table look-up memory algorithms based on convex set projection theory. This paper describes a processor for performing one of these algorithms. Unlike other iterative recall memories, the proposed processor operates at light speed.

PRELIMINARIES

In a previous paper, the author described a class of table look-up artificial neural networks for generating a vector in a specified library when given only a portion of that vector¹. We outline here one of these nets.

Let $\mathcal{G} = \{\vec{f}_n \mid 1 \leq n \leq N\}$ denote a set of N real continuous level element library vectors of length L . We form the library matrix

$$E = [\vec{f}_1 \mid \vec{f}_2 \mid \dots \mid \vec{f}_N]$$

and the neural net interconnect matrix

$$I = E (E^T E)^{-1} E^T$$

Let $\vec{f} \in \mathcal{G}$. With knowledge of the first P elements of \vec{f} , we wish to extrapolate the remainder. For a given \vec{f} , define the vector operator

$$\underline{n} \vec{a} = [\vec{f}_P \mid \vec{a}_0]^T$$

where \vec{f}_P denotes a vector containing the the first $P < L$ elements of \vec{f} and \vec{a}_0 contains the last $Q=L-P$ elements of \vec{a} . Then the iteration

$$\vec{s}_{M+1} = \underline{n} \underline{I} \vec{s}_M \quad (1)$$

will, for any initialization, converge to \vec{f} if $P \geq N$ and the first P rows of E form a matrix of full rank.

Some of the operations performed in () are not used since the $\underline{\eta}$ operator replaces the first P vector elements with \vec{f}_P . Thus, (1) can equivalently be written as

$$\vec{s}_{M+1,0} = I_0 [\vec{f}_P | \vec{s}_{M,0}]^T \quad (2)$$

where $\vec{s}_{M,0}$ is the vector of the last Q elements of \vec{s}_M and the matrix I_0 consists of the last Q rows of I . The neuron operator, $\underline{\eta}$, is not required in this equation since, for the last Q nodes, it is an identity operator.

OPTICAL IMPLEMENTATION

The iterative neural net memory described by (2) can be straightforwardly implemented on an optical processor. While similar iterative memories have been implemented optically²⁻⁴, each requires either electronics or slow optics (e.g. phase conjugation mirrors) in the processor. As we will show, there are architectures for implementing (2) that are completely optical. This is because no nonlinear operations need to be performed in the processor's feedback path.

The basic processor architecture is shown in Fig.1. The processor input corresponding to \vec{f}_p is supplied by a linear array of P point source LED's. The feed-forward path consists of a standard optical matrix-vector multiplier⁵⁻⁹. (The astigmatic spreading and focusing optics have been deleted from the figure for presentation clarity.) The processor output, $\vec{s}_{m,o}$, is then fed back to the input through fibers. Once the input array is on, the iteration in (2) is thus performed at an optical speed.

The astute reader will immediately notice three fundamental problems with this processor: (1) there is no provision made for detecting the processor output (2) absorbtive and other losses can significantly inhibit performance and (3) we require bipolar multiplication and addition operations rather than just the non-negative operations directly available in processors such as ours. We now address and offer solutions for each of these problems.

Although the feedback is linear, the processor is not. For any constant c , for example, $c\underline{n} \neq \underline{n}c$. The homogeneity property of linearity is thus violated. For linear processors, the mask and the input can be scaled independently. The corresponding multiplicative proportionality constant at the output is equivalently altered by either. For the processor in Fig.1, on the other hand, we are allowed only one scaling parameter. If, for example, the mask transmittance is scaled so that no gain is required, the input LED irradiance must be similarly scaled.

We now address the problem of negative number operations in the processor. Methods of encoding both positive and negative (and even complex) number operations on incoherent algebraic processors have been proposed⁶⁻⁹. Such an extension of our processor is shown in Fig2. We decompose the \underline{I}_0 matrix as

$$\underline{I}_0 = \underline{I}_0^+ + \underline{I}_0^-$$

where all of the elements of \underline{I}_0^+ are nonnegative and those in \underline{I}_0^- are nonpositive. All negative elements in \underline{I}_0 , for example, are set to zero to form \underline{I}_0^+ . Similarly, we can write

$$\vec{f}_0 = \vec{f}_0^+ + \vec{f}_0^-$$

and

$$\vec{s}_{0,M} = \vec{s}_{0,M}^+ + \vec{s}_{0,M}^-$$

FINAL REMARKS

We have proposed an architecture for an all optical table look-up processor based on a neural net model presented previously by the author¹. Once the optical input is made available, each iteration is performed at the time it takes light to circle the processor once.

Numerous variations on the processor are possible and are in need of further study. The optical couplers in Fig.2, for example, can be avoided by placing fiber pairs together at the input to simulate a single point source. Also, the feedback can be performed with planar¹ or concave^{1*} mirrors rather than fibers.

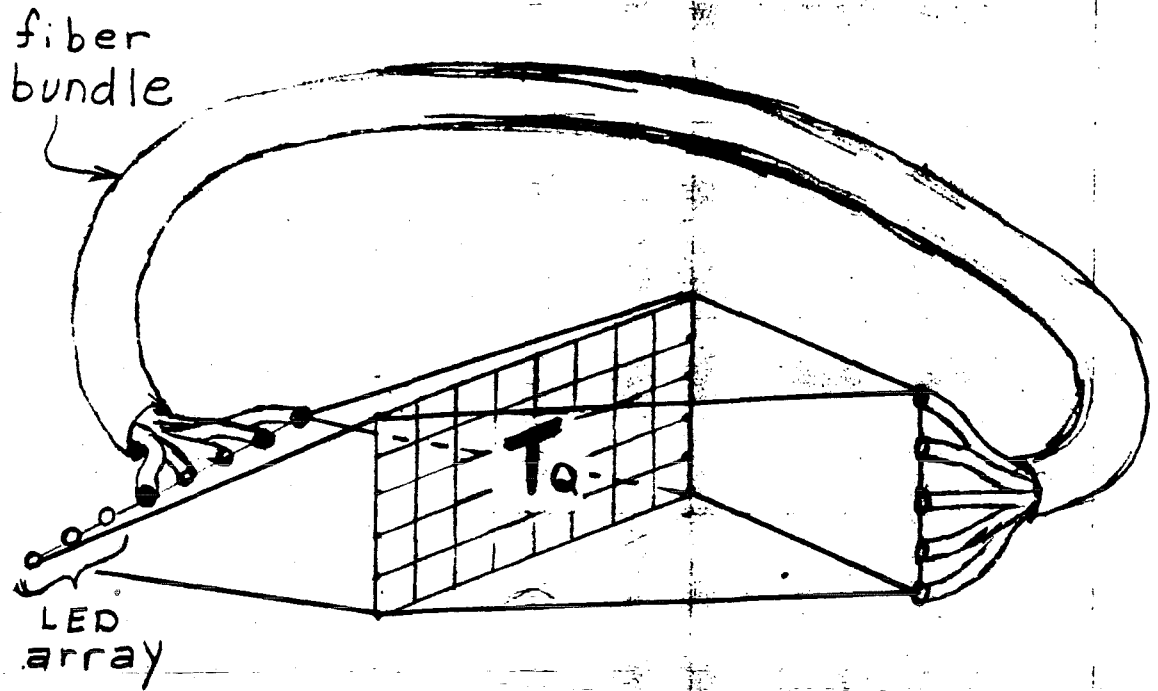


FIG 1

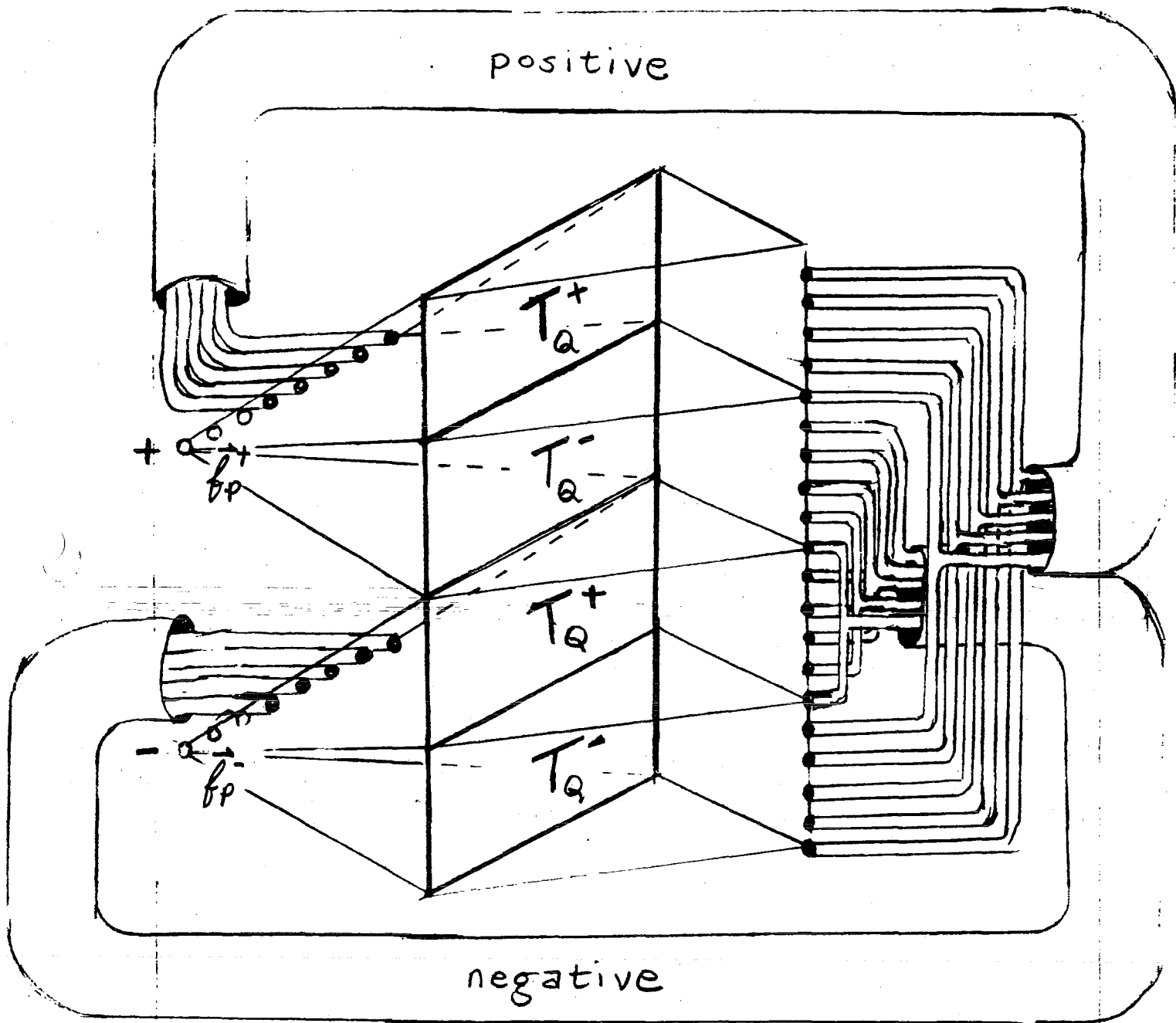


FIG. 2

A Class of Continuous Level Neural Nets and Their Optical Implementation

Robert J. Marks II
ISDL
University of Washington

Contents

1. POCS - What is it ?
 - example CS's
 - projections
 - interative POCS

2. Application to Neural Nets
 - a continuous level neural net based on POCS
 - relaxation for accelerated convergence
 - other convex constraints

3. Optical Implementation
 - fundamental architecture
 - alteration for absorbtive losses
 - alteration for bipolar operations

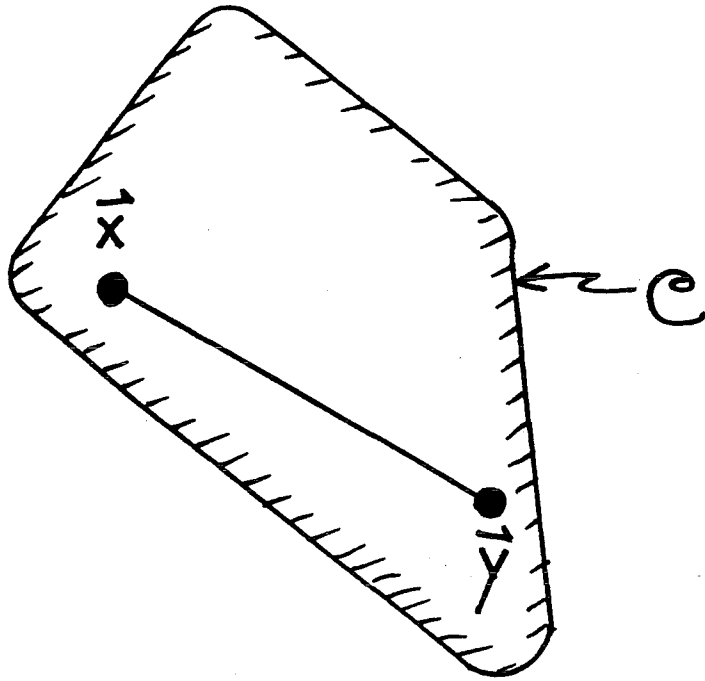
Q: What is POCS?

A: CS

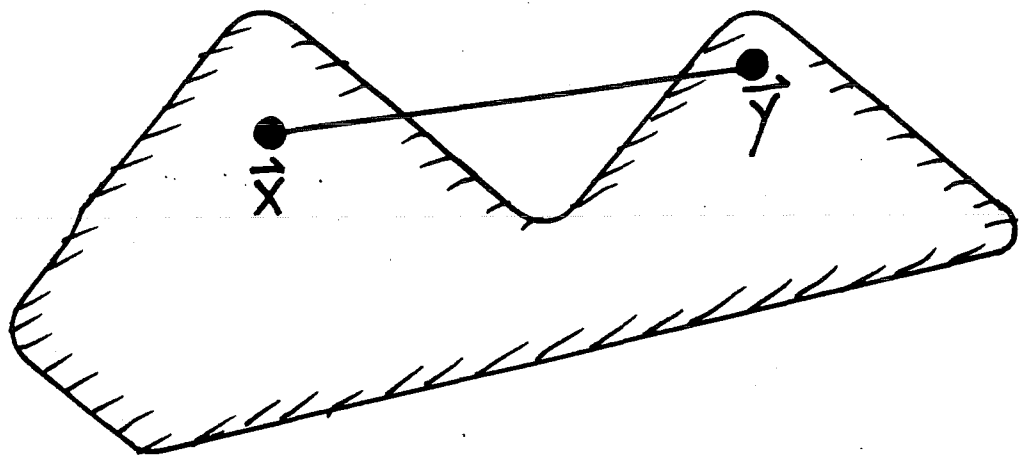
In Hilbert space, \mathcal{H} , the set \mathcal{C} is convex if, for $0 \leq \alpha \leq 1$,

$$\alpha \vec{x} + (1-\alpha) \vec{y} \in \mathcal{C} \quad \forall \vec{x}, \vec{y} \in \mathcal{C}$$

i.e.:



not convex:



Example Convex Sets in \mathbb{R}^L :

★ Subspace : Given N vectors,
 $\{\vec{f}_n \mid 1 \leq n \leq N < L\}$

Then

$$C = \{ \vec{x} \mid \vec{x} = \underline{F}^T \vec{\alpha} \}$$

where

$$\underline{F} = [\vec{f}_1 \mid \vec{f}_2 \mid \dots \mid \vec{f}_N]$$

★ Ball: $C = \{ \vec{x} \mid \|\vec{x}\| \leq R \}$

★ Box: $C = \{ \vec{x} \mid |x_\ell| \leq 1, 1 \leq \ell \leq L \}$

★ Linear Variety:

Given $\{f_1, f_2, \dots, f_P\}, P < L$

$$C = \{ \vec{x} \mid x_\ell = f_\ell, 1 \leq \ell \leq P \}$$

★ First Orthant:

$$C = \{ \vec{x} \mid x_\ell > 0, 1 \leq \ell \leq L \}$$

★ Bandlimited vectors: Given

$P < L$ integers between $1 \leq l \leq L$:

$$C = \{ \vec{x} \mid (\underline{D}\vec{x})_\ell = 0; \ell \in \text{Integers} \}$$

where

$$\underline{D} = \text{DFT matrix}$$

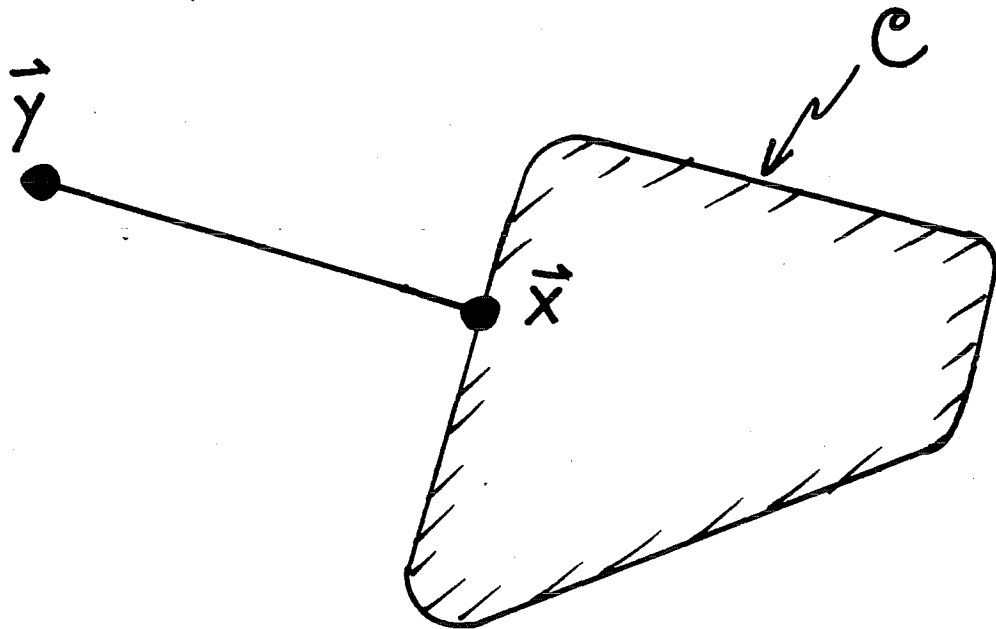
etc.

POCS

$\vec{x} \in \mathcal{C}$ is the projection of $\vec{y} \in \mathbb{R}^L$ onto \mathcal{C} if

$$\inf_{\vec{x} \in \mathcal{C}} \|\vec{x} - \vec{y}\| = \|\vec{x} - \vec{y}\|$$

i.e., \vec{x} is the closest element in \mathcal{C} to \vec{y} :



Notation:

$$\vec{x} = \mathcal{P}_{\mathcal{C}} \vec{y}$$

If $\vec{y} \in \mathcal{C}$, then $\mathcal{P}_{\mathcal{C}} \vec{y} = \vec{y}$

Iterative POCS

★ Case 1: Intersecting CS's*
M convex sets, $\{C_m \mid 1 \leq m \leq M\}$

Define

$$C_1 \cap C_2 \cap \dots \cap C_M = C \neq \emptyset$$

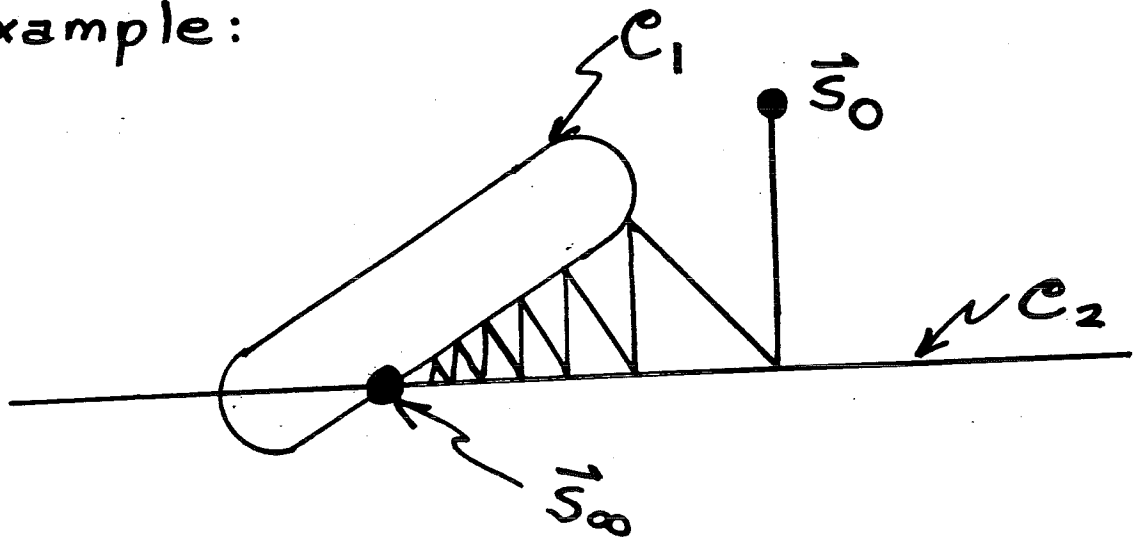
Then, if

$$\vec{s}_{N+1} = P_1 P_2 \dots P_M \vec{s}_N$$

we are assured that

$$\lim_{N \rightarrow \infty} \vec{s}_N \in C$$

Example:



Result may or may not be unique.

* Youla & Webb

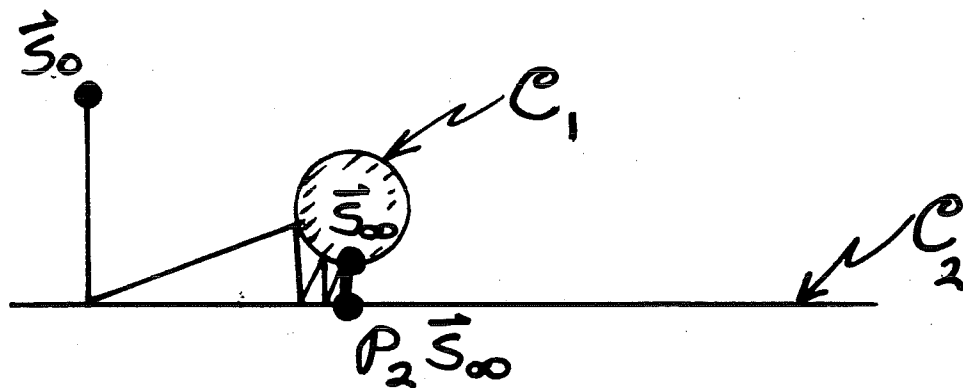
★ Case 2: Two Nonintersecting CS's*:

$$\vec{s}_{M+1} = P_1 P_2 \vec{s}_M$$

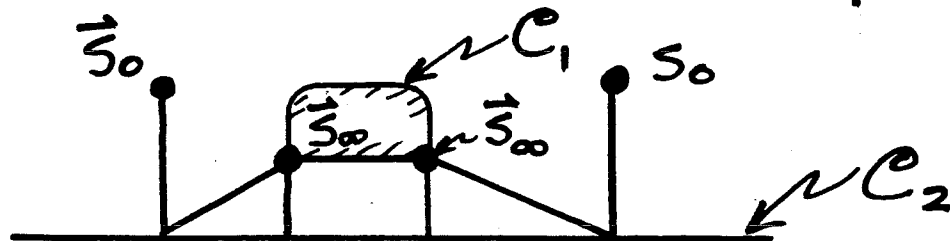
Convergence is to:

$$\inf_{\substack{\vec{x}_1 \in C_1 \\ \vec{x}_2 \in C_2}} \|\vec{x}_1 - \vec{x}_2\| = \|\vec{s}_\infty - P_2 \vec{s}_\infty\|$$

That is, convergence is to the closest distance between C_1 and C_2 :

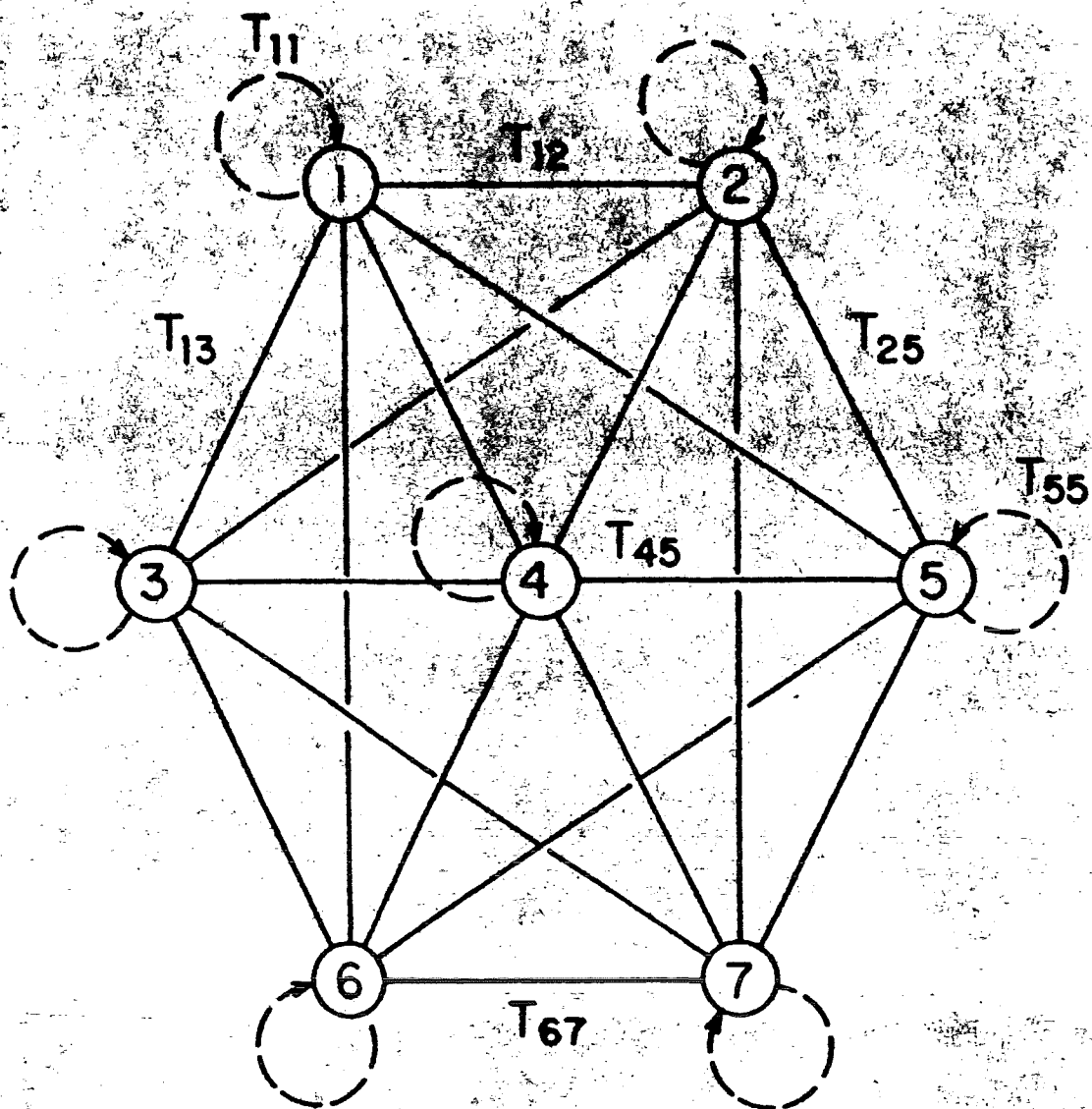


Convergence may not be unique:



* Goldberg and Marks

Neural Nets: A neural net model



L neurons, $t_{ij} = t_{ji}$

\vec{S}_M = neural state at time M

Synchronous Operation: $\vec{S}_{M+1} = \underline{\mathcal{N}} \underline{T} \vec{S}_M$

\underline{T} = matrix of interconnects

$\underline{\mathcal{N}}$ = pointwise vector operator
(performed at nodes)

A Continuous Level Neural Net

N continuous level library vectors:

$$\mathcal{F} = \{ \vec{f}_n \mid 1 \leq n \leq N \}$$

Library matrix:

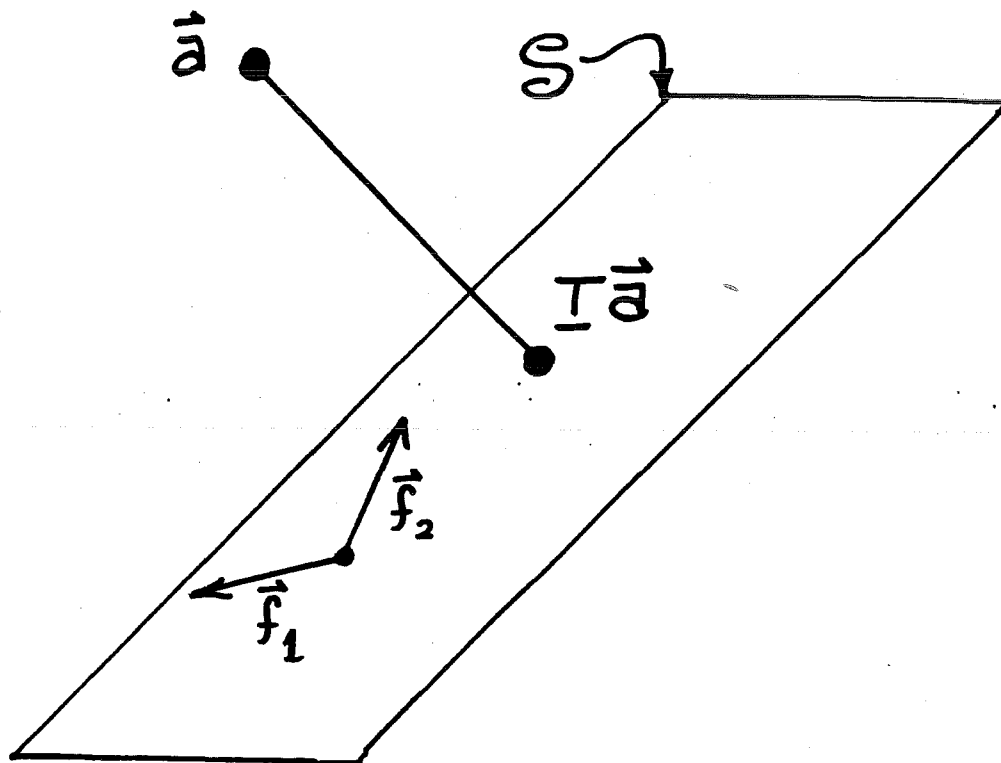
$$\underline{E} = [\vec{f}_1 \mid \vec{f}_2 \mid \dots \mid \vec{f}_N]$$

Projection interconnect matrix:

$$\underline{I} = \underline{E} (\underline{E}^T \underline{E})^{-1} \underline{E}^T$$

\underline{I} projects any $\vec{a} \in \mathbb{R}^L$ onto

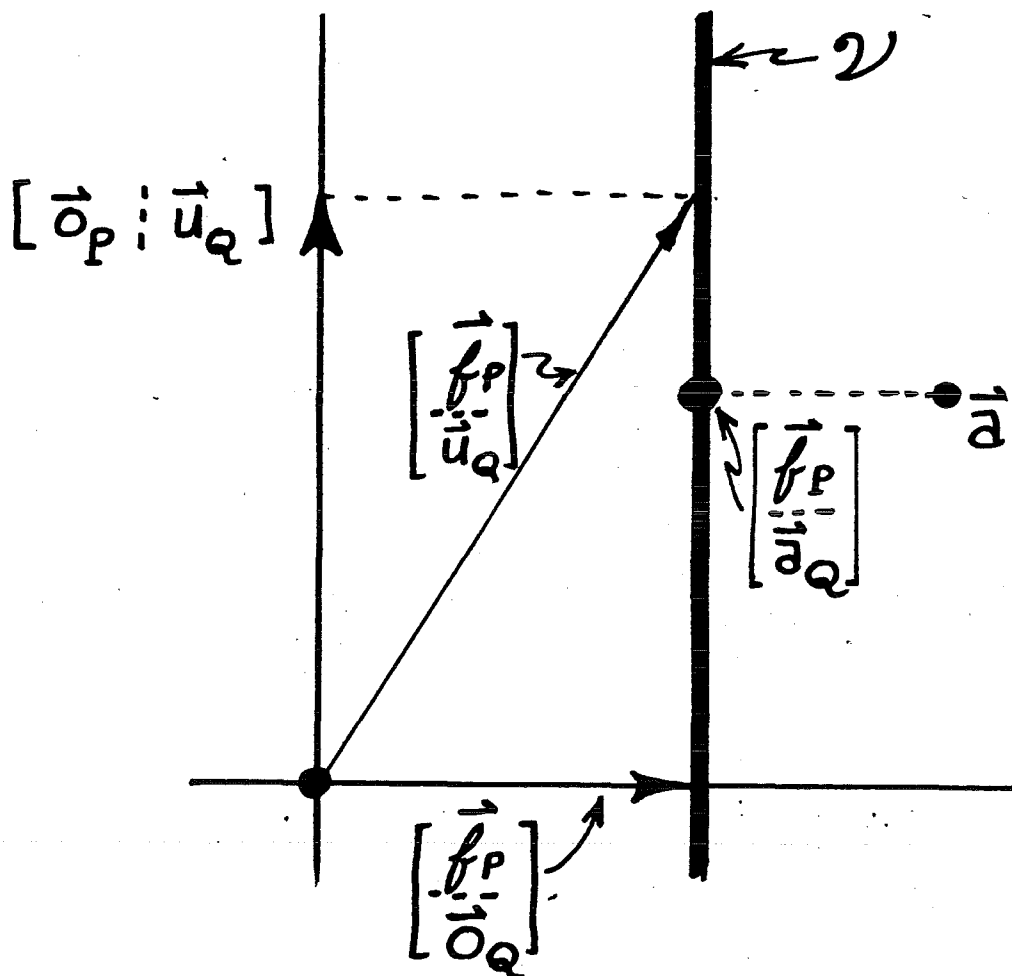
$\mathcal{S} = [\mathcal{F}]$ = subspace generated by \mathcal{F} :



The set

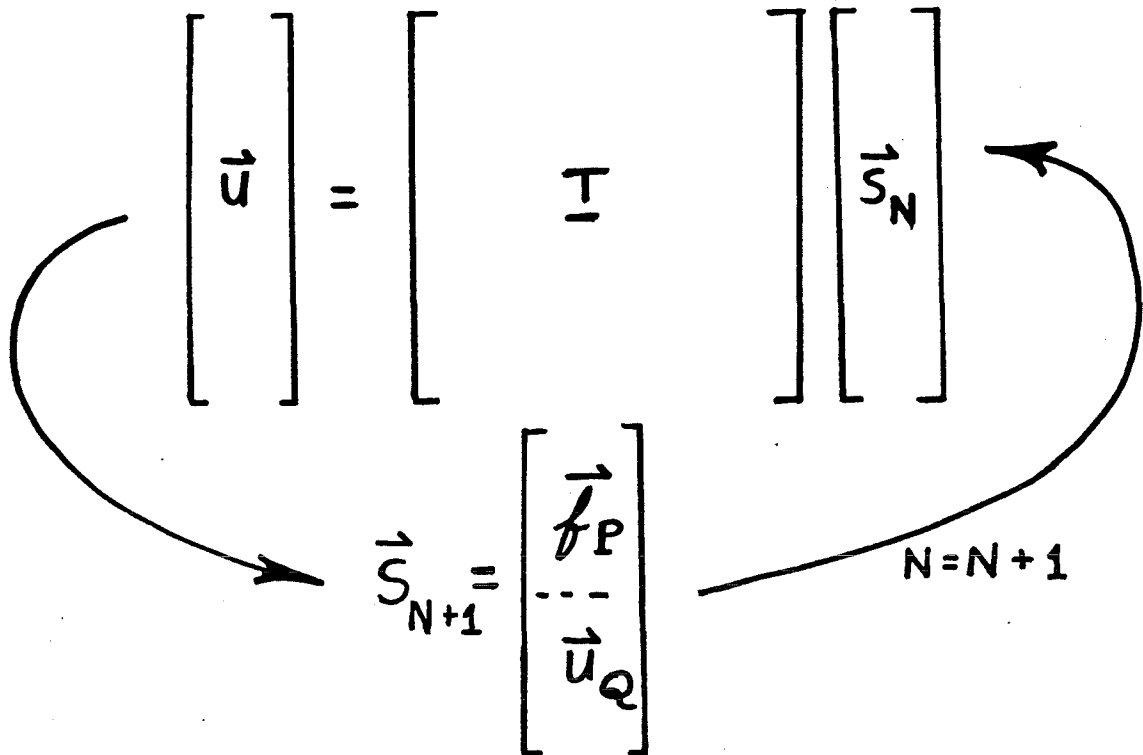
$\mathcal{V} = \{ \vec{x} \mid \vec{x} = [\vec{f}_P; \vec{u}_Q]^T; \vec{u}_Q \in \mathbb{R}^{L-P} \}$
is a linear variety.

$\mathcal{N} \vec{a}$ projects $\vec{a} \in \mathbb{R}^L$ onto \mathcal{V}

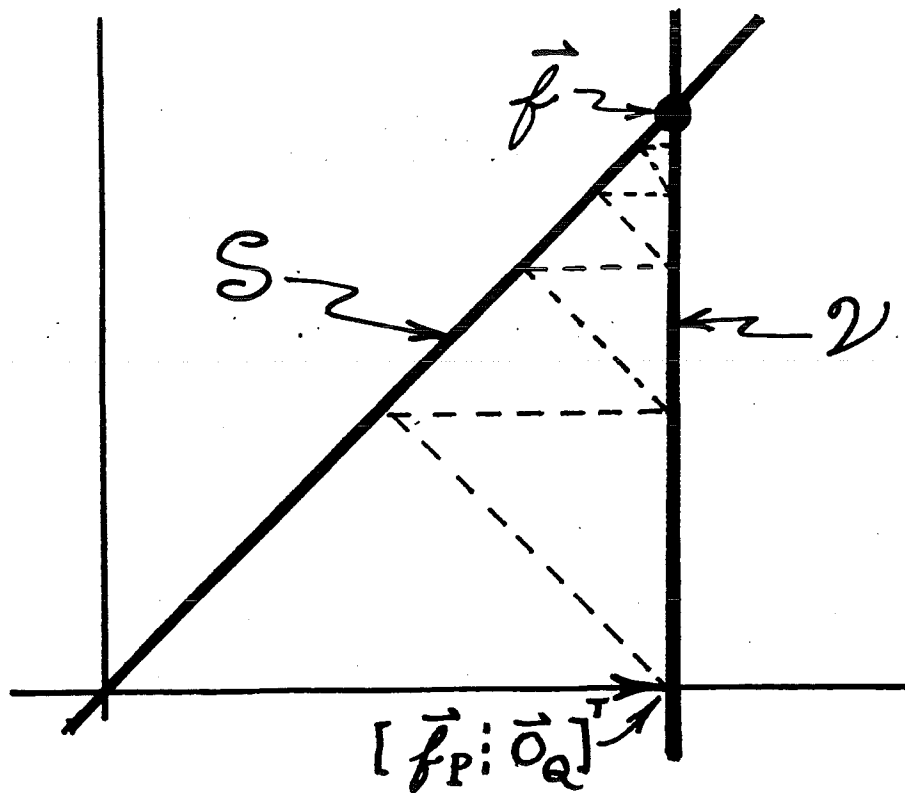


Synchronous Net Operation

Let $\vec{f} \in \mathcal{F}$, $\vec{s}_0 = [\vec{f}_P : \vec{0}_Q]^T$



View in \mathbb{R}^L :



Clearly:

$$\vec{f} \in \mathcal{V} \cap \mathcal{S}$$

Q: When is $\vec{f} = \mathcal{V} \cap \mathcal{S}$?

(Then convergence is unique.)

A: Sufficient conditions:

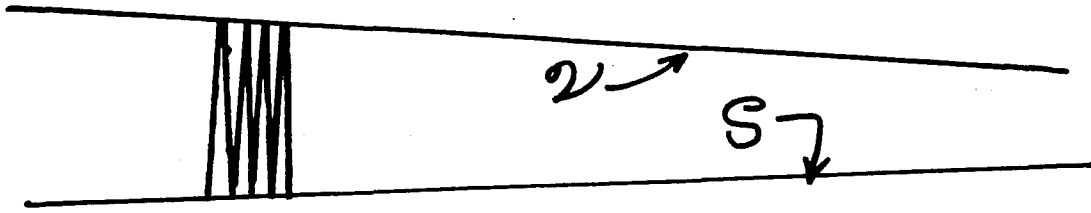
(1) $P = \text{number of known elements}$
 $\geq N = \text{number of library vectors.}$

(2) $E_p = [\vec{f}_{1p} \mid \vec{f}_{2p} \mid \dots \mid \vec{f}_{Np}]$

is full rank.

Relaxation Parameters

Problem: Slow convergence:



A Solution: Relaxation Parameters

$$\underline{T}_r = (1 - \lambda_T) \underline{I} + \lambda_T \underline{T}$$

$$\underline{n}_r = (1 - \lambda_n) \underline{I} - \lambda_n \underline{n}$$

(neurons now have memory)

$$0 < \lambda_T, \lambda_n < 2$$

Other Convex Constraints

e.g. Box $B = \{ \vec{x} \mid \max |x_n| \leq 1 \}$
 $\forall \vec{f} \in \mathcal{F}, |f_n| \leq 1$

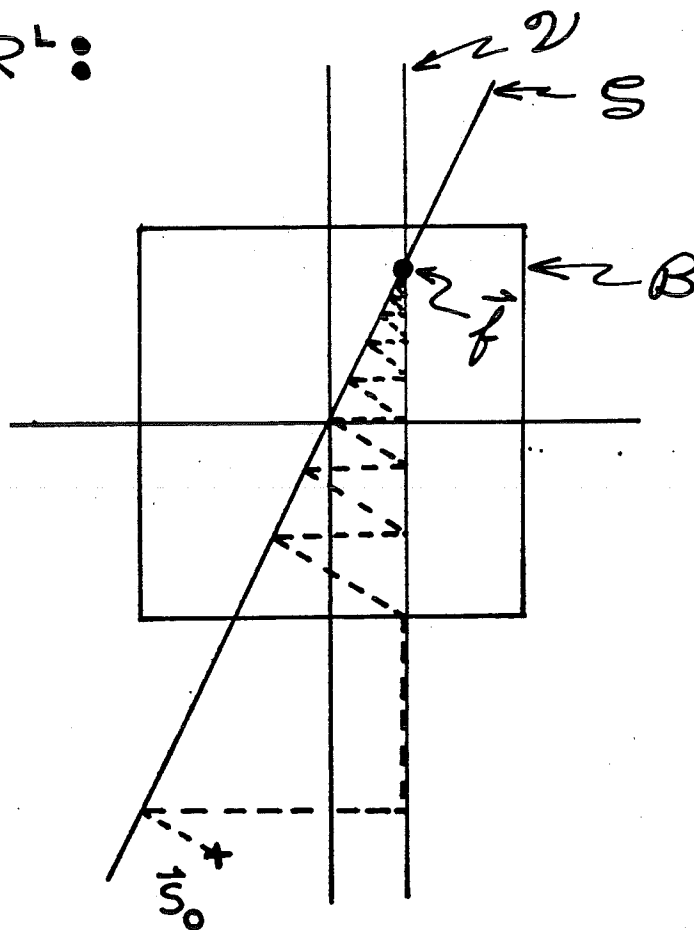
Revised operator

$$\tilde{\mathcal{N}} \vec{a} = [\vec{f}_P ; \vec{b}_Q]^T$$

where:

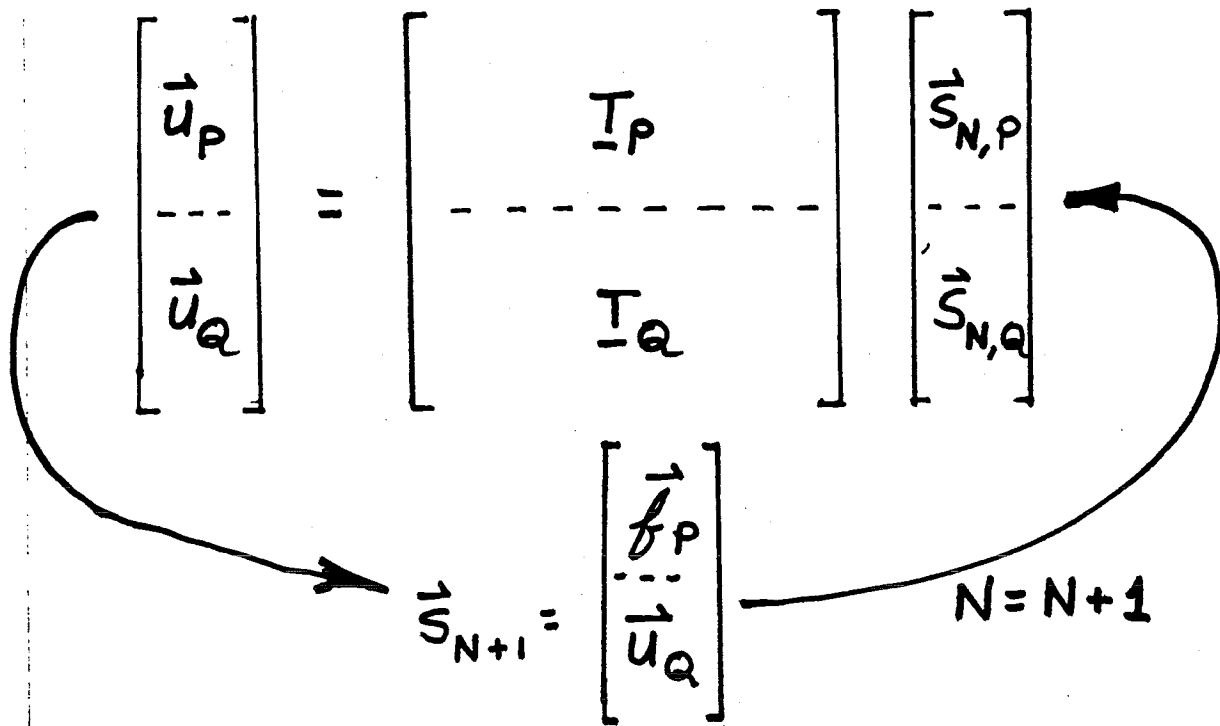
$$(b_Q)_n = \begin{cases} a_n & ; |a_n| \leq 1 \\ 1 & ; a_n \geq 1 \\ -1 & ; a_n \leq -1 \end{cases}$$

View in \mathbb{R}^L :

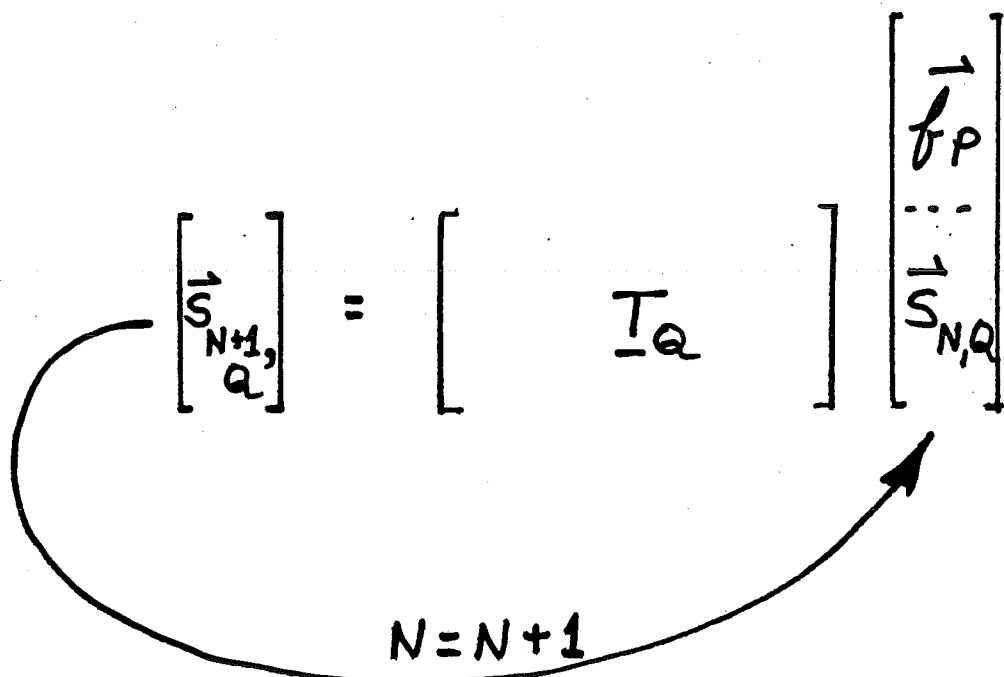


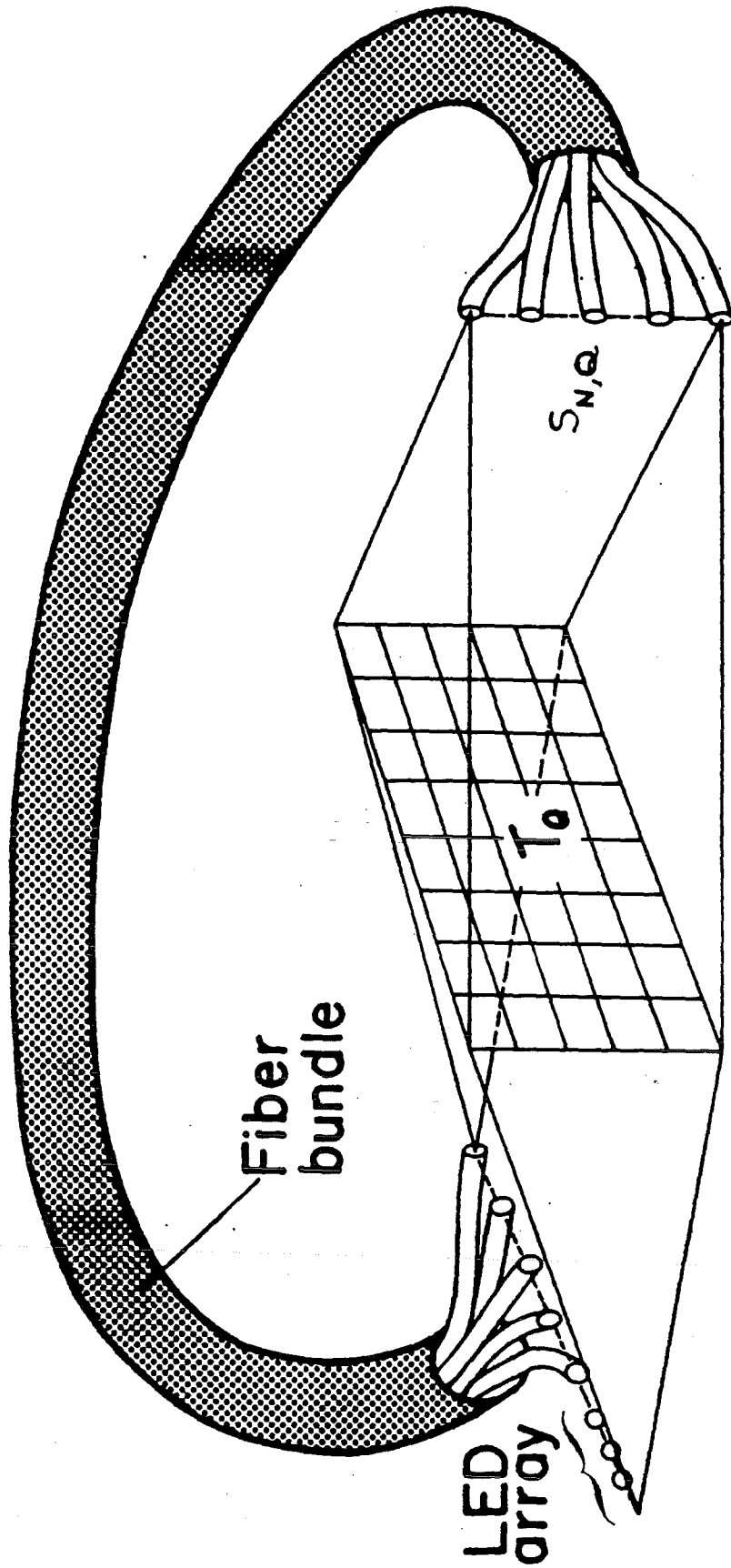
Optical Implementation

- A table look up net:



If the same nodes are always used for the input (table look-up), an equivalent iteration is:





● Question # 1:

(a) How do we detect the result?

(b) What about absorbtive losses?

Answer # 1:

Scale the mask

● Question # 2:

How do we handle bipolar operations?

Answer:

Seperate + & - operations:

$$T_Q = T_Q^+ + T_Q^-$$

$$\vec{f}_P = \vec{f}_P^+ + \vec{f}_P^-$$

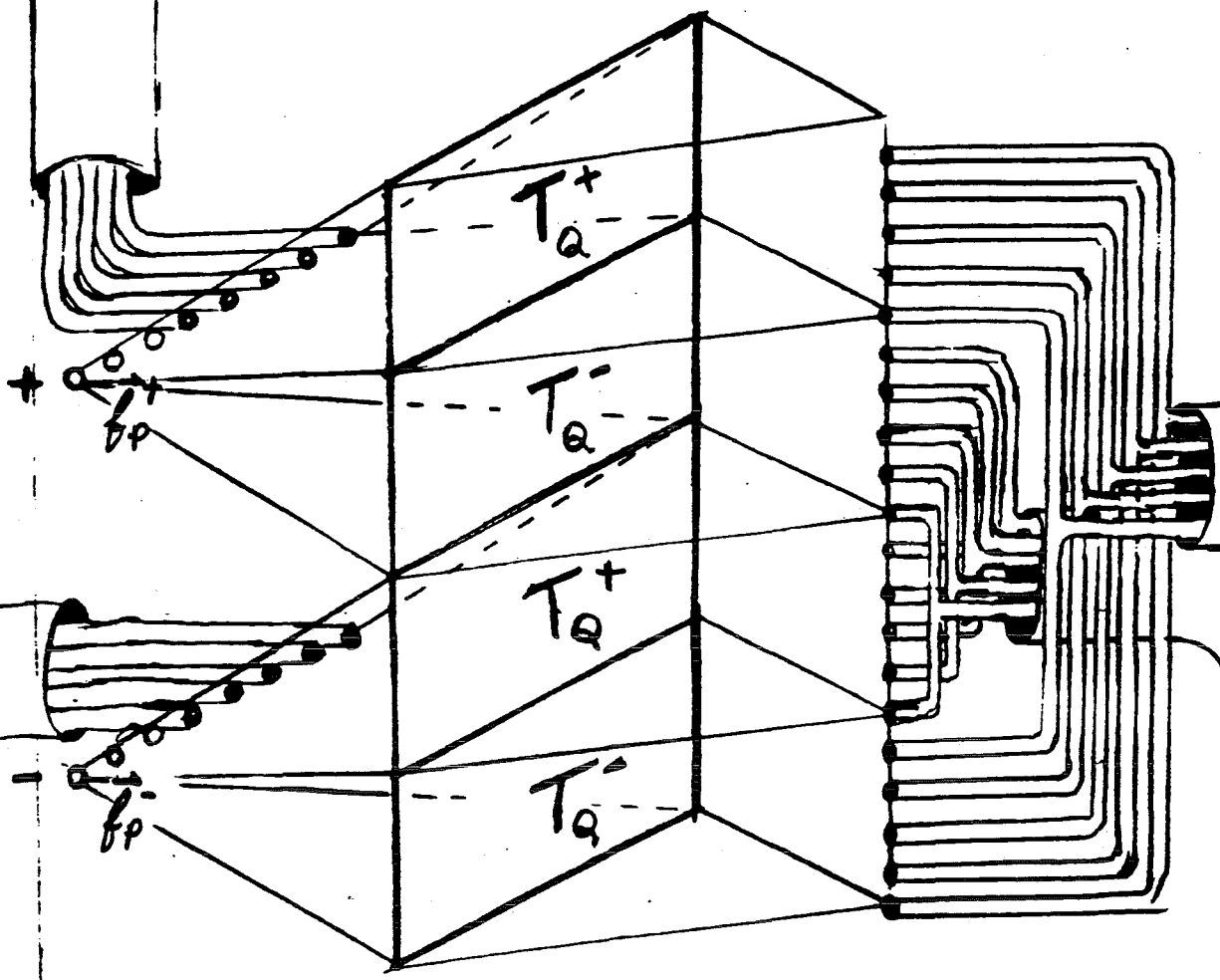
$$\vec{S}_{Q,M} = \vec{S}_{Q,M}^+ + \vec{S}_{Q,M}^-$$

Then: $\vec{S}_{Q,M+1} = T_Q [\vec{f}_P ; \vec{S}_{Q,M}]'$

becomes:

$$\vec{S}_{Q,M+1}^\pm = T_Q^\pm [\vec{f}_P^\pm ; \vec{S}_{Q,M}^\pm]' + T_Q^\mp [\vec{f}_P^\mp ; \vec{S}_{Q,M}^\mp]'$$

positive



negative

Future Work :

1. Underdetermined continuous NN performance.
2. Effects of input noise and inaccurate processing on convergence.
3. Use of stochastic processing (BHTC).
4. (a.) Identification of optical architecture.
(b.) Prototype.
5. Imposing other convex constraints.

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

Interactive Systems Design Laboratory
Department of Electrical Engineering, FT-10
Telephone: (206) 543-6990 or 543-6061

July 20, 1988

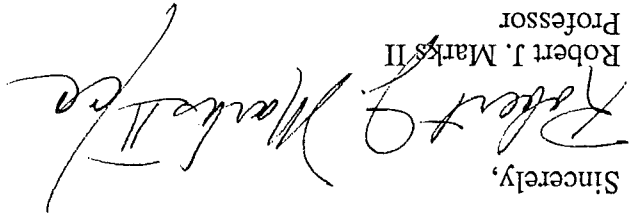
Hal Phillip
Phillip Technologies Corporation
13219 Northup Way, Suite 208
Bellevue, WA 98005

Dear Hal:

Here is a copy of a paper outlining our competition (see the table on P. 41).
Digest with a grain of salt; however, Hecht-Nielsen is a real snake oil salesman.

I'll try to find out more information.

Sincerely,


Robert J. Marks II
Professor

RJM:cc

Attachment

P.S. Update: I just talked to a neural net graduate student that says the Anza used an MC
6800 CPU and is serial. The delta floating-point processor also uses a serial chip
made by Bipolar Integrated (Something) Co. He's getting more information for us.

NON-DISCLOSURE AGREEMENT

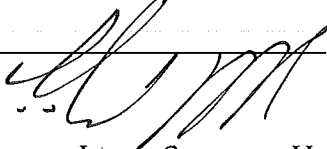
In connection with planned discussions and exchange of information between Prof. Les E. Atlas (hereinafter, LEA) and Philipp Technologies Corp., Bellevue, Washington, (hereinafter, PTC) it is understood that certain information concerning three dimensional (or volumetric) artificial neural networks as disclosed in the documents titled PATENT DISCLOSURE: THREE DIMENSIONAL ARTIFICIAL NEURAL NETWORK ARRAY dated 9/22/88 and VOLUME ARCHITECTURES FOR ARTIFICIAL NEURAL NETWORKS: A WHITE PAPER dated 12/21/88, both considered confidential by PTC, will be disclosed to LEA. LEA and PTC wish to avoid any possible misunderstanding with respect to the disclosure of confidential information and, accordingly, agree as follows:

1. All disclosures of confidential information will be in writing and marked "confidential" at the time such writings are first furnished to the other party.
2. LEA shall maintain the identified confidential information, including the document titles, in confidence for a period not exceeding three (3) years after receipt. During this period, LEA shall not divulge such information to any third party or use such information for its own benefit without the prior written consent of PTC. LEA shall treat such information with the same degree of care as he accords to his own confidential information.
3. It is understood by the parties hereto that this obligation of confidentiality shall not apply to information which is or becomes published or otherwise becomes generally available to the public through no breach of this Agreement by LEA, or information which LEA can show was properly in its possession prior to receipt of the disclosure from PTC, or becomes available to LEA from an independent source without breach of agreement or violation of law.

4. Confidential information regarding the technology disclosed hereunder shall remain the property of PTC. No license under any patent, copyright, trademark or trade secret is granted or implied.
5. Promptly after a receipt of a written request from PTC, and in the absence of such a request no later than thirty (30) days prior to the date of termination of this agreement as set forth below, LEA shall return all documents concerning the confidential information to the party who furnished such items and all copies of any such documents, subject to LEA's right to retain one copy of each such document in his files for legal counsel record purposes only.
6. This agreement shall be governed by and construed in accordance with the laws of the State of Washington.
7. Any controversy or claim arising out of relating to this contract, or the breach thereof, shall be settled by arbitration in accordance with the Commercial Arbitration Rules of the American Arbitration Association, and judgment or decision rendered by the arbitrator in any such proceeding may be entered in any court having jurisdiction thereof. The prevailing party in any such proceeding shall be entitled to receive from the other party all attorneys' fees incurred by such prevailing party and all costs incurred in connection therewith. The locale of the arbitration shall be Seattle, Washington.

This Agreement shall remain in force and effect for one (1) year from the effective date hereof, except to the extent provided in Paragraph (2) above. The effective date shall be determined by the date affixed hereto by the party last signing this Agreement.

 Les E. Atlas
 Date 12/22/88

 Philipp Technologies Corp.
 By 
 Title President
 Date 12/20/88

NON-DISCLOSURE AGREEMENT

In connection with planned discussions and exchange of information between Prof. Les E. Atlas (hereinafter, LEA) and Philipp Technologies Corp., Bellevue, Washington, (hereinafter, PTC) it is understood that certain information concerning three dimensional (or volumetric) artificial neural networks as disclosed in the documents titled PATENT DISCLOSURE: THREE DIMENSIONAL ARTIFICIAL NEURAL NETWORK ARRAY dated 9/22/88 and VOLUME ARCHITECTURES FOR ARTIFICIAL NEURAL NETWORKS: A WHITE PAPER dated 12/21/88, both considered confidential by PTC, will be disclosed to LEA. LEA and PTC wish to avoid any possible misunderstanding with respect to the disclosure of confidential information and, accordingly, agree as follows:

1. All disclosures of confidential information will be in writing and marked "confidential" at the time such writings are first furnished to the other party.

2. LEA shall maintain the identified confidential information, including the document titles, in confidence for a period not exceeding three (3) years after receipt. During this period, LEA shall not divulge such information to any third party or use such information for its own benefit without the prior written consent of PTC. LEA shall treat such information with the same degree of care as he accords to his own confidential information.

3. It is understood by the parties hereto that this obligation of confidentiality shall not apply to information which is or becomes published or otherwise becomes generally available to the public through no breach of this Agreement by LEA, or information which LEA can show was properly in its possession prior to receipt of the disclosure from PTC, or becomes available to LEA from an independent source without breach of agreement or violation of law.

4. Confidential information regarding the technology disclosed hereunder shall remain the property of PTC. No license under any patent, copyright, trademark or trade secret is granted or implied.

5. Promptly after a receipt of a written request from PTC, and in the absence of such a request no later than thirty (30) days prior to the date of termination of this agreement as set forth below, LEA shall return all documents concerning the confidential information to the party who furnished such items and all copies of any such documents, subject to LEA's right to retain one copy of each such document in his files for legal counsel record purposes only.

6. This agreement shall be governed by and construed in accordance with the laws of the State of Washington.

7. Any controversy or claim arising out of relating to this contract, or the breach thereof, shall be settled by arbitration in accordance with the Commercial Arbitration Rules of the American Arbitration Association, and judgment or decree upon any award or decision rendered by the arbitrator in such proceeding may be entered in any court having jurisdiction thereof. The prevailing party in any such proceeding shall be entitled to receive from the other party all attorneys' fees incurred by such prevailing party and all costs incurred in connection therewith. The locale of the arbitration shall be Seattle, Washington.

This Agreement shall remain in force and effect for one (1) year from the effective date hereof, except to the extent provided in Paragraph (2) above. The effective date shall be determined by the date affixed hereto by the party last signing this Agreement.

Philipp Technologies Corp.

By

Title

President

Date

12/20/88

Les E. Atlas

Date

12/22/88

BATTELLE MEMORIAL INSTITUTE,
PACIFIC NORTHWEST LABORATORIES
(hereinafter Battelle Northwest)

NON-DISCLOSURE AGREEMENT

In connection with planned discussions and exchange of information between representatives of Battelle Northwest and Philipp Technologies Corp., Bellevue, Washington, (hereinafter, PTC) it is understood that certain information concerning three dimensional neural networks as disclosed in the document dated 9/22/88 and considered confidential by PTC, will be disclosed to Battelle Northwest's representatives. Battelle Northwest and PTC wish to avoid any possible misunderstanding with respect to the disclosure of confidential information and, accordingly, agree as follows:

1. All disclosures of confidential information will be in writing and marked "confidential" at the time such writings are first furnished to the other party.

2. Battelle Northwest and its representative(s) shall maintain the identified confidential information in confidence for a period of three (3) years after receipt. During this period, Battelle Northwest shall not divulge such information to any third party or use such information for its own benefit without the prior written consent of PTC. Battelle Northwest shall treat such information with the same degree of care as it accords to its own confidential information.

3. It is understood by the parties hereto that this obligation of confidentiality shall not apply to information which is or becomes published or otherwise becomes generally available to the public through no breach of this Agreement by Battelle Northwest, or information which Battelle Northwest can show was properly in its possession prior to receipt of the disclosure from PTC, or becomes available to Battelle Northwest from an independent source without breach of agreement or violation of law.

4. Confidential information regarding the technology disclosed hereunder shall remain the property of PTC. No license under any patent, copyright, trademark or trade secret is granted or implied.

5. Promptly after a receipt of a written request from PTC, and in the absence of such a request no later than thirty (30) days prior to the date of termination of this agreement as set forth below, Battelle Northwest shall return all documents concerning the confidential information to the party who furnished such items and all copies of any such documents, subject to Battelle Northwest's right to retain one copy of each such document in the files of its law department or outside legal counsel for record purposes only.

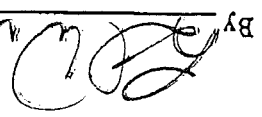
6. This agreement shall be governed by and construed in accordance with the laws of the State of Washington.

7. Any controversy or claim arising out of relating to this contract, or the breach thereof, shall be settled by arbitration in accordance with the Commercial Arbitration Rules of the American Arbitration Association, and any court having jurisdiction thereof. The prevailing party in any such proceeding shall be entitled to receive from the other party all attorneys' fees incurred by such prevailing party and all costs incurred in connection therewith. The locale of the arbitration shall be Seattle, Washington.

This Agreement shall remain in force and effect for one (1) year from the effective date hereof, except to the extent provided in Paragraph (2) above. The effective date shall be determined by the date affixed hereto by the party last signing this Agreement.

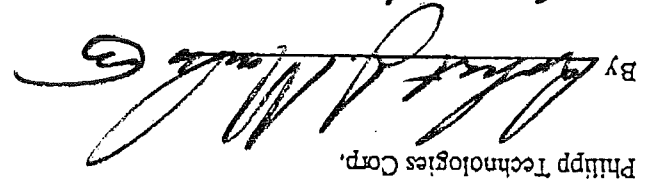
BATTELLE MEMORIAL INSTITUTE
PACIFIC NORTHWEST LABORATORIES

XXXXXXXXXXXX

By 

Title Contracting Officer

Date 9-30-88

By 

Philipp Technologies Corp.

Title Consultant

Date 9-30-88

A G R E E M E N T

We, the undersigned, Pieter J. van Heerden, Robert J. Marks II, and Seho Oh agree that the owner's rights and the financial benefits of the patent we will apply for "A Computer Chip Realizing Learning in a Digital Computer," will be distributed in the following manner: P. J. van Heerden 60%, R. J. Marks 20%, and Seho Oh 20%.

This means, of course, that the cost of obtaining the patent, specifically the cost of the patent lawyers, will be subtracted from the income from potential licensing rights and the sale of the patent.

Signed:



P. J. van Heerden

March 30, 1988



R. J. Marks II

April 4, 1988



Seho Oh

April 4, 1988

PATENT APPLICATION

A Computer Chip Realizing Learning in Digital Computers. Pieter J. van Heerden, Robert J. Marks II, and Seho Oh.

Introduction

The invention relates to a computer chip which is the 'brain' of a digital computer which can learn, that is improves and perfects its performance from previous experience. The computer may be designed to operate any kind of mechanical or electronic equipment normally operated by human beings. If we call the learning to improve the operation "Intelligence," then we may call the computer an intelligent machine. Intelligence is a quality observed in living beings, humans and animals. Since, in learning, the machine imitates the learning behavior of living beings, its operation is based on a theory of human and animal behavior. This is the theory of psychology of William MacDougall, given in his book "An Introduction to Social Psychology" Barnes and Noble, N.Y. 1960 (originally 1908). This book is out of print, but the present champion of this theory is Margaret Boden with her book "Purposive Explanation in Psychology" Harvard Un. Press 197....

The psychological theory is that man or animal does everything with a purpose. They have drives or instincts which want to be satisfied. The simplest example is hunger. Hunger drives the intelligent being to seek means to satisfying its hunger, by eating. How else would an animal know that, speaking physiologically, its body needs food to stay alive? Internal observations on the body, which could be for instance cells which measure the sugar content of the blood, are communicated, as a communication channel, to the brains. The measurement of a low-level of the sugar content of the blood results in a feeling of hunger.

MacDougall hypothesis is that man has, besides hunger, a whole spectrum of drives which want to be satisfied. These drives take care, not only of the bodily needs to the living being, but also of its social needs. Examples of these drives are: The curiosity propensity, which is the instinct to explore strange places and things; the self-assertive propensity, the instinct to domineer, to lead, to assert oneself over, or display oneself before one's fellows; the submissive propensity, the instinct to defer, to obey, to follow, to submit in the presence of others who display superior powers; the gregarious propensity, the instinct to remain in the company with fellows, and, if isolated, to seek that company; the anger propensity, which is the instinct to resent and forcibly break down any thwarting or resistance offered to the free exercise of any other propensity; the fear propensity, the instinct to flee for cover in response to violent impressions that inflict or threaten pain or injury; the constructive propensity, the instinct to construct shelters and implements; the acquisitive propensity, the instinct to acquire, possess, and defend whatever is found useful or otherwise attractive; the sex propensity, the instinct to court and mate; the parental or protective propensity, the instinct to feed, protect and shelter the young; the laughter propensity, the instinct to laugh at the defects and failures of our fellow creatures. In Boden are listed 18 propensities, which list is pretty complete, but obviously not claimed to be exhaustive or final. Many of these instincts are readily observed in animals, and, just like hunger, it is hard to imagine how the individual would survive without having these instincts. P. J. van Heerden has translated this psychological theory into a quantitative mathematical theory of intelligence. "The Foundation of Empirical Knowledge, with a Theory of Artificial Intelligence" Wistik, Wassenaar, Netherland 1968. The book is out of print, but available in many college libraries.

The brain is postulated to be a computer. Its only function is to process mathematically the input of information into an output. The input information consists of two kinds of information channels. The first kind is a spectrum of drives;

they are all of the same kind. They represent functions of time, $f_1(t)$, which represent some internal observation of an aspect of the state of the body, like hunger observes that the body needs food. In a primitive animal, the needs are few. In an animal of higher intelligence, these needs are differentiated and more refined. But their mathematical representation is always the same, $f_1(t)$, a drive which drives the individual to action, to satisfy that drive.

The second kind of input information channels, $f_2(t)$, is from the senses like eye and ear. It is obvious that without eyes and ears, the individual could never carry out intelligent actions. The senses therefore, in general, are observations on the state of the outside world which help the instincts $f_1(t)$ which are observations on the internal world of the individual, to satisfy their purposes. The output information $f_3(t)$ is of one kind. It gives commands, through the nerves, to the muscles of the body, of hand, foot and mouth. Speaking is also action, and it is hard to imagine intelligent human life without this means of communication with its fellow men. Of course, in our society, it is partly replaced by muscle action of the hand in writing.

Consequently, the mathematical theory translates the psychological theory of MacDougall into a universal quantitative theory of intelligence, and this theory is valid for all intelligence, whether in man, animal or machine, the brain is an organ, like the heart is an organ which pumps blood. The brain is a computer. It takes the input information, the drives and the senses, $f_1(t)$ and $f_2(t)$, and processes this information to arrive at an output, which is a command to the muscles to action. This command to the muscles has always only one purpose, which is to satisfy, silence, the drive which happens to be active. The sense information, $f_2(t)$, helps in performing intelligent action, that is action which leads to a better, easier, quicker satisfaction of the drives. In the case of a machine, the mathematical theory simply imitates the behavior of intelligence which we observe in man and animals. The output function $f_3(t)$ can be a typewriter, but in general any kind of machinery which one wants

the computer to drive. The input functions $f_2(t)$, the senses, are any kind of information about the outside world one wants to make available to the computer. In the most advanced machine intelligence, this could be a television camera to see, and a microphone to hear. The functions $f_1(t)$, the drives, are conceptually the most difficult part of the machine. The builder of the machine of course wants to put in functions $f_1(t)$ which satisfy the builder's purpose, but makes the machine so that it works independently, without further instructions. The more refined the drives, the higher the intelligence the machine will be able to realize. Let us say that we consider the design of the functions $f_1(t)$ as an open art, which requires psychological insight in what motivations lead to higher intelligence in humans. But, to construct machines, we should be able to give a simple function $f_1(t)$ which accomplishes the basic goal of a primitive intelligence. This is the element of reward and punishment, without which no intelligent being could survive. And we know of no better way of demonstrating this than in an eating experience, of humans or animals. When we see an apple, and we know apples as delicious to be eaten, and as satisfying, stilling our hunger drive, we are inclined to take a bite. If we take a bite, then the first bite gives us the delightful taste of the apple juices flowing in our mouth. This experience encourages us to take a second bite. On the other hand, a bite of an apple which, by some treatment, is foul tasting, or is rotten, or has a worm in it, it will cause us to spit it out, throw the apple far from us, give us a warning to look carefully at an apple before we take a bite. This is the clearest, simplest example of reward or punishment in human situations which requires no further explanation. It is a rock-bottom experience. Yet, when we think about putting the principle in a machine, we are stopped in our tracks: how do you reward a machine, how do you punish a machine? Unless we can do this, we cannot build an intelligent machine. We want to propose a way in which this simple human, or animal sensation is realized in a machine. We think that in humans or animals, the reward causes a feeling of well being, which at all times encourages muscle actions, while punishment causes a feeling of diminished well

being, which therefore discourages muscle action. In all our actions, there is a desire to act before it reaches a level of actions. Humans and animals always have a certain caution, a feeling that action might be harmful, dangerous. Unless the individual has a high confidence level, partly caused by a clear signal from the eyes and ears that nothing is wrong, that the situation is clear cut, the individual will not act. An example is that at night, under circumstances of less visibility, we won't drive our car with as much confidence as in daylight. Another example is that one wants to ask a question at a public meeting, but hesitates to ask it because one thinks one has not properly understood the situation, and might make a fool of one self. Confidence, desire for action, is partly caused by internal factors, for instance feeling robust and healthy, partly by a clear recognition, by the senses, of the situation one is in. We therefore want to propose a state of the body always present, of a bias level or rather the reverse, a "boldness level," that discourages action when it is low, and encourages action when it is high. It is like the grid voltage on a power tube which operates the motion of a power tool. At high voltage, the power tube works at full power, at low voltage, the current in the tube is cut off, the tube does not generate power.

This "boldness level," which like all our mechanical descriptions of humans, is felt as a sensation, is raised by a positive experience, that is felt as a reward; It is lowered by a bad experience, that is felt as a punishment. A bite of the apple, which tastes good, heightens the activity of the muscles in what we are doing, eating. A bad tasting bite lowers the activity of the eating muscles. So the taste cells in our mouth and nose, we hardly realize this in everyday life, form an essential mechanism given us by nature to raise our intelligence, because it increases our power of discrimination between good and bad. So, in general, such observation cells, which discriminate between good and bad action, are essential in an intelligent machine.

In teaching humans, we use reward and punishment by a show of either approval and affection, or disapproval. There is no

reason why in teaching intelligent machines we cannot use the same principle. A proper operation raises the boldness level; a bad operation, a barrier, a wrong executed motion, lowers the boldness level. But, also a simple push button, operated by the teacher, raises the boldness level. Another button lowers the boldness level.

So, in general, we must realize that in all human intelligent action, there are present multiple drives. While in eating the general drive is hunger, the very pleasure of eating something wholesome, appetite encourages the action of eating. In reaching out for instance to grasp an object, the brain cannot give precise instructions to the muscles, because the distance and the shape of the object is unknown. The eye monitors the movement of the arm to the object, and when the hand touches, the feeling of touch guides the muscles of the fingers to perform the proper grip action. This may signify that there is a certain innate pleasure of gripping, similar to appetite in eating, which has to be simulated in a machine to achieve proper learning. In the same way, in speech, hearing one's own speech, and the pleasure derived from it, may be an essential element in proper speech, and therefore has to be incorporated, simulated, in a machine.

In all human intelligent actions we have this principle that action of the muscles cannot be prescribed in detail at the beginning of the action, but must be specified in the course of the action by closer observations, and, as we have shown this may require the postulation of special appetite-like mechanisms. One may compare this with the order of an army general to move his army forward to engage and defeat the enemy. This order is intelligent action, motivated by his sense of patriotism, loyalty to his country, or personal ambition. However, he cannot prescribe to the individual platoon, or to the individual soldier how he has to move forward, since this depends on the exact terrain, and the conditions under which the soldier operates. There must therefore be a detailed motivation of the soldier, a desire for individual combat, a desire to avoid enemy fire, a willingness to obey orders, or, at the lowest level, to put his

foot forward without stumbling. But all this detailed action is integrated in the total intelligent action of the general, to defeat the enemy.

This discussion shows clearly that, before we can properly describe intelligent action in detail, and therefore incorporate its imitation in machines, a lot of thought and experimentation will still be necessary. However, this does not influence in the least the nature and purpose of our invention, which merely is aimed at finding the information which is appropriate to the situation at hand, no matter what the instincts, no matter what the information from the senses, no matter what the muscle action. We claim that intelligence is nothing but the modification of the behavior, from the automatic reactions of a newborn babe to its inborn instincts, by its life experiences that its instincts, like hunger, can be satisfied in different ways. Learning is always choosing the best experience in the past, and of the past, those experiences are stored in the permanent memory which have best satisfied the inborn instincts. The object of our invention therefore is merely a circuit which can carry out a search of the past experience, as fast, and therefore through as many parallel channels, as possible. Such a circuit, for information in the binary form, can be made on the surface of a silicon crystal, by the modern, conventional methods used for instance in the well-known manufacturing methods of a RAM device. The methods are micro photography of the circuit, or parts thereof on a photo-sensitive layer on the surface, etching, oxidation, vapor deposition or vapor reaction, etc. The idea is simply to process information in as many parallel channels as possible.

To sum it up, we believe that intelligent action in animals and man is not based on a mysterious, magic principle but on learning, that is rational information processing. It is only the enormous amount of information processing involved in the brain which is difficult to comprehend. Certainly the details of the kind of information, and information processing, and in particular, the information from drives and senses, will require a great deal of experimental and theoretical research. But it

seems clear that at the center of the operation of intelligence lies the fast search for accurate information of positive experiences of the past, and that the circuit of our invention carries that out in a machine. How far the operation of such a circuit alone will go in achieving intelligence in a machine, whether only of a simple nature or advanced, only future investigations can show.

The Mathematical Principle Involved

The mathematical principle involved in the theory was described in the book "The Foundation of Empirical Knowledge" mentioned before and realized in the optical machine described there and in two U.S. patents of P. J. van Heerden, #3,296,594 and #3,492,652. It is basically the principle that at every moment of the life of an intelligent individual the brain carries out a rapid search in its permanent, or temporary, storage for that positive life experience which most closely matches the present situation it finds itself in. If the information is given in the form of three functions $f_1(t)$, $f_2(t)$ and $f_3(t)$, mentioned in the introductions, then this search produces automatically the function $f_3(t)$, the muscle commands, for hand, foot or tongue, which the individual needs in the present situation. This muscle motion $f_3(t)$, which, in a specific case for human beings, may be nothing more than speaking the right words, as learned from a previous experience, is now produced automatically. It is but the muscle motion $f_3(t-k)$ which was successful in a previous satisfying experience k units of time ago. It is the claim of the theory that all muscle motions are learned by practice as successful in satisfying the drive function $f_1(t)$.

The optical machine, with a hologram, described in the book mentioned and the patents, formed a fast and accurate way in which this rapid search for matching information could be carried out in a large memory. It was believed, at the time, that an equivalent search for information could never be carried out by a digital computer. At present, because of the great development of making complicated digital circuitry on the surface of a semiconductor, the digital computer has the capability of matching the performance of the hologram principle in the optical machine. Stating it mathematically, if the number of binary digits equivalent to the information storage in a memory of an intelligent individual is the number n , then the number of elementary algebraic operations for search ((like $(A+B)=C$, when A , B and C are binary digits)) in a life time is n^2 . At present, the informa-

tion stored in the human brain in a lifetime is estimated of the order of 10^9 binary bits (Th.K. Landauer, Cognitive Science 10,477,1986). This means that 10^9 to 10^{10} elementary binary operations have to be carried out every second. With a clock time of a micro second (10^{-6} sec), 10^6 such operations can be carried out in one channel per second, and therefore a computer chip with 10^4 parallel channels is necessary for carrying out the equivalent of the information processing that goes on, according to our theory, in the human brain.

The complexity of the circuit of our invention, described here, necessary to carry out this amount of information processing is estimated to be like that of a 400K RAM, which units at present are manufactured on a large scale. The circuit of our invention therefore can be manufactured in the conventional ways of making the complex circuitry required for modern computers. Improvements in technology, leading to more parallel channels, and a faster clock time, will improve machine operations.

The Circuit

The circuit has as an input one binary time series $f(t)$, which therefore is a series of standard pulses, which represent a "one," or the absence of a standard pulse, which represents a "zero." However, according to our mathematics, this procedure can also be reversed, in that a "zero" can be represented by a standard pulse, and a "one" by the absence of a standard pulse. The function $f(t)$ is periodically divided in the three functions $f_1(t)$, $f_2(t)$ and $f_3(t)$ mentioned before, so that for instance a fixed period of a sequence of pulses represents $f_1(t)$, a second sequence represents $f_2(t)$, and a third sequence represents $f_3(t)$, a fourth sequence represents again $f_1(t)$, a fifth sequence $f_2(t)$, a sixth sequence $f_3(t)$, a seventh sequence again $f_1(t)$, and so forth, so that $f(t)$ represents the full history of positive experiences of the machine. This history of experiences of the machine will be stored, temporarily or permanently, on a magnetic tape or disc or other electronic storage medium. As we will see,

memories of a different content will have to be present in the machine, and our invention does not cover the wiring of together of these memories. Our invention only covers the circuit on the chip, which is the same independent of the kind of information it processes. It is always involved in search for the best match with the content of the memory.

The complete function $f(t)$ is fed into the circuit on the chip, and the operation of the circuit is to form the binary functions $(1+D^1) f(t)$, $(1+D^2) f(t)$, $(1+D^3) f(t)$, and so on, in general $(1+D^k) f(t)$. Here "+," for the binary function, means "plus modulo two": $1+1=0$; $1+0=1$; $0+1=1$, $0+0=0$, and D , the so called "Huffman operator," is defined as $D^1 f(t) = f(t-1)$, $D^k f(t) = f(t-k)$. Here the limit of time is chosen the clock time, the time in which a pulse, or "absence of a pulse," repeats itself. Therefore $(1+D^k) f(t)$ represents the new function $[f(t)+f(t-k)]$.

The operation of the circuit is now to select that function $(1+D^{k^0}) f(t)$ which produces the highest percentage of zero's (over one's) in the most recent past interval, the length of which interval can be specified by appropriate circuitry and may be variable. This selection of the function $(1+D^{k^0}) f(t)$ is realized in the circuit drawings by the method used for selecting the best player in tennis tournaments. One first matches each two players, and then matches the winners again two by two, and so on, until finally one winner emerges of the tournament. So, in the circuit of our invention, each pair of adjacent functions, $(1+D^k) f(t)$ and $(1+D^{k+1}) f(t)$ are compared on this excess of zero's content, by having a counter for each function counts the excess of zero's content, and then have the highest excess of zero's counter determine the switch setting. The "winner" is thus admitted to the "next round," and the winner of the "pairing of winners of the first round" is determined by the B- or comparator circuits. These comparators B again operate switches to admit the winners to the next round. Finally, one winner $(1+D^{k^0}) f(t)$ emerges as the one who has produced the largest excess of zeros of all the circuits $(1+D^k) f(t)$ on the chip.

This function is added, in the binary way, to $f(t)$, according to the formula $(1+D^{k^{\circ}})f(t)+f(t)=D^{k^{\circ}}f(t)=f(t-k)$. The part $f_3(t-k)$ represents the "muscular" output of the machine. It operates the mechanical or electric apparatus one wants the intelligent computer to operate.

However, there is a mismatch of the operations of the computer circuit and the operation of intelligence in man and animals. This mismatch is the fact that a computer may have a clock time of one micro second, and processes information at a speed of 10^6 digits per second, while the human intelligence receives only a fraction of 10^6 digits per second. Let us say 10^4 , 10^3 , 10^2 or less digits. That would mean that the circuit would have to be idle the larger part of the time. Without further tricks, to match the inherent speed of the computer with the estimated capability of the human brain to process 10^9 to 10^{10} digits per second, the very speed of the computer would be useless.

Therefore, some of these circuits present in an intelligent machine work not in real time, but from a memory of storage of past information, call it $g(t)$, to differentiate it from the real lifetime experiences $f(t)$, in such a way that it scans successive segments of $g(t)$, given by a segment of pulses $g(t-a)$ to $g(t)$, $g(t-2a)$ to $g(t-a)$, $g(t-3a)$ to $g(t-2a)$, $g(t-4a)$ to $g(t-3a)$ and so on. However, the winner $f(t)+D^{k^{\circ}}g(t)$ in each segment, has to be added to $f(t)$ to give an output $D^{k^{\circ}}g(t)=g(t-k)$. Therefore, at the point in the circuit T we have for every segment repeat the function $f(t-a)$ to $f(t)$, while in the shift register D should appear, successively, the segments of $g(t)$, to wit $g(t-a)$ to $g(t)$, $g(t-2a)$ to $g(t-a)$, and so on. Clearly, we can only achieve a smooth operation if the number a (in units of time of one micro second) is equal to the number of processing units in the shift register, on a simple fraction there of ($1/2$, $1/3$, $1/4$, etc.).

Discussion

We must realize that in achieving human intelligence - and even much more primitive intelligence in animals - in computers a great deal of experimenting and thought will be necessary. And the kind of experimenting involved will have to be both of an engineering nature, in the way the circuits execute their purposes, and the way they are connected in the general organization of the intelligent machine, and of a scientific nature, on what kind of drive - and sense - information is conducive to develop intelligent behavior.

For instance, one must realize that in intelligent action different kinds of intelligence are involved (as in our example of the general and the soldier), which therefore require different kinds of information storage, permanent or temporary, and circuits searching them. The claim is however that in this general organization of an intelligent machine the circuit of our invention plays the essential role. The action of the circuit is to search fast for that kind of information that is applicable to the present situation. That is all it does, and that is, as is claimed by our theory, the basic element in developing intelligence. No doubt, by this principle, the machine will learn, since it will recollect past positive experiences, and apply them to the present. Like in all scientific theories, future experiments with these circuits will show us the level of intelligence that can be reached.

Intelligence is learning by experience; that is learning by experience that kind of actions - muscular activity, including speech - which satisfies the drives Nature has endowed us with. These drives of course are "survival instincts." They are necessary for the individual to survive, and thrive, in its surrounding. And this surrounding can be its group, its tribe, its society. No doubt, life developed these drives in the millions of years of evolution of life, and its changing circumstances.

We think that all drives the baby has at birth are accompanied by an automatic response, in muscle actions. When a baby is hungry, it cries; when it is offered the mother's breast, it sucks. Then, in the course of time, it discovers, by experience, that there are other ways to satisfy its hunger.

While, in the life of the intelligent individual, the response to a drive has to be modified, to reach the age of a mature individual, and some drives may develop in adolescence, other drives present in the baby stage may not require modification. Let me give two examples. When an object comes close to the eyes, or touches the eyeball, we will automatically blink to protect the eyes. This is an automatic and intelligent response which does not need modification (except in exceptional cases, like a prize fighter who is taught not to blink when he sees a fist coming). When we touch a hot object with the fingers, so we burn ourselves, we will automatically pull back our hands. This is again an automatic response which does not need modification. P. J. van Heerden, in his book, has also pointed out that learning to see what it sees, a baby requires the curiosity drive from birth to direct the eyeballs, and focus the eye lenses, to an object appearing in its field of view. All these mechanisms may be imitated in machines, or mechanisms may be invented to serve the particular purpose.

It is clear that those inborn intelligent responses, and also the learned intelligent responses, like in speech, or moving the hands and fingers as we learn it in the crafts, professions and sports, a fast response is necessary. To scan the full stored intelligent memory in that short a time is physically impossible. A limited search, through a smaller memory, is necessary. This makes it clear that in intelligent machines, as in intelligent living beings, several kinds of memory storage are necessary, in which also different scanning times, and zero count integration times, are required. For instance, in speech, the amount of information flowing from our lips may be expressed in hundreds or tens, of binary digits per second. According to our theory, this is accomplished by a circuit with a fast counting integration

time scanning a small memory containing only words, and short sequences of words, of the language. But it is hard to imagine that this fast circuit also would contain the main purpose a person has with his conversation. Therefore, one must imagine that speech is guided by several circuits, or rather (since circuits only deal with intelligence in machines), one should say several memories. One memory controls the details of correct speech, from a limited content, and one which controls the overall purpose of the conversation one has. In the efforts to build intelligent machines, this must be taken into account.

To sum it up, our theory of intelligence is like a theory of human flight. The Wright brothers proved that flight is possible. The essential elements were: a wing to support the airplane, a motor-driven propeller to give it speed, and a steering mechanism to guide its motion. However, nobody could build a Boeing 747 in 1905. That would take humanity 70 years of learning. But, the same elements used by the Wright brothers are still the elements of flight: wing, motor and steering, except that now we use jet engines. In the same way, our theory of intelligence says that three elements are necessary to make intelligent machines: drives, senses and ^{commands to} muscles, and that the essential operation is learning from past experience. In the digital computer, this learning is carried out in circuits of our design, and the availability of these circuits will be vital for developing intelligence in digital computers.

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON

*Interactive Systems Design Laboratory
Department of Electrical Engineering, FT-10
Telephone: (206) 543-6990 or 543-6061*

November 4, 1988

Dr. Pieter J. van Heerden
18217 - 145th Court, N.E.
Woodinville, WA 98072

Dear Pieter:

I hope the enclosed meets the needs of the patent attorney. If not, please let me know.

Best personal regards,



Robert J. Marks II
Professor

RJM:cc

Enclosure

cc: S. Oh

ELECTRONIC CIRCUITS

Digital and Analog

CHARLES A. HOLT

Virginia Polytechnic Institute and State University

JOHN WILEY & SONS, New York • Chichester • Brisbane • Toronto

that the \bar{K} input is inverted ahead of the K terminal. This feature is useful in counting and sequence-generation applications. When the JK inputs are connected together, the first stage is a D flip-flop, which is examined in the next section. All flip-flops are master-slave (MS) types with *static* operation, in contrast to the *dynamic* mode of multiphase configurations such as that of Fig. 8-7 of Sec. 8-3.

9-4. CMOS FLIP-FLOP CIRCUITS

D FLIP-FLOP

RS, JK, and T flip-flops have been examined. Another configuration of importance is the *D type*, with D representing *delay*. The output after a clock pulse equals the input before the pulse. In Fig. 9-15 are shown the symbol and characteristic table. Optional are the clear-preset inputs and the complement \bar{Q} of the output. A D flip-flop can be made from a JK flip-flop by connecting the J and \bar{K} inputs, with the connection serving as the data input. When periodic pulses are applied to the clock input, the output is that of the input delayed by one clock pulse.

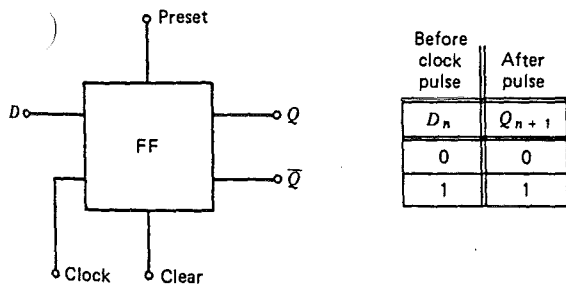


Figure 9-15 D flip-flop and characteristic table.

A clocked *D latch* differs from a D flip-flop in that the one-bit delay is eliminated. The network is designed so that when the clock pulse triggers the gate, the output is coupled directly to the input *D*, and *Q* equals *D*. The output is then held, or *latched*, in this state until the next pulse triggers the gate. The clock simply acts as an enable input to the latch. It has important applications in registers, especially for temporary data storage.

The logic diagram of a CMOS clocked D latch is shown in Fig. 9-16. Transmission gate TG_2 turns *off*, and TG_1 then turns *on*, during a clock-pulse rise from low to high. The reason TG_2 turns *off* is to prevent the output at Q_2 from interacting with the data input. When TG_1 turns *on*, the input bit enters

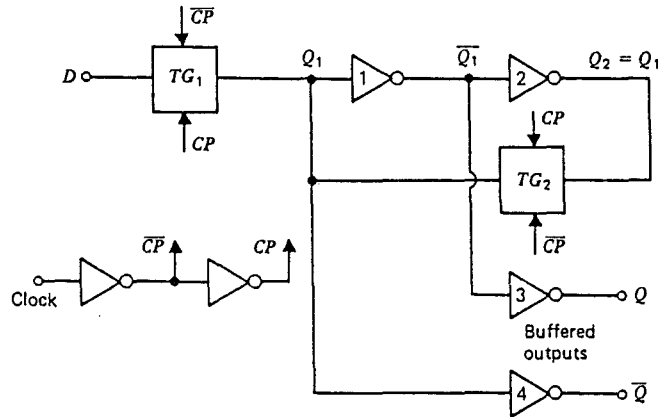


Figure 9-16 CMOS clocked D latch.

the latch. This stored bit appears at the buffered output terminal Q with very little delay. The propagation delays of the inverters are small compared with the clock period.

When the clock pulse drops from high to low, TG_1 cuts off, TG_2 then turns on, and the bit remains stored until the next pulse appears. The reason for the inclusion of inverter 2 and TG_2 is to maintain the proper stored charge on the insulated gate terminals of inverters 1 and 4. If they were eliminated, any charge stored on these insulated gates would soon be lost by leakage. With TG_2 on, inverters 1 and 2 constitute a cross-coupled latch. Transmission gates are used instead of NOR gates to control the operation. The use of two inverters for the clock circuitry provides buffering to reduce the loading of the clock and to improve the pulse waveforms.

Integrated circuit CD4042A is classified as a CMOS quad clocked D latch. It consists of four separate D latches, each strobed by a common clock. The configuration is that of Fig. 9-16. A polarity circuit of two cascaded inverters can be used to program the pulse transition, either positive or negative, that switches the output. The gate propagation delay is typically 50 ns with a 10 V supply and a load capacitance of 15 pF, corresponding to a fan-out of three. In the low state the gate can sink about 2 mA while maintaining an output less than 0.5 V, and in the high state it can supply 2 mA with the voltage held above 9.5. A toggle frequency up to about 8 MHz is reasonable. Typical applications include buffer storage and use as a holding register in digital systems.

A D type master-slave (MS) flip-flop can be made simply by cascading two D latches of the form of Fig. 9-16, with the transmission gates clocked so that only one latch receives data at a time. By replacing inverters 1 and 2 of each latch with NOR gates, preset and clear controls can be added, which are often referred to as *set-reset* controls. Such a configuration is shown in Fig. 9-17, along

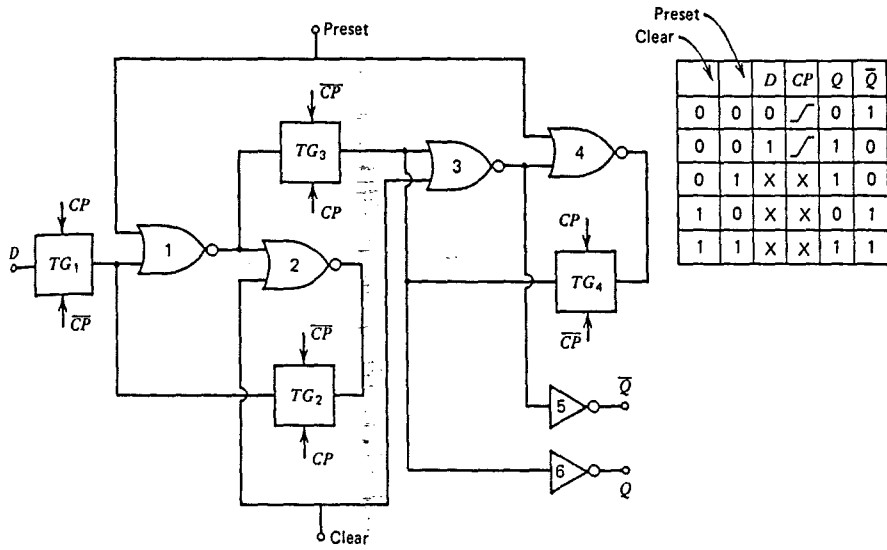


Figure 9-17 Logic diagram of D-type master-slave flip-flop.

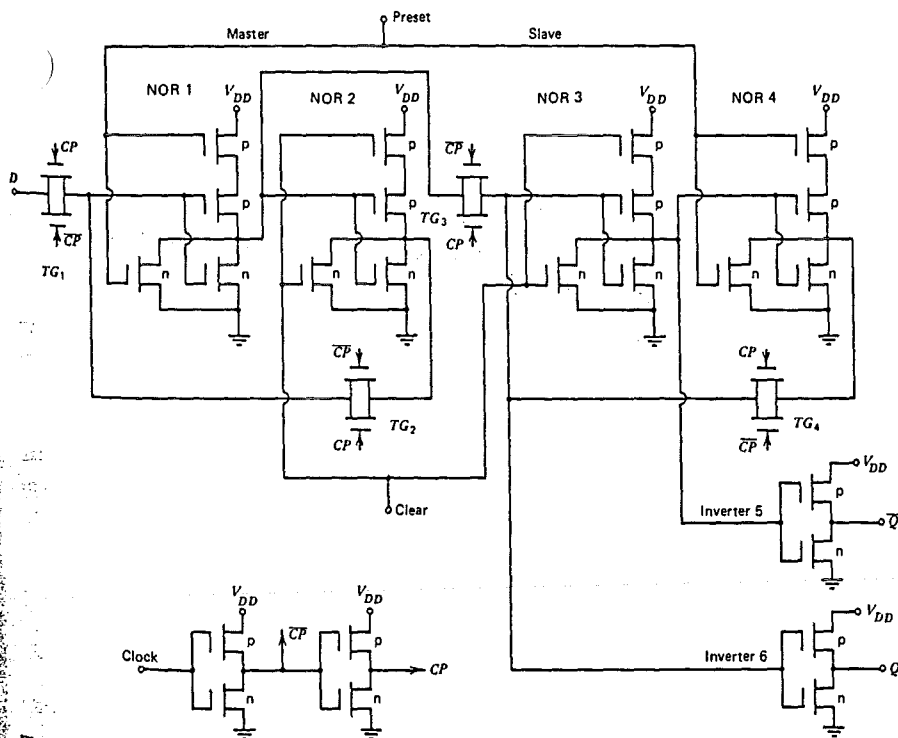


Figure 9-18 D-type master-slave flip-flop circuitry.

with the truth table. When a clock pulse rises from low to high, which is a positive transition, the logic level present at the D input becomes the Q output. Data enter the master on negative transitions and are transferred to the slave on positive transitions.

The first two rows of the truth table are those of a D flip-flop, with the symbols under the clock column indicating the level change at which the D input becomes the Q output. The bottom three rows simply represent the truth table for the case in which one or both of the preset-clear inputs is a logical 1. The states and clock transitions marked X have no effect on the output. These are referred to as *don't-care* conditions. When a logical 1 is present at a preset or clear input, the output is independent of the data input and the clock pulses.

CMOS FLIP-FLOPS

Circuitry accomplishing the logic of Fig. 9-17 is shown in Fig. 9-18 with each gate identified. Both the numbers and the relative positions of the gates of Fig. 9-18 correspond with those of Fig. 9-17. Note the symbol used for the transmission gate. Because transmission through a gate is possible in both directions, the position of the gate terminal is centered. All transistors are enhancement-mode devices.

Integrated circuit CD4013A consists of dual D type flip-flops. Each of the two identical flip-flops has the circuitry of Fig. 9-18. Operation is static, rather than dynamic, with the state of the flip-flop retained indefinitely when the clock input is constant at either a high-level or low-level voltage. A toggle rate of about 8 MHz is typical with a 10 V supply, and the respective high-level and low-level output impedances are typically 400 and 200 ohms. The dc supply V_{DD} should be between 3 and 15 V. By connecting the \bar{Q} output to the D input the flip-flop toggles at each clock pulse. Applications include shift registers, counters, and control circuits.

A D flip-flop can be converted to a JK configuration in a number of ways, one of which is shown in Fig. 9-19. In addition to the logic arrangement, the figure includes the characteristic table, assuming zero preset and clear inputs. Let us consider the first row of the table. With the present state of Q equal to 0 and the input at J a logical 1, the outputs of gates 1 and 2 are logical zeros regardless of K , which is a don't-care state. Therefore, the output of NOR gate 3 is 1. As this is the D input, Q becomes 1 after the positive pulse transition. Verification of the other rows is left as an exercise. The process is simplified by recognizing that the output D of NOR gate 3 is given by

$$D = \overline{KQ + J + Q} \quad (9-1)$$

with Q denoting the present state. Output D is the next state of Q .

Figure 9-20 shows suitable circuitry that performs the logic of (9-1), along with the proper connections to the D flip-flop. All p-channel transistors have a

University of Washington Correspondence

INTERDEPARTMENTAL

=====

*Interactive Systems Design Laboratory; Department of Electrical Engineering, FT-10;
(206) 543-6990 (office), 543-6061 (secretary), 543-2150 (main office), 776-8995 (home), 543-3842 (FAX).
marks@blake.acs.washington.edu*

5-1-90

TO: Prof. Tom Seliga, Chair *TDS*
Department of Electrical Engineering

Dr. Ray Bowen, Dean
College of Engineering

Prof. Ed Stear, Director
Washington Technology Center

Peter Odabashian, External Affairs Director
Washington Technology Center

FROM: Robert J. Marks II
SUBJECT: Patent Disclosure *RAMS*

This attached disclosure requires approval from Profs. Seliga, Bowen and Stear. If approval is given, please forward this memo to the next person on the list. Otherwise, it should be returned to me.

The subject of this disclosure is a computational procedure to adapt training in an artificial neural network to nonstationary training data. The technology was developed by Prof. El-Sharkawi, Mr. D.C. Park and me. The work was performed under the sponsorship of *Puget Sound Power and Light Company* and was motivated by the need to adapt load forecasting to the changing load profiles. The adaptive technique, however, is potentially applicable to a large number of similar problems where the training data for the neural network comes from a slowly varying nonstationary process.

cc: Prof. M. El-Sharkawi
D.C. Park
M.L. Bruce, *Puget Sound Power and Light Company*

Volume Architectures for Artificial Neural Networks

a White Paper

by
H. PHILIPP AND R. J. MARKS II

12/21/88

there are nine pages plus six figures in this document

Introduction

Artificial neural networks (ANN's) attempt to simulate the architecture and operation of their biological counterparts. While considerable effort has been made to create implementation electronics, most efforts to date have either [1]

- ◆ concentrated on using conventional high speed serial computers designed on a highly planar structure or

- ◆ have used high speed planar analog electronics.

The serial electronics have been primarily marketed as simulation tools. ANN's, however, are inherently parallel and serial implementation severely degrades potential speed possibilities. The analog electronics approach is superb for certain ANN operations (e.g. Hopfield ANN's [2-4] or recall from a trained ANN), but the poor accuracy of analog operations makes it ill suited for the high precision required of currently used adaptive and learning algorithms (e.g. back propagation [5-6]).

Furthermore, the planar approach to both serial digital and analog ANN VLSI architectures is in contrast to the parallel three dimensional structures found in many biological neural systems. The high connectivity available in three dimensions is clearly not available in two.

This white paper outlines a method for overcoming these problems. Specifically, we propose development of an electronic architecture for volumetric artificial neural networks (VANN's) with the following characteristics:

- ◆ Required electronics are currently available.
- ◆ Modular structure.
- ◆ Architecturally* fault tolerant.
- ◆ Volumetric interconnect density capability (and thus an extremely high speed density factor).
- ◆ Flexible
- ◆ in choice of algorithm.
- ◆ in configurability.
- ◆ in connectivity.

Other positive attributes of the VANN are those normally associated with digital implementation and include

* additional fault tolerance may be inherent in the ANN algorithm.

CONFIDENTIAL

- ◆ Algorithm programmability
- ◆ Non volatility
- ◆ Ease in establishment of architecture fault tolerance
- ◆ No thermal drift in operating characteristics

The observations to this point strongly suggests a digital three dimensional ANN as the preferred architecture for adaptation and learning. The remainder of this white paper addresses more in detail how a VANN meets these objectives.

Volumetric Artificial Neural Network Description

Architecture

The VANN architecture is based conceptually on a cellular building block approach. The basic construction element is three-dimensional. Such a neural cell is most easily visualized as use a cube, but other arbitrary three-dimensional shapes (such as are found in crystal lattices) can also be used. A hexagonal cell, for example, is shown in Figure 1. Each cell contains a processing element such as a microcomputer and, in general, has the ability to simulate a number of neurons. A cell is directly connected electrically to each cell to which it is in physical contact. These connections carry information relating to the state of one or more neural cells, plus electrical power to permit the cells to function.

These cells may be stacked in volumetric fashion, *e.g.* the 8x5x4 cubic array as shown in Figure 2. Other arbitrary stackings may be obtained by simply ordering cubes differently. Nor is it necessary to have three stacking dimensions; an array could be laid out as a planar geometry, for example as simply 5x5x1, or as a linear array, for example 5x1x1. Neither do we require the same number of neurons in each layer. The resulting dimensions of the ANN is dictated only by the geometry of the basic construction element.

External Interface

Signals external to the array must be interfaced in such a manner as to permit large amounts of data throughput. The sides of the array and the open connections found on the sides may be so used. Both data input and output may be so facilitated. It is also possible to focus an image of data on one or more sides of the array by incorporating photodetectors and appropriate detection electronics into neurons on each such side. Alternatively, special cells may be affixed to each such side with photoreceptive properties, and little or no neural simulation ability.

Cell Connectivity

How high of a cell connectivity can be achieved? If every other layer in the cubic cellular structure was phased as illustrated in the top of Figure 3, then each cube makes physical contact with 12 adjacent cubes. Sides of 14 adjacent cubes can be made to have physical contact if adjacent rows in a layer are phased as is illustrated at the bottom of Figure 3. If similar phasing is applied to the hexagonal structure in Figure 1, then each unit will also make contact with 14 other units.

CONFIDENTIAL

Operation

Operation Modes

The VANN will operate in three modes: programming, learning and operational:

- (1) The type of ANN algorithm to be used is established in the programming mode. The operations here include establishment of the set of neurons to which a given neuron is (directly or indirectly) connected and the (sigmoidal) nonlinearity to be used by the neuron.
- (2) In the learning mode, the interconnect weights among neurons are established using training data or, in certain applications such as combinatorial search problems [7-8], some training algorithm. When training data are used, some or all of the neurons are assigned certain states. The interconnect weights are then determined internal to the VANN by algorithms both known and yet to be discovered. In certain training algorithms, the initial interconnect weights are algorithmically specified by, say, a random number generator.
- (3) In the operational mode, the neuron cubes perform three primary functions:
 - a) computation of the neuron state which is a function of the neurons to which it is connected,
 - b) conversion of the neuron's state into an electrical signal,
 - c) retransmission of neuron states from other adjacent neurons to yet other neurons in a message passing type of procedure.

Inter-Cell Communication

The interconnects from a neuron to the set of neurons with which it communicates are stored within the neural cell with the corresponding cell addresses. In the learning process, these values are established algorithmically (possibly iteratively) as a function of the states desired in the operational mode. This is done internally to the VANN, for example, by imposing desired states on a class of neural cells, letting the ANN compute the states at some other group of cells, and computing the difference of this value and the states desired. This error is then used to alter the interconnect weights to reduce or compensate for this error.

A neuron's state is typically computed as the (interconnect) weighted sum of connected neural states nonlinearly altered using some memoryless nonlinearity such as a sign function or a (biologically motivated) sigmoid. The conversion to an electrical signal of the state possibly involves scaling of the state value and generation of a destination address (each cell contains within it an address locator number which may be used to designate its position within the cell array) if required. Retransmission of adjacent state signals is done using a messenger function. They are employed to distribute state signals from a first cell which generates the signal to another cell (or a number of neurons) not adjacent to the first neuron.

The function of retransmission is employed to simulate the action of biological neurons which have a high degree of connectivity to numerous other neurons, some at a

CONFIDENTIAL

great distance from the source neuron. In any physical geometry of electronic neurons, this connectivity aspect represents a real problem. Allowing autoconnects, for example, in a 10x10x10 neuron array, it is possible to require up to one million interconnection paths in some algorithms. Wiring such a set of interconnections is clearly extremely difficult physically.

In the structure outlined here, all interconnects among non-adjacent neural cells are performed by having other neurons retransmit the sending state signal until the signal reaches its destination. Additionally, it is possible for a signal to be broadcast to a defined subset of all neurons, or even all neurons, via specially encoded messages. This is taken care of in the address portion of the signal.

Each cell must contain a communications handler whose purpose is to receive, redirect, and generate state signals. Each cell must also contain a computational element for computing state changes, and for applying weights to signals received from other neurons and also perhaps to weight its own outgoing signal. It must contain memory for program storage, which may be in the form of read-write, read-only, or read-mostly memory. It must contain read-write memory for storing parameters associated with changes in state and state weighting functions.

Neuron addresses may be either programmed permanently into each neuron prior to assembly of the array, or, preferably, would be self-programmed on power-up of the array. For example, a neural cell in the top left corner could through internal software ascertain its position simply via the fact that certain of its sides are not connected to other cells. It could then communicate to adjacent cells its position, allowing adjacent cells to determine their locations and hence addresses. The process can propagate automatically through the entire array until completed and all cells have assigned themselves addresses. The addresses would be stored in read-write memory or read-mostly memory in each neuron.

The flow of signals must be organized in such a fashion as to avoid collision of moving packets of information. For ANN algorithms that require each neuron to communicate with every other neuron, this can be achieved by alternating signal flow directions as is illustrated in Figure 4. At one instance, communication can be with neuron elements in a specified direction. In the next communication cycle, this direction would change. The technique can also be modified for the less severe case to algorithms where a neuron is only required to be connected to each neuron in an adjacent layer.

Downloading and Uploading Features of the VANN

Since cells imbedded deeply in the array are unreachable by direct electrical contact, the program may be 'downloaded' into each neuron via the retransmission process, or into just a subset of the array. A single neuron may be used as an entry node to facilitate the downloading. The programs may be loaded into the array via a conventional computer. Weights and communications paths may also be loaded into the array on a neuron by neuron basis if required by a similar process.

The ability to download neural information may be complemented by an 'upload' feature used to extract all neuron state and program information, especially information and programming of a variable nature. This is a critical feature for saving neural state information permanently onto hard media, such as a magnetic or optical disk. On power down of the network, all such information may be otherwise lost. Also, if a neural

CONFIDENTIAL

network is to be replicated in mass production with specific programming, such uploads are crucial to extracting the information required for duplication. Only then can the extracted information be reprogrammed into one or more other similar neural networks which, for example, may utilize a higher speed operational mode dedicated architecture or be fabricated using analog VLSI. If this process were not performed, it would be necessary to teach each network individually, a process which can be tedious and impractical. The upload/download techniques are a form of cloning akin to software duplication of a conventional computer's programs and information.

Neuron per Cell Ratio

Since each neuron contains a digital computing element, it is possible and indeed, desirable, for each neuron to simulate a number of neurons at once. The $8 \times 5 \times 4$ array shown may actually be made to simulate not 160 neurons but 640 neurons if each neuron cube simulates the action of four neurons. Communications among such 'internal' neurons may be facilitated with appropriate software. Communications among neurons would be quite similar except that additional burden would be placed on the inter-cell electrical connections.

Fault Tolerance

Another related issue is fault tolerance. If thousands of neurons are employed in a network, failures of neurons are inevitable. The software in each neuron must be designed to tolerate failures. For example, a communications failure of a single neuron may block transmission of messages among many other neurons. Considerable thought must be given to making communications automatically reroutable if such failures occur. It is possible to design a neuron algorithm such that an adjacent neuron could 'take over' the functioning of a bad cell or neuron.

Performance

The potential performance of a VANN is illustrated by the following analysis. We assume:

- ◆ A message handler can decode and route a byte or other parallel word of data and move it from one of the faces or edges of a neural cell to another face or edge to which it has physical contact at a constant rate, V bytes/second. Alternately, at this same rate, the handler can intercept a word and queue it to a neuron inside a neural cell.
- ◆ The VANN has linear dimension of N and thus is composed on the order of N^3 neural cells.
- ◆ A cell has K connection faces to adjacent cells.
- ◆ Each data packet travels an average distance of D cells from source to destination corresponding to D intercell transfers.

CONFIDENTIAL

From these assumptions, it follows that:

- ◆ At any given moment there can be a maximum of $K N^3$ bytes of information pending within the VANN communication interfaces.
- ◆ At an intercell transfer rate of V , there exists a $V K N^3$ bytes/second maximum transfer limit, and a limit of

$$T = V K N^3 / (L D)$$

on the number of packets/second transmitted and delivered where L is the communications packet length in bytes.

In order to better appreciate this analysis, let's assume we require $L = 72$ bits/packet (= 9 bytes/packet using 8 bit bytes) parsed as follows:

- 24 bits of destination address or specific destination code.
- 16 bits of data (neural state)
- 24 bits of source address
- 8 bits of special handling code information (multiple destinations, etc.)

Let's further assume that

- $N = 10$
- $V = 10^7$
- $K = 12$
- $L = 9$
- $D = N/2$ (average)

Then the effective transfer rate in terms of messages transmitted and received is:

$$T = 2.22 \times 10^9 \text{ per second (maximum)}$$

If we assume the reasonable inefficiency factor of 2 due to collisions, a realistic transfer rate would be

$$T \approx 10^9 \text{ messages/second delivered}$$

Assume further that each cell contains 1,000 artificial neurons. Then there would be a total of 10^6 neural simulations per second. This would only leave time for each neural simulation to be computed and retransmitted in only one microsecond. The neural computer imbedded in each cell would thus need to process 10^6 neural simulations per second, requiring perhaps an optimized DSP chip for the task or even several DSP chips running in tandem.

The problem then becomes inverted relative to more traditional ANN hardware: the communications, using conventional CPU hardware, becomes faster than the ability to compute.

In reality, data transfers can be made at least twice as fast as our example (50 nsec/byte) using relatively slow low power CMOS logic. With ECL logic, transfers can easily be made in about 10 nsec. As we have indicated, however, the transfer rates seem not to be the relevant issue with VANN's until processing speed can approach the sustainable transfer rates.

CONFIDENTIAL

Packaging

Electronic coupling via mechanical joined electrical contacts is highly unreliable and thus not suitable for use in avionics. There are at least three potentially attractive alternatives:

- ◆ Highly reliable capacitive coupling can be achieved using an appropriate thin layer of dielectric for the cell walls.
- ◆ If the physical dimensions of the array are fixed, interconnects can simply be hard wired.
- ◆ Communication among neural cells can be done optically. (Note that, however, unless power can be provided internal to the construction element or through some other externally applied field, alternate interconnects would still be required to provide power.) As is shown in Figure 5, optical sources, such as LED's, would be aligned to optical detectors at the construction element's surface through a skin of optically transparent material. Inter-element communication could be established by any one of a number of commonly used modulation techniques.

Power Dissipation

It may be seen that as each neuron cube consumes power, the power is converted to heat which must be dissipated in some manner. The geometry of the basic construction element can be modified to commit a large percentage of the volume to coolant flow. An example that can be used in lieu of the cube cell is shown in Figure 6. A single construction element is shown on top. A 2x2 array of these elements is shown on the bottom.

Final Remarks

The volumetric artificial neural network (VANN) is a neural network packaging with potentially high accurate performance capabilities using conventional electronics. We hope to propose a three year program wherein the VANN can be developed as a highly flexible and reliable computational tool for avionic and other applications. The milestones for this project are:

YEAR 1: Detailed performance of the VANN using state-of-the art electronics, including comparison with other more abstract connectionist architectures such as hypercubes and multicubes [9]. Initiate development of VANN software.

YEAR 2: Packaging study including materials, reliability analysis, cell coupling techniques and heat dissipation. Software finalization.

YEAR 3: Prototype the VANN.

CONFIDENTIAL

References

1. R. Hecht-Nielsen, "Neurocomputing: picking the brain", **IEEE Spectrum**, pp. 36-41 (1988).
2. J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities" **Proc. of the Nat'l. Academy of Sciences, USA**, Vol.(79), 2554-2558 (1982).
3. J.J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons" **Proc. of the Nat'l. Academy of Sciences, USA**, Vol.(81), 3088-3092.(1984).
4. R.J. Marks II and L.E. Atlas "Geometrical interpretation of Hopfield's content addressable memory neural network" **Northcon/88 Conference Record, vol.II**, pp.964-977, Seattle WA, October 1988 (Western Periodicals Co., North Hollywood, CA) - invited paper.
4. J.J. Hopfield, J.J. & D. Tank (1985). 'Neural' computation of decisions in optimization problems (Biol. Cybern.), Vol.(52), 141-152.
5. D.E. Rumelhart, J.L. McClelland and the PDP research group, **Parallel distributed processing: explorations in the microstructure of cognition**, (Bradford Books, Cambridge, MA.,1986)
6. D.E. Rumelhart, G.E. Hinton & R.J. Williams "Learning representations by back-propagating errors", **Nature**, Vol.(323), 533-536.(1986).
7. D.W. Tank & J.J. Hopfield "Simple 'neural' optimization networks: an A/D converter, signal decision circuit, and a linear programming circuit" **IEEE Trans on CAS**, Vol.(CAS-33), 533-541 (1986).
8. J.G. McDonnell, R.J. Marks II and L.E. Atlas "Neural networks for solving combinatorial search problems: a tutorial" **Northcon/88 Conference Record, vol.II**, pp.868-876, (Western Periodicals Co., North Hollywood, CA), Seattle WA, October 1988 - invited paper.
9. J.R. Goodman & P.J. Woest, "The Wisconsin multicube: a new large-scale cache-coherent multiprocessor", **Proceedings of the 15th Annual International Symposium on Computer Architecture**, May-June 1988, Honolulu (IEEE Computer Society Press), pp.422-431.

CONFIDENTIAL

FIGURE CAPTIONS:

FIGURE 1: GEOMETRICAL SHAPES SUCH AS THE HEXAGONAL ONE SHOWN HERE CAN BE USED AS A NEURAL CELL.

FIGURE 2: AN 8X5X4 ARRAY OF CUBIC NEURAL CELLS. POSSIBLE GEOMETRIES ARE DICTATED ONLY BY THE SHAPE OF THE NEURAL CELL.

FIGURE 3: (TOP) PHASING THE LAYERS OF A CUBIC NEURAL CELL ALLOWS EACH CELL TO INTERACT WITH THE 12 OTHER CELLS THAT IT TOUCHES. (BOTTOM) ADDITIONAL PHASING OF ADJACENT ROWS ALLOWS A CELL TO DIRECTLY CONNECT TO 14 OTHER CELLS.

FIGURE 4: ILLUSTRATION OF CYCLICALLY CHANGING SIGNAL FLOW DIRECTIONS. THE TECHNIQUE IS USED TO REDUCE COLLISIONS OF TRAVELING INFORMATION PACKETS. (ALL REQUIRED DIRECTION FLOWS FOR INTENSE INTERCONNECTION ARE NOT SHOWN.) ALTERNATELY, THE DIRECTION OF FLOW IN ADJACENT LAYERS CAN BE DIFFERENT AT DIFFERENT POINTS OF TIME.

FIGURE 5: ILLUSTRATION OF THE MANNER THAT ADJACENT CELLS CAN BE OPTICALLY COUPLED

FIGURE 6: (LEFT) AN EXAMPLE OF A CONSTRUCTION ELEMENT THAT ALLOWS AMPLE COOLANT FLOW. (RIGHT) A 2X2 ARRAY OF THESE ELEMENTS.

CONFIDENTIAL

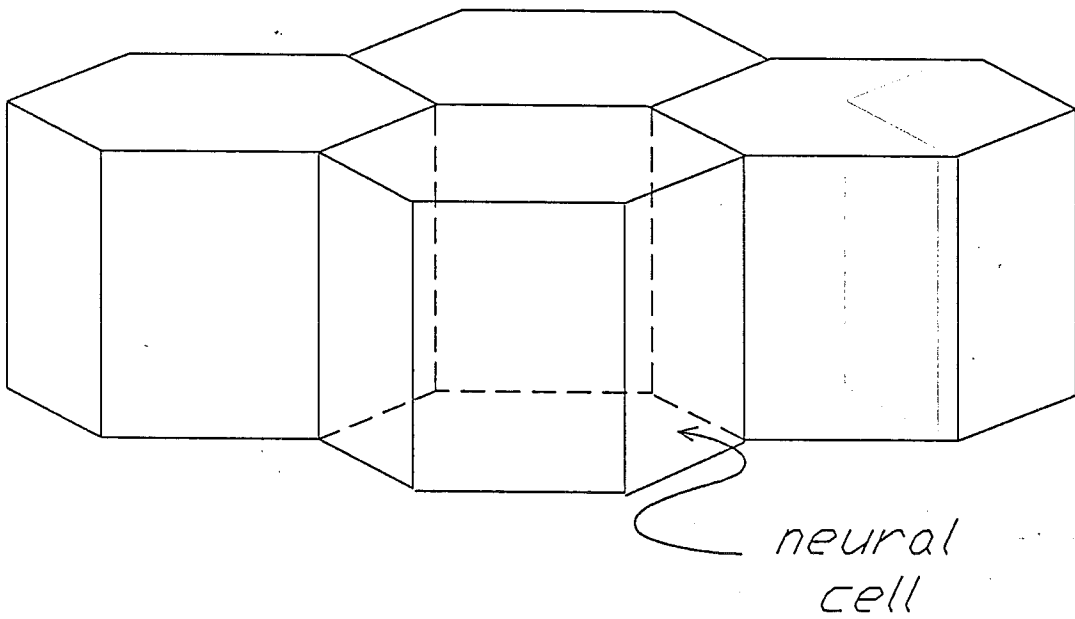


Figure 1

CONFIDENTIAL

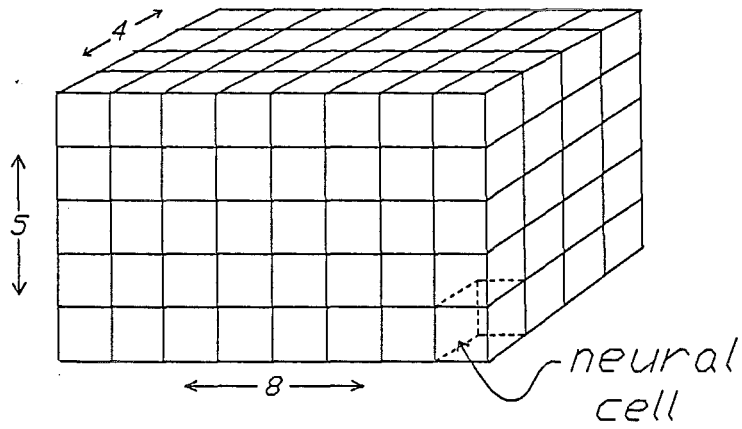


Figure 2

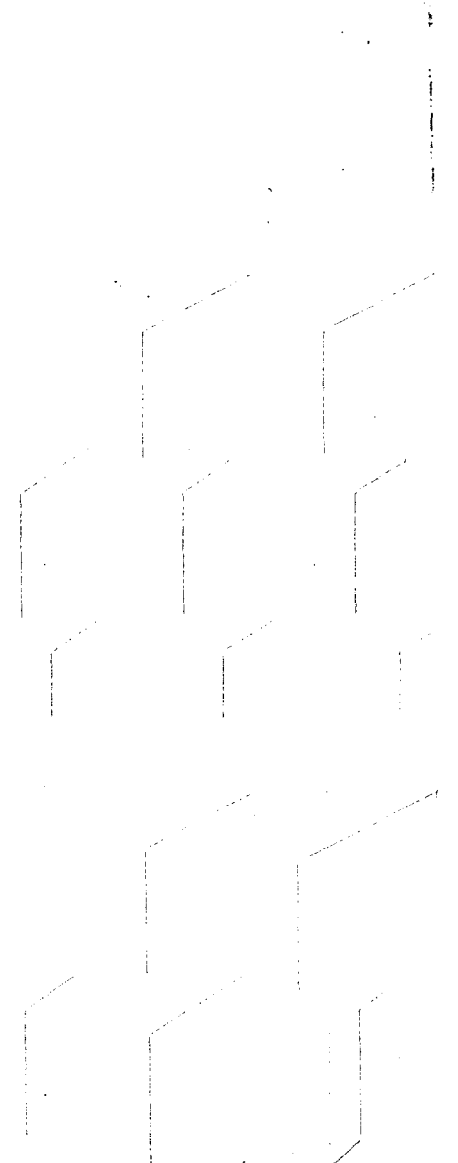
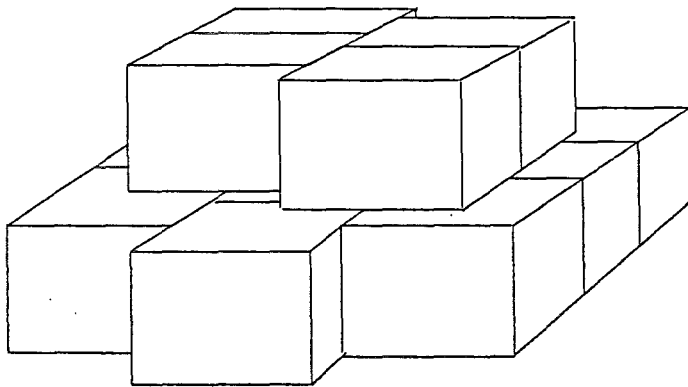
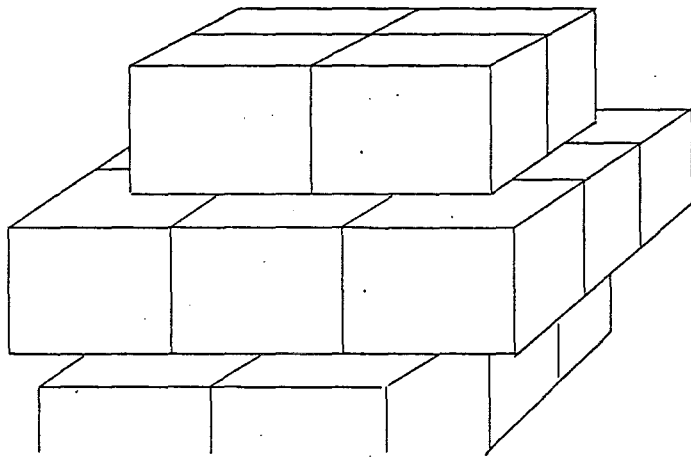


Figure 3

CONFIDENTIAL

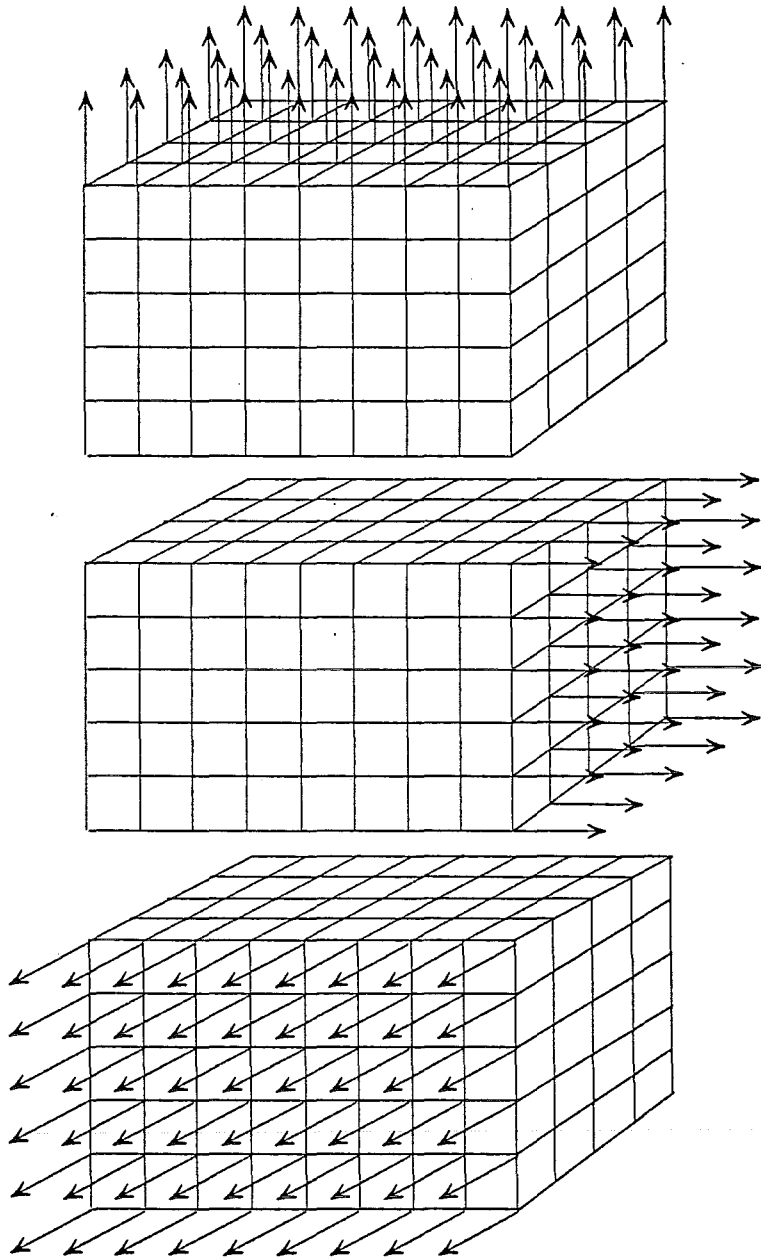


Figure 4

CONFIDENTIAL

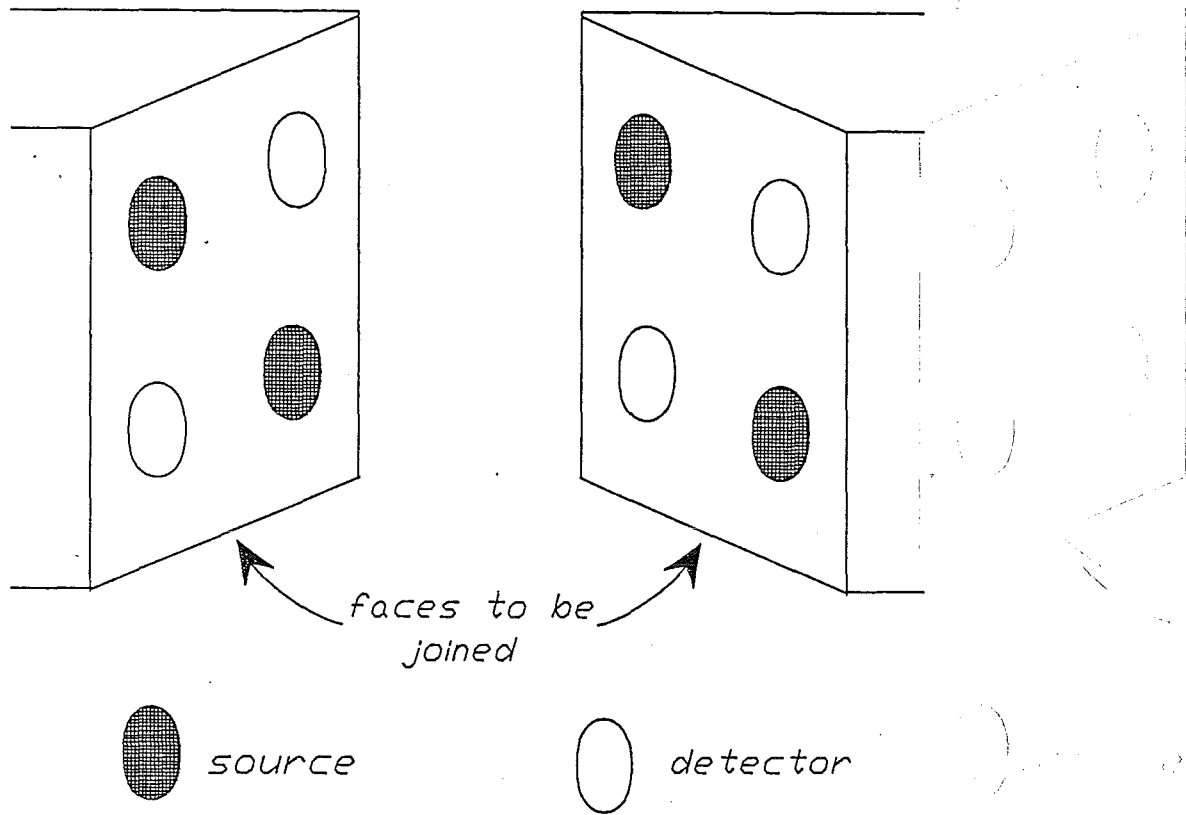


Figure 5

CONFIDENTIAL

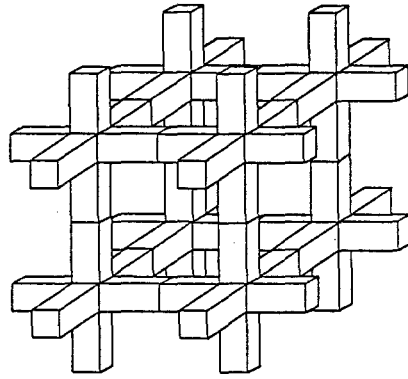
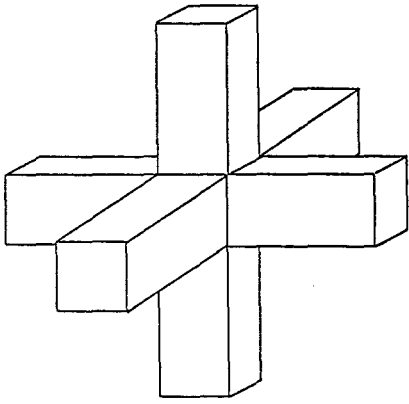


Figure 6

PATENT DESCRIPTION:
THREE DIMENSIONAL ARTIFICIAL NEURAL NETWORK ARRAY

(Confidential & Proprietary)
(contains 6 pages plus seven figures)
9-22-88

Harald Philipp
Robert J. Marks II

In this description, we present a new method of constructing electronic neural networks that permits modular three dimensional fabrication. Artificial neural networks (ANN's) attempt to simulate the construction and operation of their biological counterparts. While considerable effort has been made to create such electronics, most efforts to date have concentrated on using conventional high speed serial computers designed on a highly planar structure. This is in contrast to the parallel three dimensional structures found in many biological neural systems. As a result, a primary obstacle to manufacturing more complex electronic ANN's is the degree of interconnectivity required by a large number of neurons. This disclosure describes a method for overcoming these problems.

In this disclosure, a three dimensional ANN architecture is described which is based on a building block approach. The basic construction element is three-dimensional. For sake of discussion we will use a cube (Figure 1) but spheres, polyhedron, or other arbitrary three-dimensional shapes can also be used. A hexagonal neural construction unit is shown in Figure 2. As is illustrated in the cube example in Figure 1, each such construction element contains a processing element such as a microcomputer. This cube has at each of its edges or sides or both a series of electrical connections which are used to communicate with adjacent neurons. Such connectors carry information relating to the state of one or more neurons, plus electrical power to permit the neurons to function.

These cubes may be stacked in volumetric fashion, e.g. the $8 \times 5 \times 4$ cubic array as shown in Figure 3. Other arbitrary stackings may be obtained by simply ordering cubes differently. Nor is it necessary to have three stacking dimensions; an array could be laid out as a planar geometry, for example as simply $5 \times 5 \times 1$, or as a linear array, for example $5 \times 1 \times 1$. Neither do we require the same number of neurons in each layer. The resulting dimensions of the ANN is dictated only by the geometry of the basic construction element.

It may be seen that as each neuron cube consumes power, the power is converted to heat which must dissipated in some manner. The neuron cubes may be modified to permit air or coolant channels (Figure 1) when stacked. As shown, these channels would be designed to automatically couple when the units are connected. Alternately, the geometry of the basic construction element can be modified to commit a large percentage of the volume to coolant flow. An example that can be used in lieu of the cube is shown in Figure 4. A single construction element is shown of top. A 2×2 array of these elements is shown on the bottom.

A stack of neurons with springy interconnections must be somehow made to compress together to make good electrical contacts through-out. This can be accomplished with external pressure plates from all sides of the array (Figure 1). Dummy construction elements containing no electronics can be used to fill out the geometry to a rectangular box to allow for better pressurized mechanical coupling.

Another mechanical method of interconnecting such arrays is to have each cube snap together with adjacent cubes, obviating the need for external pressure plates. Cubes may

also be simply cemented together or adhered via any of a number of commercially available means, or through the attraction of magnets imbedded in each cube.

The ANN will operate in three modes: programming, learning and operational:

(1) The type of ANN architecture to be used is established in the programming mode. The operations here include establishment of the set of neurons to which a given neuron is (directly or indirectly) connected and the (sigmoidal) nonlinearity to be used by the neuron.

(2) In the learning mode, the interconnect weights among neurons are established using training data or, in certain applications such as combinatorial search problems, some training algorithm. When training data are used, some or all of the neurons are assigned certain states. The interconnect weights are then determined internal to the ANN by algorithms both known and yet to be discovered. In certain training algorithms, the initial interconnect weights are algorithmically specified by, say, a random number generator.

(3) In the operational mode, the neuron cubes perform three primary functions: a) computation of the neuron state which is a function of the neurons to which it is connected, b) conversion of the neuron's state into an electrical signal, c) retransmission of neuron states from other adjacent neurons to yet other neurons in a message passing type of procedure.

The interconnects from a neuron to the set of neurons with which it communicates are stored within the neuron cube with the corresponding cube addresses. In the learning process, these values are established algorithmically (possibly iteratively) as a function of the states desired in the operational mode. This is done internally to the ANN, for example, by imposing desired states on a class of neuron cubes, letting the ANN compute the states at some other group of neuron cubes, and computing the difference of this value and the states desired. This error is then used to alter the interconnect weights to reduce or compensate for this error.

A neuron state is typically computed as the (interconnect) weighted sum of connected neuron states nonlinearly altered using some memoryless nonlinearity such as a sign function or a (biologically motivated) sigmoid. The conversion to an electrical signal of the state possibly involves scaling of the state value and generation of a destination address (each neuron contains within it an address locator number which may be used to designate its position within the neuron array) if required. Retransmission of adjacent state signals is done using a messenger function. They are employed to distribute state signals from a first neuron which generates the signal to another neuron (or a plurality of neurons) not adjacent to the first neuron.

The function of retransmission is employed to simulate the action of biological neurons which have a high degree of connectivity to numerous other neurons, some at great distance from the source neuron. In any physical geometry of electronic neurons, this connectivity aspect represents a real problem. Allowing autococonnects, for example, in a $10 \times 10 \times 10$ neuron array, it is possible to require up to one million interconnection paths in some algorithms. Wiring such a set of interconnections is clearly extremely difficult physically.

In the structure outlined here, all interconnects among non-adjacent neurons are performed by having other neurons retransmit the sending state signal until the signal reaches its destination. Additionally, it is possible for a signal to be broadcast to a defined subset of all neurons, or even all neurons, via specially encoded messages. This is taken care of in the address portion of the signal. As a simple example, one neuron

may transmit a signal to one full layer of the array with a single transmission properly encoded with address information. Or, it could address all elements of the array at once.

In cases where a neuron typically communicates with a very large number of other neurons, the interconnects may also provide for a global communications path. Such a path would consist of an electrical interconnection common to all neurons (or perhaps a large subset of all neurons), which would facilitate the transmission of a signal from any one neuron so connected to all other neurons on the common connection, simultaneously. The design would require fault tolerance to any failure of a neuron on the interconnect which might 'hog' or clamp the global interconnect, rendering it useless. Such fault tolerance is characteristic with biological neural networks.

Algorithms for inter-neuron communication need to be designed to facilitate such relayed state information. Alternatively, each neuron could also contain a separate communications processor, perhaps hard wired in silicon (i.e. not implemented in software) for higher speed. The microcomputer would then be free to compute its new state from its existing state and new transmissions received from other neurons.

Each neuron must thus contain a communications handler whose purpose is to receive, redirect, and generate state signals. Each neuron must also contain a computational element for computing state changes, and for applying weights to signals received from other neurons and also perhaps to weight its own outgoing signal. It must contain memory for program storage, which may be in the form of read-write, read-only, or read-mostly memory. It must contain read-write memory for storing parameters associated with changes in state and state weighting functions.

Neuron addresses may be either programmed permanently into each neuron prior to assembly of the array, or, preferably, would be self-programmed on power-up of the array. For example, a neuron cube in the top left corner could through internal software ascertain its position simply via the fact that certain of its sides are not connected to other cubes. It could then communicate to adjacent cubes its position, allowing adjacent neurons to determine their locations and hence addresses. The process can propagate automatically through the entire array until completed and all neurons have assigned themselves addresses; the addresses would be stored in read-write memory or read-mostly memory in each neuron.

The interconnects may be simple mechanical contacts, perhaps spring loaded, which touch and make contact with adjacent neurons. If, for example, every other layer in the cube structure was phased as illustrated in the top of Figure 5, then each cube makes physical contact with 12 adjacent cubes. Sides of 14 adjacent cubes can be made to have physical contact if adjacent rows in a layer are phased as is illustrated at the bottom of Figure 5. If similar phasing is applied to the hexagonal structure in Figure 2, then each unit will also make contact with 14 other units.

Alternatively, communication among construction elements can be done optically thereby eliminating the need for transmitting signals through mechanically coupled interconnects. (Note that, however, unless power can be provided internal to the construction element or through some other externally applied field, mechanical interconnects would still be required to provide power.) As is shown in Figure 6, optical sources, such as LED's, would be aligned to optical detectors at the construction element's surface through a skin of optically transparent material. Inter-element communication could be established by any one of a number of commonly used modulation techniques.

The flow of signals must be organized in such a fashion as to avoid collision of moving packets of information. For artificial neural network algorithms that require each neuron to communicate with every other neuron, this can be achieved by alternating signal flow directions as is illustrated in Figure 7. At one instance, communication can be with neuron elements in a specified direction. In the next communication cycle, this direction would change. The technique can also be modified for the less severe case to algorithms where a neuron is only required to be connected to each neuron in an adjacent layer.

One primary characteristic of a neuron is its reprogrammability, in the sense that the other neurons it communicates with may be reprogrammed to be more or less restrictive. A neuron may "grow" communications paths to other neurons during a learn cycle, or similarly destroy such paths. It may also modify state weights on its own. Also, it may be desirable to modify the actual structure of the microcomputer program, either on its own through a learning process or through external intervention. For example, during development of a neural network computer the cubes may require program modification. A human programmer may then create a new microcomputer program and load this program into the array. Since neurons imbedded deeply in the array are unreachable by direct electrical contact, the program may be 'downloaded' into each neuron via the retransmission process, or into just a subset of the array. A single neuron may be used as an entry node to facilitate the downloading. The programs may be loaded into the array via a conventional computer. Weights and communications paths may also be loaded into the array on a neuron by neuron basis if required by a similar process.

The ability to download neural information may be complemented by an 'upload' feature used to extract all neuron state and program information, especially information and programming of a variable nature. This is a critical feature for saving neural state information permanently onto hard media, such as a magnetic or optical disk. On power down of the network, all such information may be otherwise lost. Also, if a neural network is to be replicated in mass production with specific programming, such uploads are crucial to extracting the information required for duplication. Only then can the extracted information be reprogrammed into one or more other similar neural networks which, for example, may utilize a higher speed operational mode dedicated architecture. If this process cannot be performed, it may be required to unnecessarily teach each network individually, a process which can be tedious and impractical. The upload/download techniques are a form of cloning akin to software duplication of a conventional computer's programs and information.

Another related issue is fault tolerance. If thousands of neurons are employed in a network, failures of neurons are inevitable. The software in each neuron must be designed to tolerate failures. For example, a communications failure of a single neuron may block transmission of messages among many other neurons. Considerable thought must be given to making communications automatically reroutable if such failures occur. It is possible to design a neuron algorithm such that an adjacent neuron could 'take over' the functioning of a bad neuron.

Since each neuron contains a digital computing element, it is possible for each neuron to simulate a number of neurons at once. The $8 \times 5 \times 4$ array shown may actually be made to simulate not 160 neurons but 640 neurons if each neuron cube simulates the action of four neurons. Communications among such 'internal' neurons may be facilitated with appropriate software. Communications among neurons would be quite similar except that additional burden would be placed on the inter-cube electrical connections.

Signals external to the array must be interfaced in such a manner as to permit large amounts of data throughput. The sides of the array and the open connections found on

the sides may be so used. Both data input and output may be so facilitated. It is also possible to focus an image of data on one or more sides of the array by incorporating photodetectors and appropriate detection electronics into neurons on each side. Alternatively, special cubes may be affixed to each side with photoreceptive properties, and little or no neural simulation ability. Energy fields other than light may also be used, such as microwave, sound, radiation, etc.

INVENTIVE ASPECTS

The inventive aspects of the proposed neural network we believe include but are not limited to the following:

1. A design for a neural network comprising a plurality of three dimensional structures or cells, each such cell having an ability to electrically or optically interconnect on a plurality of sides or edges of each such cell and each having an ability to simulate the characteristics of a neuron to varying degrees of modification in programming, learning and operational modes.

2. An ability to construct an arbitrary stacking of such cells into an array essentially without restriction or limit except for a requirement of physical contact with adjacent cells of similar type.

3. An ability of each cell within the array to electrically or optically communicate one or more of programs, data, or commands, the cells in general having an ability to originate, retransmit, receive and reconfigure as a function of such communications.

4. Several electro-mechanical means for interconnecting cells by stacking, involving one or more of: compression mated contacts, plug-together mechanisms, adhesive mating methods, or magnetic attraction.

5. A communications interconnection among cells which permits global or large-subset transmissions among cells, without requiring the retransmission function among cells.

6. An ability of each cell to perform computations on data received from other cells within the array or external to the array. A further ability of each cell to originate communications to one or more other similar cells, the communicated data or programming being dependent on an algorithm and on the nature of communications from other cells prior to the communication.

7. An ability for cells to self-determine their locations within an array by an algorithm and the communications means.

8. An ability for such an array and its component cells to propagate programs and data from an external source, either to all cells in an array or to a subset thereof.

9. An ability for such an array and its component cells to have programs and data extracted from it via an external computer or controller, either for storage, analysis, or duplication purposes.

10. The use of specially designed or programmed interface cells on one or more faces of the array, engineered to permit communications to and/or from external sources. The further use of light or other radiative means to couple either into or out of such cells in

order to simplify the task of connection, and the use of radiatively active transducers such as phototransistors and light emitting diodes to facilitate such external interface coupling.

11. An ability of functional cells to ignore malfunctioning cells via communications methods and algorithms governing the communications paths. A further ability of other cells to simulate the functions of malfunctioning cells if required.

12. An ability of a cell to simulate more than one neuron via computational algorithms, and to communicate information from such simulations to other cells in the array via similar communications means.

FIGURE 1: A SINGLE NEURON CUBE - EDGES AND FACES MAY BE USED FOR INTERCONNECTS. COOLING CHANNELS ARE CONSTRUCTED FOR MODULAR CONNECTION. SPRING INTERCONNECTS, SHOWN HERE, ARE ONE OF A NUMBER OF AVAILABLE TECHNIQUES FOR MECHANICAL COUPLING.

FIGURE 2: OTHER GEOMETRICAL SHAPES SUCH AS THE HEXAGONAL ONE SHOWN HERE CAN BE USED AS A NEURON ELEMENT.

FIGURE 3: AN $8 \times 5 \times 4$ ARRAY OF NEURON CUBES. POSSIBLE GEOMETRIES ARE DICTATED ONLY BY THE SHAPE OF THE NEURON UNIT.

FIGURE 4: (LEFT) AN EXAMPLE OF A CONSTRUCTION ELEMENT THAT ALLOWS AMPLE COOLANT FLOW. (RIGHT) A 2×2 ARRAY OF THESE ELEMENTS.

FIGURE 5: (TOP) PHASING THE LAYERS OF A CUBIC NEURON UNIT ALLOWS EACH NEURON UNIT TO INTERACT WITH THE 12 OTHER NEURON CUBES THAT IT TOUCHES. (BOTTOM) ADDITIONAL PHASING OF ADJACENT ROWS ALLOWS A CUBE TO DIRECTLY CONNECT TO 14 OTHER CUBES.

FIGURE 6: ILLUSTRATION OF THE MANNER THAT ADJACENT CONSTRUCTION ELEMENTS CAN BE OPTICALLY COUPLED

FIGURE 7: ILLUSTRATION OF CYCLICALLY CHANGING SIGNAL FLOW DIRECTIONS. THE TECHNIQUE IS USED TO REDUCE COLLISIONS OF TRAVELING INFORMATION PACKETS. (ALL REQUIRED DIRECTION FLOWS FOR INTENSE INTERCONNECTION ARE NOT SHOWN.) ALTERNATELY, THE DIRECTION OF FLOW IN ADJACENT LAYERS CAN BE DIFFERENT AT DIFFERENT POINTS OF TIME.

Figure 1

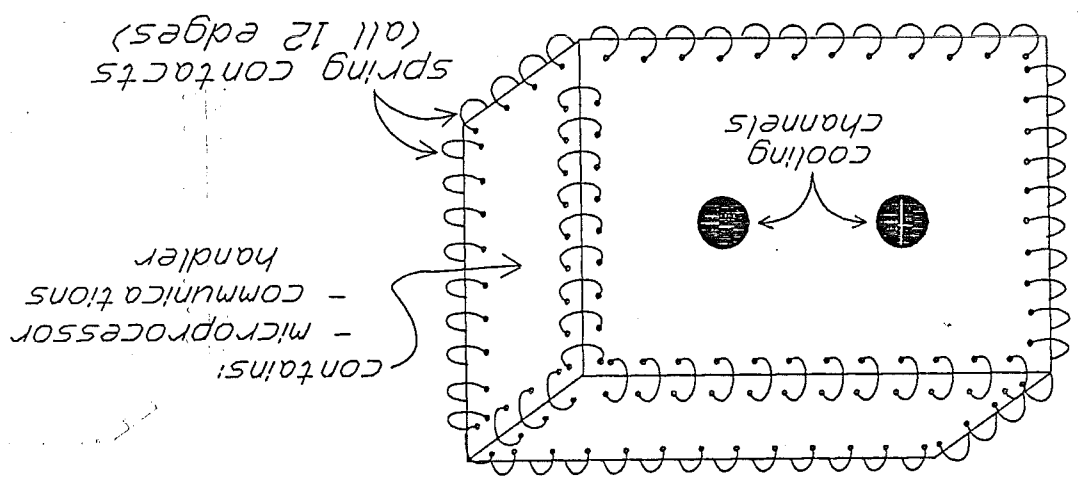


figure 2

neuron unit

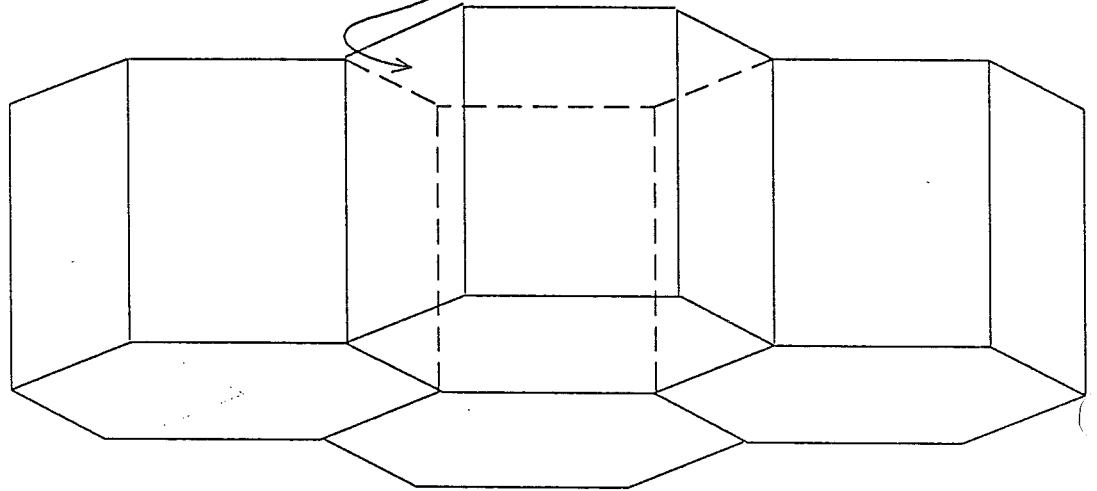


Figure 3

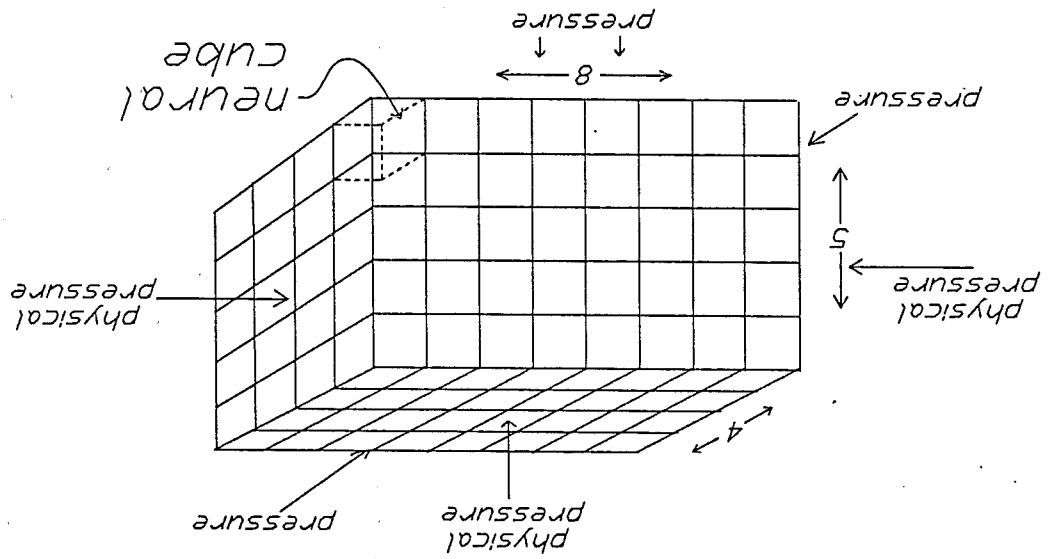
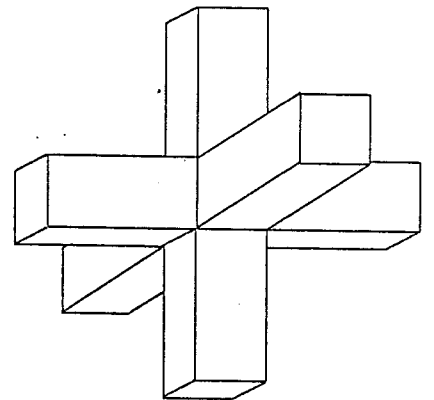
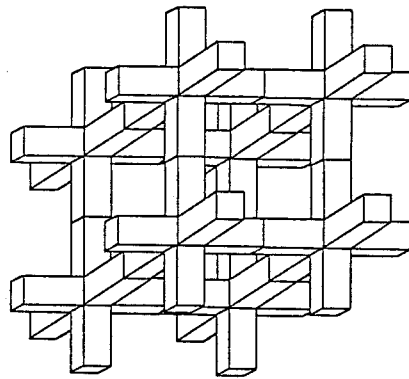


Figure 4



CONFIDENTIAL

Figure 5

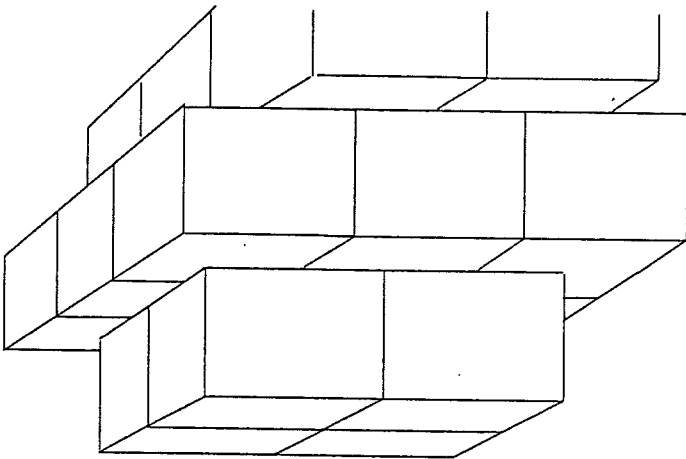
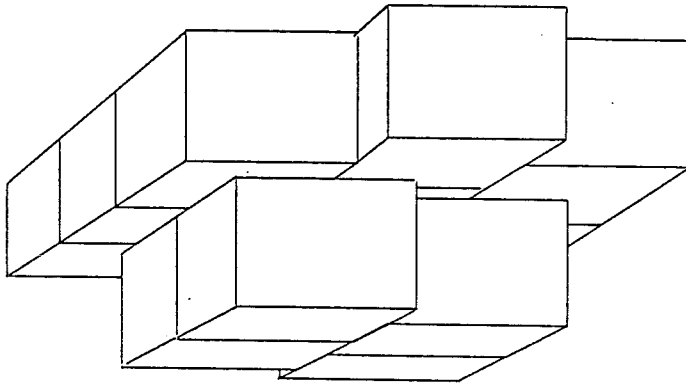


Figure 6

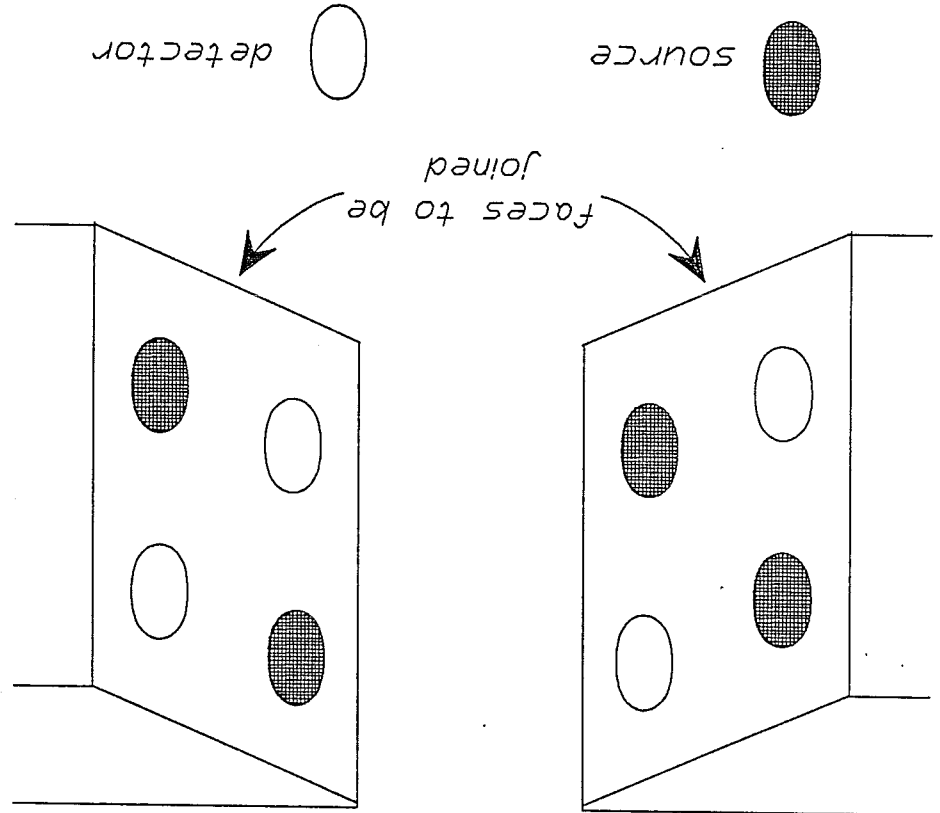
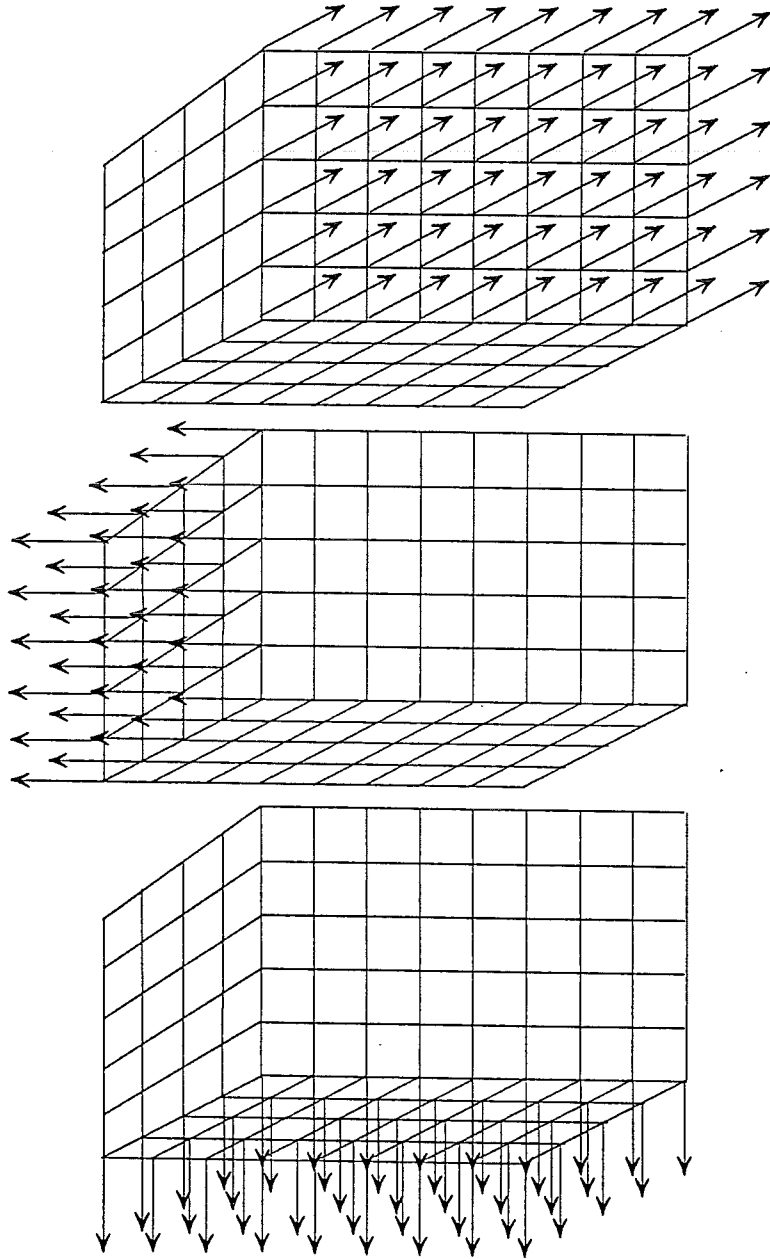


Figure 7



U.S. DEPARTMENT OF DEFENSE
SMALL BUSINESS INNOVATION RESEARCH (SBIR) PROGRAM
PHASE 1—FY 1989
PROJECT SUMMARY

Topic No. AF89-241

Military Department/Agency AFOSR

Name and Address of Proposing Small Business Firm
Multidimensional Systems Associates
13219 Northrup Way, Suite #203
Bellevue, WA 98003

Name and Title of Principal Investigator

Dr. Robert J. Marks II, Research Associate

Proposal Title

Volumetric Architectures

Technical Abstract (Limit your abstract to 200 words with no classified or proprietary information/data.)

We propose a massively parallel architecture for implementation of artificial neural networks. The result is a highly flexible, architecturally fault tolerant electronic system that can be implemented with currently available electronics. We propose, in Phase I, a feasibility study of the architecture and initial prototyping.

Anticipated Benefits/Potential Commercial Applications of the Research or Development
Interest in artificial neural networks (ANN's) has grown tremendously in this decade. The proposed architecture is potentially compatible with all of the ANN algorithms thus far proposed and may eventually evolve into a staple for ANN implementation.

At a maximum of 8 Key Words that describe the Project.
Artificial Neural Networks, Artificial Intelligence, VLSI, Connection Machines

C. IDENTIFICATION AND SIGNIFICANCE OF THE PROBLEM OR OPPORTUNITY

Introduction

Artificial neural networks (ANN's) attempt to primitively simulate the architecture and operation of their biological counterparts. While considerable effort has been made to create implementation electronics, most efforts to date have either [1]

- ◆ concentrated on using conventional high speed serial computers designed on a highly planar structure or
- ◆ have used high speed planar analog electronics.

The serial electronics have been primarily marketed as simulation tools. ANN's, however, are inherently parallel and serial implementation severely degrades potential speed possibilities. The analog electronics approach is superb for certain ANN operations (e.g. Hopfield ANN's [2-4] or recall from a trained ANN), but the poor accuracy of analog operations makes it ill suited for the high precision required of currently used adaptive and learning algorithms (e.g. back propagation [5-6]).

Furthermore, the planar approach to both serial digital and analog ANN VLSI architectures is in contrast to the parallel three dimensional structures found in many biological neural systems. The high connectivity available in three dimensions is clearly not available in two.

To overcome these limitations, we propose investigation and initial development of an electronic architecture for volumetric artificial neural networks (VANN's) with the following characteristics:

- ◆ Required electronics are currently available.
- ◆ Modular structure.
- ◆ Architecturally fault tolerant.
- ◆ Volumetric interconnect density capability (and thus an extremely high speed density factor).
- ◆ Flexible
- ◆ in choice of algorithm.
- ◆ in configurability.
- ◆ in connectivity.

* Other positive attributes of the VANN are those normally associated with digital implementation# and include

- ◆ Algorithm programmability
- ◆ Non volatility
- ◆ Ease in establishment of architecture fault tolerance
- ◆ No thermal drift in operating characteristics

* additional fault tolerance may be inherent in the ANN algorithm.
The architecture we propose for the VANN can also clearly be used with analog circuitry resulting in faster yet less accurate and less flexible processing ability.

The observations to this point strongly suggest a digital three dimensional ANN as the preferred architecture for adaptation and learning. The remainder of this proposal addresses more in detail how a VANN meets these objectives.

*** VANN Architecture**

The VANN architecture is based conceptually on a cellular building block approach. The basic construction element is three-dimensional. Such a neural cell is most easily visualized as a cube, but other arbitrary three-dimensional shapes (such as are found in crystal lattices) can also be used. A hexagonal cell, for example, is shown in Figure 1. Each cell contains a processing element such as a microcomputer and, in general, has the ability to simulate a number of neurons. A cell is directly connected electrically to each cell to which it is in physical contact. These connections carry information relating to the state of one or more neural cells, plus electrical power to permit the cells to function.

These cells may be stacked in volumetric fashion, e.g. the 8x5x4 cubic array as shown in Figure 2. Other arbitrary stackings may be obtained by simply ordering cubes differently. Nor is it necessary to have three stacking dimensions; an array could be laid out as a planar geometry, for example as simply 5x5x1, or as a linear array, for example 5x1x1. Neither do we require the same number of neurons in each layer. The resulting dimensions of the ANN is dictated only by the geometry of the basic construction element.

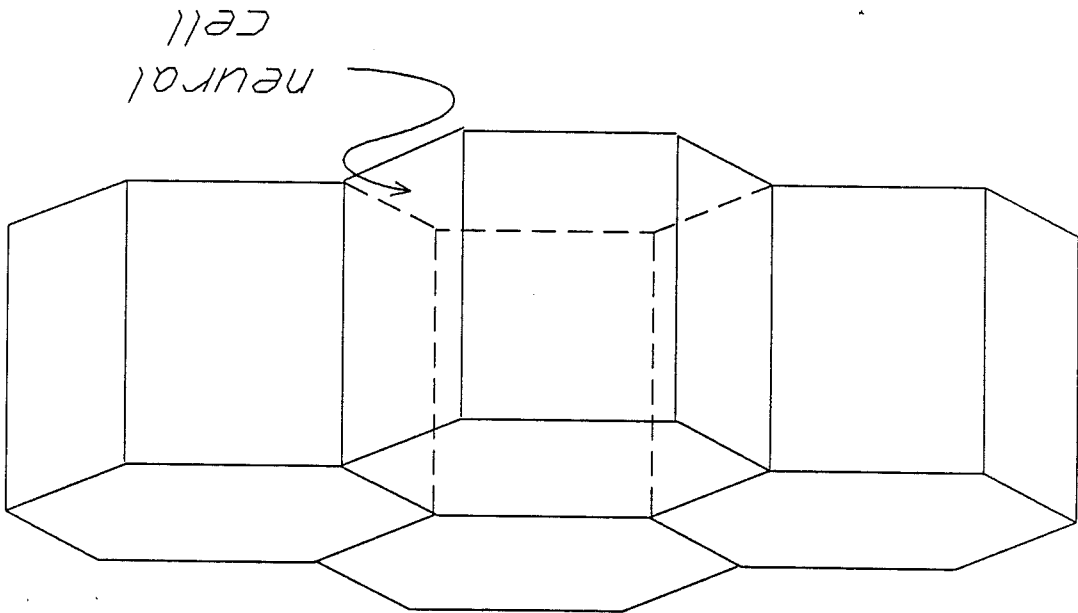
D. PHASE I TECHNICAL OBJECTIVES

The technical objective of Phase I will be to evaluate feasibility aspects of the VANN including the following:

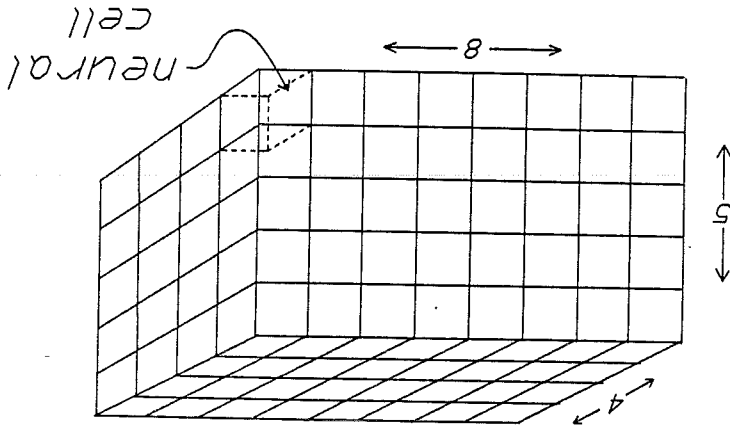
- * ♦ Operational (software) capabilities of the VANN including
 - * • programming, learning and recall modalities.
 - * • inter-cell communication abilities and limitations.
 - * • uploading and downloading capabilities.
 - * • programmable fault tolerance.
 - * ♦ Performance impact of architectural packaging options such as
 - * • cell shape and connectivity.
 - * • connection technology comparisons.
 - * • performance advantages over planar geometries.
 - * • cell shape and size effects on power dissipation.
 - * • external interfacing.
 - * ♦ Overall performance analysis evaluation
 - * • using currently available electronics.
 - * • in comparison with other current implementation technologies.

Following a more specific description of the VANN, the remainder of this section is devoted to a more in depth discussion of these proposed areas of research.

* FIGURE 1: GEOMETRICAL SHAPES SUCH AS THE HEXAGONAL ONE SHOWN HERE CAN BE USED AS A NEURAL CELL.



* FIGURE 2: AN 8X5X4 ARRAY OF CUBIC NEURAL CELLS. POSSIBLE GEOMETRIES ARE DICTATED ONLY BY THE SHAPE OF THE NEURAL CELL.



** Operation*

** Operation Modes*

The VANN will operate in three modes: programming, learning and recall:

(1) The type of ANN algorithm to be used is established in the programming mode. The operations here include establishment of the set of neurons to which a given neuron is (directly or indirectly) connected and the (sigmoidal) nonlinearity to be used by the neuron.

(2) In the learning mode, the interconnect weights among neurons are established using training data or, in certain applications such as combinatorial search problems [7-8], some training algorithm. When training data are used, some or all of the neurons are assigned certain states. The interconnect weights are then determined internal to the VANN by algorithms both known and yet to be discovered. In certain training algorithms, the initial interconnect weights are algorithmically specified by, say, a random number generator.

(3) In the recall mode, the neuron cubes perform three primary functions:

- * a) computation of the neuron state which is a function of the neurons to which it is connected,
- * b) conversion of the neuron's state into an electrical signal,
- * c) retransmission of neuron states from other adjacent neurons to yet other neurons in a message passing type of procedure.

** Inter-Cell Communication*

The interconnects from a neuron to the set of neurons with which it communicates are stored within the neural cell with the corresponding cell addresses. In the learning process, these values are established algorithmically (possibly iteratively) as a function of the states desired in the operational mode. This is done internally to the VANN, for example, by imposing desired states on a class of neural cells, letting the ANN compute the states at some other group of cells, and computing the difference of this value and the states desired. This error is then used to alter the interconnect weights to reduce or compensate for this error.

A neuron's state is typically computed as the (interconnect) weighted sum of connected neural states nonlinearly altered using some memoryless nonlinearity such as a sign function or a (biologically motivated) sigmoid. The conversion to an electrical signal of the state possibly involves scaling of the state value and generation of a destination address (each cell contains within it an address locator number which may be used to designate its position within the cell array) if required. Retransmission of adjacent state signals is done using a messenger function. They are employed to distribute state signals from a first cell which generates the signal to another cell (or a number of neurons) not adjacent to the first neuron.

The function of retransmission is employed to simulate the action of biological neurons which have a high degree of connectivity to numerous other neurons, some at a great distance from the source neuron. In any physical geometry of electronic neurons, this connectivity aspect represents a real problem. Allowing autoconnects, for

confidential proprietary information

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

* example, in a 10x10x10 neuron array, it is possible to require up to one million * interconnection paths in some algorithms. Wiring such a set of interconnections is * clearly extremely difficult physically.

* In the structure outlined here, all interconnects among non-adjacent neural cells * are performed by having other neurons retransmit the sending state signal until the * signal reaches its destination. Additionally, it is possible for a signal to be broadcast to * a defined subset of all neurons, or even all neurons, via specially encoded messages. * This is taken care of in the address portion of the signal.

* Each cell must contain a communications handler whose purpose is to receive, * redirect, and generate state signals. Each cell must also contain a computational * element for computing state changes, and for applying weights to signals received from * other neurons and also perhaps to weight its own outgoing signal. It must contain * memory for program storage, which may be in the form of read-write, read-only, or * read-mostly memory. It must contain read-write memory for storing parameters * associated with changes in state and state weighting functions.

* Neuron addresses may be either programmed permanently into each neuron prior * to assembly of the array, or, preferably, would be self-programmed on power-up of the * array. For example, a neural cell in the top front left corner could through internal * software ascertain its position simply via the fact that certain of its sides are not * connected to other cells. It could then communicate to adjacent cells its position, * allowing adjacent cells to determine their locations and hence addresses. The process * can propagate automatically through the entire array until completed and all cells have * assigned themselves addresses. The addresses would be stored in read-write memory or * read-mostly memory in each neuron.

* The flow of signals must be organized in such a fashion as to avoid or minimize * collision of moving packets of information. For ANN algorithms that require each * neuron to communicate with every other neuron, this can be achieved by alternating * signal flow directions as is illustrated in Figure 4. At one instance, communication can * be with neuron elements in a specified direction. In the next communication cycle, this * direction would change. The technique can also be modified for the less severe case to * algorithms where a neuron is only required to be connected to each neuron in an * adjacent layer.

Downloading and Uploading Features of the VANN

* Since cells imbedded deeply in the array are unreachable by direct electrical * contact, the program may be 'downloaded' into each neuron via the retransmission * process, or into just a subset of the array. A single neuron may be used as an entry * node to facilitate the downloading. The programs may be loaded into the array via a * conventional computer. Weights and communications paths may also be loaded into * the array on a neuron by neuron basis if required by a similar process.

* The ability to download neural information may be complemented by an 'upload' * feature used to extract all neuron state and program information, especially information * and programming of a variable nature. This is a critical feature for saving neural state * information permanently onto hard media, such as a magnetic or optical disk. On * power down of the network, all such information may be otherwise lost. Also, if a * neural network is to be replicated in mass production with specific programming, such * uploads are crucial to extracting the information required for duplication. Only then

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

* can the extracted information be reprogrammed into one or more other similar neural networks which, for example, may utilize a higher speed operational mode dedicated architecture or be fabricated using analog VLSI. If this process were not performed, it would be necessary to teach each network individually, a process which can be tedious and impractical. The upload/download techniques are a form of cloning akin to software duplication of a conventional computer's programs and information.

* Neuron per Cell Ratio

* Since each neuron contains a digital computing element, it is possible and indeed, desirable, for each neuron to simulate a number of neurons at once. The 8x5x4 array shown may actually be made to simulate not 160 neurons but 640 neurons if each neuron cube simulates the action of four neurons. Communications among such 'internal' neurons may be facilitated with appropriate software. Communications among neurons would be quite similar except that additional burden would be placed on the inter-cell electrical connections.

* Packaging Impact on Performance

* There are at least four potentially attractive techniques to couple neural cells:

- * ♦ Direct electrical contact, although relatively unreliable, is an obvious interconnect option.
- * ♦ Highly reliable capacitive coupling can be achieved using an appropriate thin layer of dielectric for the cell walls.
- * ♦ If the physical dimensions of the array are fixed, interconnects can simply be hard wired.

* ♦ Communication among neural cells can be done optically. (Note that, however, unless power can be provided internal to the construction element or through some other externally applied field, alternate interconnects would still be required to provide power.) As is shown in Figure 5, optical sources, such as LED's, would be aligned to optical detectors at the construction element's surface through a skin of optically transparent material. Inter-element communication could be established by any one of a number of commonly used modulation techniques.

* Power Dissipation

* It may be seen that as each neuron cube consumes power, the power is converted to heat which must be dissipated in some manner. The geometry of the basic construction element can be modified to commit a large percentage of the volume to coolant flow. An example that can be used in lieu of the cube cell is shown in Figure 6. A single construction element is shown on top. A 2x2 array of these elements is shown on the bottom.

* Fault Tolerance

confidential proprietary information

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

* Another related issue is fault tolerance. If thousands of neurons are employed in a network, failures of neurons are inevitable. The software in each neuron must be designed to tolerate failures. For example, a communications failure of a single neuron may block transmission of messages among many other neurons. Considerable thought must be given to making communications automatically reroutable if such failures occur. It is possible to design a neuron algorithm such that an adjacent neuron could 'take over' the functioning of a bad cell or neuron.

External Interface

* Signals external to the array must be interfaced in such a manner as to permit large amounts of data throughput. The sides of the array and the open connections found on the sides may be so used. Both data input and output may be so facilitated. It is also possible to focus an image of data on one or more sides of the array by incorporating photodetectors and appropriate detection electronics into neurons on each such side. Alternatively, special cells may be affixed to each such side with photoreceptive properties, and little or no neural simulation ability.

Cell Connectivity

* How high of a cell connectivity can be achieved? If every other layer in the cubic cellular structure was phased as illustrated in the top of Figure 3, then each cube makes physical contact with 12 adjacent cubes. Sides of 14 adjacent cubes can be made to have physical contact if adjacent rows in a layer are phased as is illustrated at the bottom of Figure 3. If similar phasing is applied to the hexagonal structure in Figure 1, then each unit will also make contact with 14 other units.

*** Overall Performance**

* Once reliable operation is assured and the architectural limitations of the VANN ascertained, the overall potential performance of the VANN can be evaluated. In this section, a sample calculation showing the potential performance of the VANN is given. We assume:

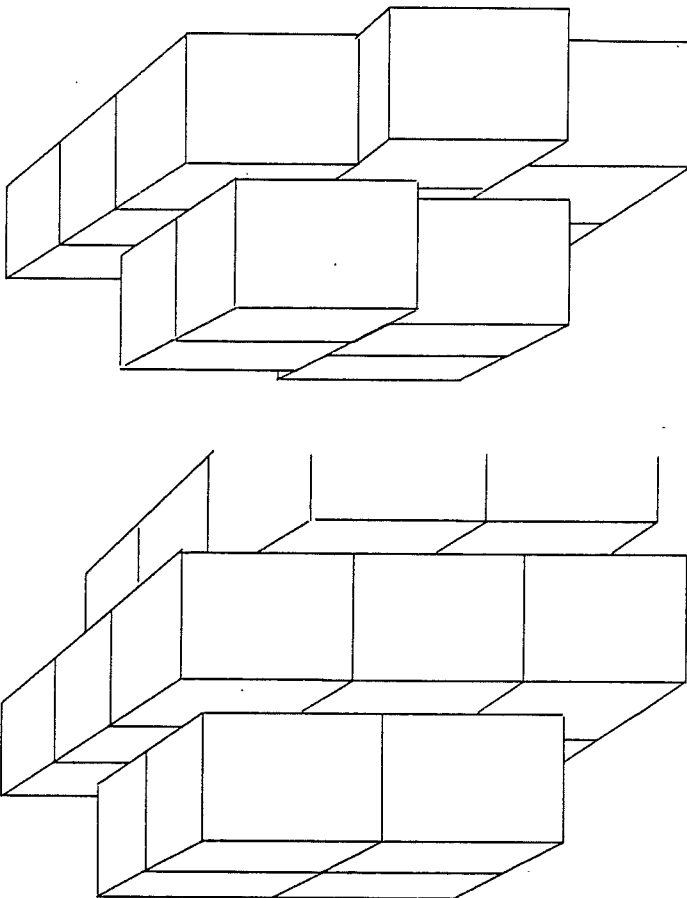
- * ♦ A message handler can decode and route a byte or other parallel word of data and move it from one of the faces or edges of a neural cell to another face or edge to which it has physical contact at a constant rate, V bytes/second. Alternately, at this same rate, the handler can intercept a word and queue it to a neuron inside a neural cell.
- * ♦ The VANN has linear dimension of N and thus is composed on the order of N^3 neural cells.
- * ♦ A cell has K connection faces to adjacent cells.

* ♦ Each data packet travels an average distance of D cells from source to destination corresponding to D intercell transfers.

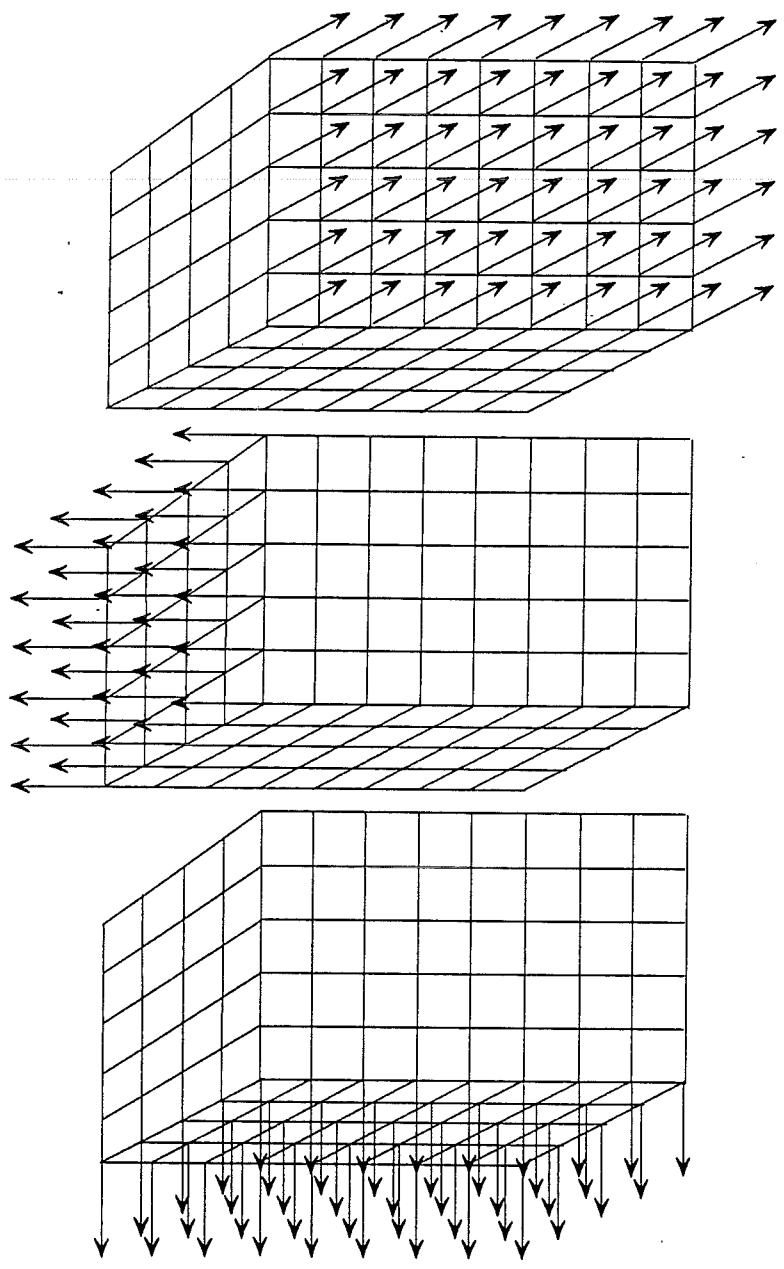
* From these assumptions, it follows that:

confidential proprietary information
Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the
restriction on the cover page of this proposal.

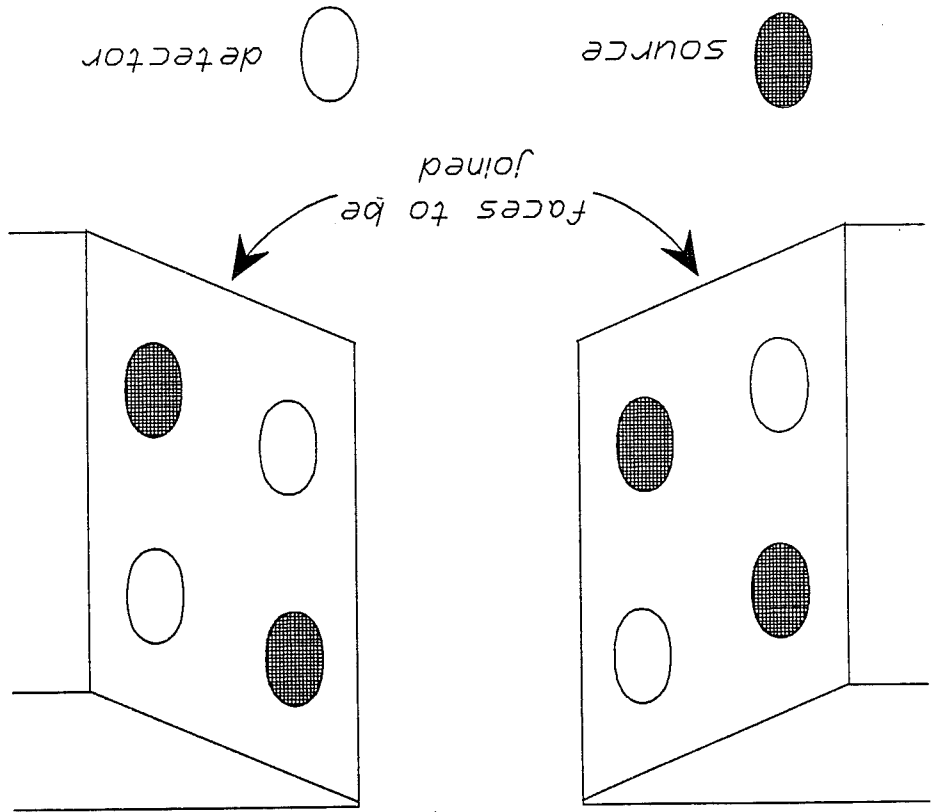
* FIGURE 3: (TOP) PHASING THE LAYERS OF A CUBIC NEURAL CELL ALLOWS EACH CELL TO INTERACT WITH
THE 12 OTHER CELLS THAT IT TOUCHES. (BOTTOM) ADDITIONAL PHASING OF ADJACENT ROWS ALLOWS A
CELL TO DIRECTLY CONNECT TO 14 OTHER CELLS.



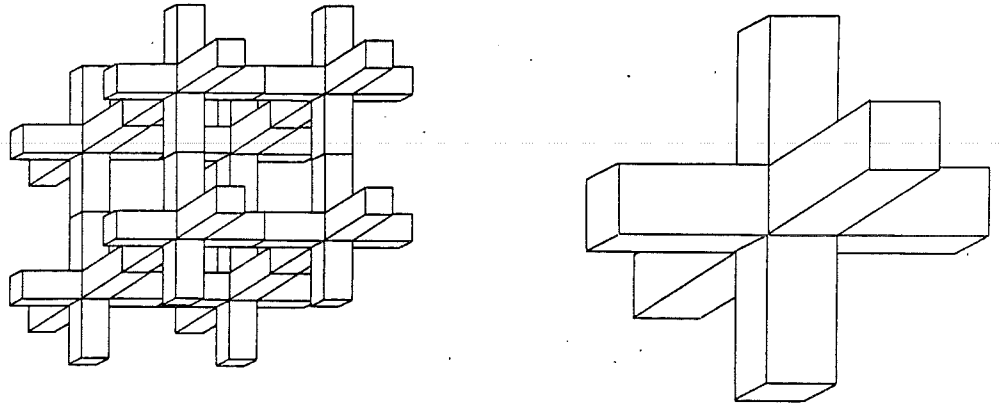
* FIGURE 4: ILLUSTRATION OF CYCLICALLY CHANGING SIGNAL FLOW DIRECTIONS. THE TECHNIQUE IS USED TO REDUCE COLLISIONS OF TRAVELING INFORMATION PACKETS. (ALL REQUIRED DIRECTION FLOWS FOR INTENSE INTERCONNECTION ARE NOT SHOWN.) ALTERNATELY, THE DIRECTION OF FLOW IN ADJACENT LAYERS CAN BE DIFFERENT AT DIFFERENT POINTS OF TIME.



* FIGURE 5: ILLUSTRATION OF THE MANNER THAT ADJACENT CELLS CAN BE OPTICALLY COUPLED



* FIGURE 6: (LEFT) AN EXAMPLE OF A CONSTRUCTION ELEMENT THAT ALLOWS AMPLE COOLANT FLOW. (RIGHT) A 2X2 ARRAY OF THESE ELEMENTS.



confidential proprietary information

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

- * ♦ At any given moment there can be a maximum of $K N^3$ bytes of information pending within the VANN communication interfaces.
- * ♦ At an intercell transfer rate of V , there exists a $V K N^3$ bytes/second maximum transfer limit, and a limit of

$$* T = V K N^3 / (L D)$$

* on the number of packets/second transmitted and delivered where L is the number of packets/second transmitted and delivered where L is the

* In order to better appreciate this analysis, let's assume we require $L = 72$ bits/packet (= 9 bytes/packet using 8 bit bytes) parsed as follows:

- * ♦ 24 bits of destination address or specific destination code.
- * ♦ 16 bits of data (neural state)
- * ♦ 24 bits of source address
- * ♦ 8 bits of special handling code information (multiple destinations, etc.)

* Let's further assume that

- * ♦ $N = 10$ cells per edge
- * ♦ $V = 10^7$ bytes per second (transfer rate)
- * ♦ $K = 12$ (the number of adjacent cells with which a cell is in physical contact)
- * ♦ $L = 9$ communications packet length
- * ♦ $D = N / 2$ (average transmission distance)

* Then the effective transfer rate in terms of messages transmitted and received is:

$$* T = 2.22 \times 10^9 \text{ per second (maximum)}$$

* If we assume the reasonable inefficiency factor of 2 due to collisions, a realistic transfer rate would be

$$* T \approx 10^9 \text{ messages/second delivered}$$

* Assume further that each cell contains 1,000 artificial neurons. Then there would be a total of 10^6 neural simulations per second. This would only leave time for each neural simulation to be computed and retransmitted in only one microsecond. The neural computer imbedded in each cell would thus need to process 10^6 neural simulations per second, requiring perhaps an optimized DSP chip for the task or even several DSP chips running in tandem.

* The problem then becomes inverted relative to more traditional ANN hardware: the communications, using conventional CPU hardware, becomes faster than the ability to compute.

* In reality, data transfers can be made at least twice as fast as our example (50 nsec/byte) using relatively slow low power CMOS logic. With ECL logic, transfers can easily be made in about 10 nsec. As we have indicated, however, the transfer rates seem not to be the relevant issue with VANN's until processing speed can approach the sustainable transfer rates.

* Through simulation, analysis and first order prototype, we hope to establish in Phase I a detailed performance analysis of the VANN using state-of-the-art electronics.

confidential proprietary information
 Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.
 * including a comparison with other more abstract connectionist architectures such as hypercubes and multicubes [9]

References

1. R. Hecht-Nielsen, "Neurocomputing: picking the brain", *IEEE Spectrum*, pp. 36-41 (1988).
2. J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities" *Proc. of the Nat'l. Academy of Sciences, USA*, Vol.(79), 2554-2558 (1982).
3. J.J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons" *Proc. of the Nat'l. Academy of Sciences, USA*, Vol.(81), 3088-3092.(1984).
4. R.J. Marks II and L.E. Atlas "Geometrical interpretation of Hopfield's content addressable memory neural network" *Northcon/88 Conference Record, vol.II*, pp.964-977, Seattle WA, October 1988 (Western Periodicals Co., North Hollywood, CA) - invited paper.
4. J.J. Hopfield, J.J. & D. Tank (1985). 'Neural' computation of decisions in optimization problems (Biol. Cybern.), Vol.(52), 141-152.
5. D.E. Rumelhart, J.L. McClelland and the PDP research group, *Parallel distributed processing: explorations in the microstructure of cognition*, (Bradford Books, Cambridge, MA.,1986)
6. D.E. Rumelhart, G.E. Hinton & R.J. Williams "Learning representations by back-propagating errors", *Nature*, Vol.(323), 533-536.(1986).
7. D.W. Tank & J.J. Hopfield "Simple 'neural' optimization networks: an A/D converter, signal decision circuit, and a linear programming circuit" *IEEE Trans on CAS*, Vol.(CAS-33), 533-541 (1986).
8. J.G. McDonnell, R.J. Marks II and L.E. Atlas "Neural networks for solving combinatorial search problems: a tutorial" *Northcon/88 Conference Record, vol.II*, pp.868-876, (Western Periodicals Co., North Hollywood, CA), Seattle WA, October 1988 - invited paper.
9. J.R. Goodman & P.J. Woest, "The Wisconsin multicube: a new large-scale cache-coherent multiprocessor", *Proceedings of the 15th Annual International Symposium on Computer Architecture*, May-June 1988, Honolulu (IEEE Computer Society Press), pp.422-431.

F. RELATED WORK

* The Principle Personnel have recently disclosed to the United States Patent Office * a fundamental description of the VANN (Disclosure Document #199167 dated 8/12/88) * and, in parallel with this proposal, are pursuing a patent filing for the VANN * architecture.

As is evident from the biographical information in sections i and k, the key personnel and consultants in this project have extensive experience in artificial neural networks, signal processing and consumer & industrial electronics. The reader is referred to these sections for specifics including patents, publications and related professional activities.

The Principle Investigator has participated extensively in University level funded research in artificial neural networks and related topics. He was one of two PI's on the industrial grant

- "Analysis and application of neural nets", Boeing High Technology Center (1986-88), with L.E. Atlas.

And, with the other PI's listed, is currently involved in the following projects:

- "Increasing the accuracy of optical processors", SDI/IST through ONR and the Optical Systems Lab at Texas Tech University (1986-89).

- "Neural network computer architectures", The Washington Technology Center (1987-89) with L.E. Atlas.

- "Power Systems Stability and Security Assessments Using Artificial Neural Networks" NSF (1988-1990) with M.A. El-Sharkawi, M. Damborg & L.E. Atlas.

The total cumulative funding of the above projects is well in excess of one half million dollars.

G. RELATIONSHIP WITH FUTURE RESEARCH OR R&D

* If the Phase I effort exposes the VANN as a highly reliable architecture with the * computational capabilities discussed in Section G, the VANN will be the most cost * effective, flexible ANN architecture with high accuracy that has been proposed to date. * Furthermore, the computational speed of the VANN will increase as the speed of state- * of-the-art electronics increases. In this sense, the VANN is a fundamental non-Von * Neumann architecture which is adaptive to advances in technology.

* Phase I of this project is concerned with detailed performance and feasibility * analysis of the VANN using state-of-the-art electronics, including comparison with * other more abstract connectionist architectures such as hypercubes and multicubes [9]. * First order prototyping to establish proof-of-principal is also proposed. Phase II * priorities will include:

- * Development of VANN software for the more commonly used ANN algorithms.
- * Performance of a packaging study including materials, reliability analysis, cell coupling and heat dissipation. The Principle Personnel have recently initiated dialog with *Stratos Products Development Group* of Seattle concerning the study of these packaging characteristics.
- * Development of a more sophisticated VANN prototype.
- * Investigation of potential use of the VANN architecture in other aspects of supercomputing.
- * Initiation of industrial dialog in anticipation of Phase III.

H. POTENTIAL POST APPLICATIONS

Rarely in the young history of electronic computer technology has there been an explosion of research interest as intense as that seen in this decade in regard to ANN's. In the last few years, the United States Federal Government, through DOD, NSF and other agencies has introduced multi-million dollar research initiatives in the field of ANN's. Other countries, including Japan, England and West Germany have recently established similar governmentally sponsored programs. Two major annual conferences on ANN's have been initiated in the last three years. They have, to date, continually drawn over 1500 international participants to each meeting. Despite this widespread interest in ANN's and the three dimensional character of many of nature's neural networks, the VANN, to our knowledge, is currently the only proposed architecture for a three dimensional ANN.

A fundamental reason for the enthusiasm for ANN's is their obvious potential to perform at least as well as their biological counterparts. Although ANN's are not new in concept, technology has advanced to the point where researchers believe they can be effectively fabricated. There have been significant demonstrations of ANN applications in areas such as vision, speech, signal processing, pattern recognition, robotics, combinatorics and even in financial fields such as mortgage brokering. Indeed, elaboration on the application of ANN's (and their efficient implementation as VANN's) is nearly as immense in scope as would be an elaboration on the applications of digital computers. Success in the currently massive sponsored research into ANN's will, frankly, propel the VANN into the status of a staple neural network computer architecture.

confidential proprietary information
Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the
restriction on the cover page of this proposal.

I. KEY PERSONNEL

ROBERT J. MARKS II, Principle Investigator

Robert J. Marks II received his PhD in 1977 from Texas Tech University in Lubbock. He is Chairman pro tem of the *IEEE Neural Networks Committee* which coordinates all of the neural network activity of *IEEE*. He is also Chair of the *IEEE Circuits & Systems Society Technical Committee on Neural Systems & Applications*. Joint with Les E. Atlas, he has twice taught the short course *Introduction to Artificial Neural Networks*, the most recent version (summer 1988) of which was video taped and is now available internationally through AMCEE. Professor Marks was awarded the *Outstanding Branch Councilor* award in 1982 by IEEE and, in 1984, was presented with an *IEEE Centennial Medal*. He is a senior member of IEEE. Dr. Marks was a co-founder and first President of the *Puget Sound Section of the Optical Society of America* and was recently elected that organization's first honorary member. Dr. Marks has over eighty archival journal and proceedings publications in the areas of detection theory, signal recovery, optical computing and artificial neural processing. He is a member of Eta Kappa Nu and Sigma XI.

PUBLICATIONS OF R.J. MARKS II IN ARTIFICIAL NEURAL NETWORKS AND RELATED AREAS:

ARCHIVAL PUBLICATIONS

- R.J. Marks II and L.E. Atlas "Composite matched filtering with error correction", *Optics Letters*, vol. 12, pp.135-137 (1987).
- R.J. Marks II "A class of continuous level associative memory neural nets", *Applied Optics*, vol. 26, pp.2005-2009 (1987).
- R.J. Marks II, J.A. Ritcey, L.E. Atlas and K.F. Cheung "Composite matched filter output partitioning", *Applied Optics*, vol. 26, pp.2274-2278 (1987).
- K.F. Cheung, L.E. Atlas, J.A. Ritcey, C.A. Green and R.J. Marks II "A comparison of conventional and composite matched filters with error correction", *Applied Optics*, vol. 26, pp.4235-4239 (1987).
- K.F. Cheung, L.E. Atlas and R.J. Marks II "Synchronous versus asynchronous behavior of Hopfield's content addressable memory", *Applied Optics*, vol. 26, pp.4808-4813 (1987).
- R.J. Marks II, L.E. Atlas, J.I. Choi, S. Oh, K.F. Cheung and D.C. Park "A performance analysis of associative memories with nonlinearities in the correlation domain", *Applied Optics*, vol. 27, pp.2900-2904 (1988).
- R.J. Marks II, L.E. Atlas and K.F. Cheung "Optical processor architectures for alternating projection neural networks", *Optics Letters*, vol. 13, pp.533-535 (1988).
- S. Oh, D.C. Park, R.J. Marks II and L.E. Atlas "Error detection and correction in multilevel algebraic optical processors", *Optical Engineering*, vol. 27, pp.289-294 (1988) - invited paper.

confidential proprietary information

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

R.J. Marks II, S. Oh, L.E. Atlas, and J.A. Ritcey "Alternating projection neural networks", to appear in *IEEE Trans. CAS*.

S. Oh, D.C. Park, R.J. Marks II and L.E. Atlas "Propagation skew in iterative optical & neural processors", *Optical Engineering* (in review) - invited paper.

PROCEEDINGS PAPERS

T. Homma, L.E. Atlas and R.J. Marks II "A neural network model for vowel classification", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1987. Published as a reprint in *Proceedings of the 1988 Connectionist Models Summer School*, (Morgan Kaufman Publishers, San Mateo, CA, 1988) pp.380-387.

J.A. Ritcey, L.E. Atlas, A. Somani, D. Nguyen, F. Holt and R.J. Marks II "A signal space interpretation of neural networks", *Proceedings of the International Symposium on Circuits and Systems*, pp.370-376, Philadelphia, May 1987.

K.F. Cheung, R.J. Marks II and L.E. Atlas "Neural net associative memories based on convex set projections", *Proceedings of the IEEE First International Conference on Neural Networks*, San Diego, June 1987.

R.J. Marks II, L.E. Atlas and K.F. Cheung "A class of continuous level neural nets", *Proceedings of the Fourteenth Congress of the International Commission for Optics*, pp.29-30, Quebec City, Quebec Canada, August 24-28, 1987.

R.J. Marks II, L.E. Atlas, S. Oh and J.A. Ritcey "The performance of convex set projection based neural networks", *Neural Information Processing Systems*, Dana Z. Anderson, editor, (American Institute of Physics, New York, 1988), pp. 534-543.

L.E. Atlas, T. Homma, and R.J. Marks II "An artificial neural network for spatio-temporal bipolar patterns: application to phoneme classification" *Neural Information Processing Systems*, Dana Z. Anderson, editor, (American Institute of Physics, New York, 1988) pp.31-40.

R.J. Marks II, L.E. Atlas and K.F. Cheung "Architectures for a continuous level neural network based on alternating orthogonal projections", *Proceedings of O-E/LASE '88 Conference on Neural Network Models for Optical Computing*, Los Angeles, January 1988, SPIE volume 882, pp 90-92.

R.J. Marks II, L.E. Atlas, J.J. Choi, S. Oh and D.C. Park "Nonlinearly requirements for correlation based associative memories", *Proceedings of O-E/LASE '88 Conference on Optical Computing and Nonlinear Materials*, Los Angeles, January 1988, SPIE volume 881, pp 179-183.

R.J. Marks II, L.E. Atlas and S. Oh, "Generalization in layered classification neural networks", *1988 IEEE International Symposium on Circuits and Systems*, pp. 503-506, Helsinki, 7-9 June, 1988.

R.J. Marks II, S. Oh, L.E. Atlas and J.A. Ritcey "Homogeneous and layered alternating projection neural networks", *Proceedings of the International Symposium on Optical Engineering and Industrial Sensing for Advanced Manufacturing Technologies*, 26-30 June 1988, Dearborn Hyatt, Michigan.

confidential proprietary information
Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

S. Oh, L.E. Atlas, R.J. Marks II and D.C. Park "Effects of clock skew in iterative neural network and optical feedback processors", **Proceedings of the IEEE International Conference on Neural Networks**, San Diego, July 24-27, 1988, vol.II, pp.429-436.

R.J. Marks II, L.E. Atlas, D.C. Park and S. Oh "The effect of stochastic interconnects in artificial neural network classification", **Proceedings of the IEEE International Conference on Neural Networks**, San Diego, July 24-27, 1988, vol.II, pp.437-442.

J.G. McDonnell, R.J. Marks II and L.E. Atlas "Neural networks for solving combinatorial search problems: a tutorial" **Northcon/88 Conference Record, vol.II**, pp.868-876, (Western Periodicals Co., North Hollywood, CA), Seattle WA, October 1988 - invited paper.

R.J. Marks II and L.E. Atlas "Geometrical interpretation of Hopfield's content addressable memory neural network" **Northcon/88 Conference Record, vol.II**, pp.964-977, Seattle WA, October 1988 (Western Periodicals Co., North Hollywood, CA) - invited paper.

R.J. Marks II, L.E. Atlas, J.J. Choi, S. Oh & D.C. Park "Parametric transformations for invariant image recognition", **Proc. of the International Congress on Optical Science & Engineering**, 24-28 April 1989, Paris - invited paper.

R.J. Marks II, S. Oh, D.C. Park and L.E. Atlas "Skew effects due to optical path length variation in iterative neural processors", **Proc. ISCAS 1989**, 9-11 May 1989, Portland - invited paper.

S. Oh and R.J. Marks II "Noise sensitivity of alternating projection neural networks", **Proc. ISCAS 1989**, 9-11 May 1989, Portland.

R.J. Marks II, M.J. Damborg and M.A. El-Sharkawi "Artificial neural networks for power system security assessment", **Proc. ISCAS 1989**, 9-11 May 1989, Portland - invited paper.

PUBLICATIONS - BOOK CHAPTERS

The invited paper [S. Oh, D.C. Park, R.J. Marks II and L.E. Atlas "Error detection and correction in multilevel algebraic optical processors" **Optical Engineering**, vol. 27, pp.289-294 (1988)] was selected for inclusion in the book **Optical Computing** edited by H. John Caulfield. The anthology is advertised as "a source book of outstanding optical engineering papers, selected from the world literature, on the subject of optical computing".

PUBLICATIONS - ABSTRACTS

C. Green, K.F. Cheung, L.E. Atlas and R.J. Marks II "Performance of conventional and composite matched filters with error correction", **Journal of the Optical Society of America A**, vol. 3, p.P13 (1986).

L.E. Atlas, J.A. Ritey, K.F. Cheung and R.J. Marks II "Improving the performance of composite matched filters", **Journal of the Optical Society of America A**, vol. 3, p.P13 (1986).

L.E. Atlas, R.J. Marks II and J.W. Taylor "Network learning modifications for multi-modal classification problems: applications to EKG patterns", *Neural Networks*, vol.1, sup. 1, p.4 (1988).

SPECIAL SESSIONS AND WORKSHOPS

Artificial Neural Systems and Applications, session organizer and co-chair, 1987 International Symposium on Circuits and Systems, Philadelphia (May 6, 1987).

Chair of Working Group on Perception at the Workshop on Optical Artificial Intelligence, Gold Lake, Colorado (3-5 August, 1987).

Artificial Neural Networks: Algorithms and Applications, organizer and chair, *Norhcon* '88, Seattle, WA, October 1988.

Artificial Neural Networks: Foundations and Implementations, organizer and chair, *Norhcon* '88, Seattle, WA, October 1988.

1989 International Joint Conference on Neural Networks Conference planning and publicity committees, Washington D.C..

Chair of the session on artificial neural networks at the International Symposium on Circuits and Systems to be held in May 1989 in Portland, Oregon.

Editor of a sequence of three articles on implementation of artificial neural networks (analog, digital and photonic) in *IEEE Circuits & Devices Magazine* (to appear in 1989).

PATENTS

R.J. Marks II, L.E. Atlas and S. Oh, "An optical neural network", assigned to the *Washington Technology Center* (pending).

Pieter J. van Heerden, Robert J. Marks II and Seho Oh, "A Computer Chip Realizing Learning in a Digital Computer", (pending).

HARALD PHILIPP, Research Associate

Harald Philipp received the B.S.E.E. degree (with honors) from Michigan Technological University in 1975. From 1975 to 1978, he was an engineer for the *National Bureau of Standards* in Maryland where he designed custom electronic instrumentation for fusion research applications and an electronic synchrotron, including EMP immune high speed timing, control, and data acquisition systems. From 1978 to 1981, he worked at Tektronix on the design of intelligent oscilloscope architectures. He was responsible for the development of the OF150 optical time-domain reflectometer for a number of firms in the Pacific Northwest in the fields of instrumentation, communications, microprocessor applications, and electro-optics. He also has experience in the development of hybrid integrated circuits, high speed state machines, surface mount electronics, local area networks, high speed laser diode pulsers and receivers, and real

confidential proprietary information
 Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

time imbedded software development. He is an honorary member of the Society of Photo-Optical Instrumentation Engineers.

Harald Philipp is currently Chairman of the Board of *Multidimensional Systems Associates*. He is also President of *Phillip Technologies Corp.* where he is involved in the developing and licensing of electronic technology in the fields of electrooptics, control systems, and data acquisition systems. Clients of *Phillip Technologies Corporation* include Tektronix Inc., John Fluke Manufacturing Co. Inc., Eaton Corporation, Photon Kinetics, Soloflex, Besam AB (Sweden) and numerous others.

Harald Philipp has 5 patents issued in the fields of database architectures, sampling systems, electro-optics, and switching amplifier technology:

- Patent #4,283,713
 - Patent #4,438,404
 - Patent #4,475,151
 - Patent #4,497,575
 - Patent #4,736,097
 - Patent Pending
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - Patent in Progress
 - * Patent in Progress
 - * Patent in Progress
 - * Patent in Progress
- * II)

Waveform Acquisition System

Signal Sampling System

Switching Amplifier Circuit

Optical Fiber Calibrator

Optical Motion Sensor

Energy Field Sensor

Morphological Signal Acquisition Conversion

Computed Successive Approximation Conversion

Logarithmic Successive Approximation Conversion

High Speed Digital Sampling Timebase

Waveform Synchronization Circuit

Volumetric Computing Structure (with R.J. Marks

RECENT PUBLICATIONS OF H. PHILIPP IN THE AREAS OF ELECTRONICS AND PHOTONICS:

H. Philipp and R. Syputa, "Fine tuning infrared sensor response" *Advanced Imaging*, (May/June 1988).

H. Philipp, "Light bridge photoelectric sensing", *Sensors* (August 1988).

H. Philipp, "The light bridge sensor", *Robotics World* (September 1988)

H. Philipp, "Photoelectric sensing", *Measurement and Control* (October 1988)

H. Philipp and R.J. Marks II "Microprocessor based light bridge sensors", *Proceedings of the International Symposium on Optical Engineering and Industrial Sensing for Advanced Manufacturing Technologies*, 26-30 June 1988, Dearborn Hyatt, Michigan - invited paper.

confidential proprietary information

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

I. PRIOR, CURRENT OR PENDING SUPPORT

There is no prior, current or pending support for a proposal similar to this one by *Multidimensional Systems Associates*. If an additional proposal with similar content is submitted to any federal agency by *Multidimensional Systems Associates* prior to a decision on this SBIR, the AFOSR office at Bolling AFB to which this proposal is submitted will be so notified immediately.

J. FACILITIES/EQUIPMENT

Multidimensional Systems Associates in affiliation with *Phillip Technologies Corporation* has the following equipment:

- Orion Unilab II microprocessor development system
- *Altera Sampius*TM state machine development system
- Gtek 7228 EPROM/microprocessor programmer
- Various electronic instrumentation (scopes, meters, etc.)

Various CAD, compiler and assembler software packages are also available.

* In order to initiate prototyping of multi-cell VANN's, we propose purchasing a *Xilinx* development system which will permit the on-sight fabrication of semi-custom IC's. The system requires the purchase of the following software and hardware:

- XC-DSS1 Programmable Gate Array Development System (\$4,950)
- XC-DSS81 Configuration PROM programmer (\$485)
- Extended memory (\$3500)

The first two items are available from *Xilinx Corporation*, 2069 Hamilton Ave., San Jose CA 95125.

* Simulation studies of the VANN will be performed on a subcontract basis to the *Interactive Systems Design Laboratory (ISDL)* at the *University of Washington*, Seattle, * under the direction of Prof. Les E. Atlas (see section K). The *ISDL*, founded in 1984, has been primarily concerned with artificial neural networks and speech processing. Several high quality A/D and D/A converters are used for signal acquisition onto either a Symbolics 3640 Lisp Machine, a Sun 3/50 computer or a SUN 4/110 system. The Symbolics computer has extensive signal processing software available including *ISDL* developed neural network simulators. The SUN computers have been programmed to simulate several auto-associative and trainable artificial neural network models. An additional simulation with interconnect filters instead of the standard interconnect scalar weights has been recently added to improve the convenience of these ANN simulation programs. Included in the lab is the following computational equipment:

- DEC Micro-VAX I
- Symbolics 3640 Lisp Machine
- SUN 3/50 Workstation
- SUN 4/100 Workstation
- Multi-Channel A/D and D/A converters
- Texas Instruments TMS32010 real-time DSP development system
- AT&T DSP32 floating-point real-time DSP development system.
- 2 SUN 386i workstations

confidential proprietary information

Use or disclosure of the proposed data on lines specifically identified by an asterisk (*) are subject to the restriction on the cover page of this proposal.

• 2 NEXT workstations

K. CONSULTANTS

LES E. ATLAS is an Associate Professor of Electrical Engineering at the University of Washington. He has received a *National Science Foundation Presidential Young Investigator Award* to support his work in artificial neural networks. As is evident upon inspection of the grant listings in section F and the publication list in section I, Prof. Atlas has worked quite closely with the Principle Investigator on a number of aspects of artificial neural networks. Other recent papers on this topic include:

L.E. Atlas "Auditory coding in higher centers of the CNS", *IEEE Engineering in Medicine and Biology Magazine*, vol.6, pp.29-32, June 1987 - invited paper.

Y. Suzuki and L.E. Atlas "A comparison of processor topologies for a fast trainable neural network for speech recognition", *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Glasgow, Scotland, May 23-26, 1989.

L.E. Atlas "Potential advantages of neural networks for automatic speech recognition" *Northcon/88 Conference Record, vol. II*, pp.877-881, (Western Periodicals Co., North Hollywood, CA), Seattle WA, October 1988 - invited paper.

DMITRY KAPLAN received his PhD in Electrical Engineering in 1988 from the University of Washington. His dissertation concerned applications of neural networks to artificial intelligence. Dr. Kaplan also has expertise in the areas of both digital & analog electronics and Fourier analysis. Since 1983, he has operated his own consulting firm with projects in the areas of adaptive image processing, two and three dimensional graphics and heuristic training. Dr. Kaplan's current interests center about application of neural networks to large artificial intelligence problems, neural architectures and training strategies.

Use or disclosure of the proposed data on lines specifically identified by an asterisk () are subject to the restriction on the cover page of this proposal.*

M. COST PROPOSAL

SALARIES

◆ PRINCIPLE INVESTIGATOR/PROJECT DIRECTOR:
Dr. Robert J. Marks II

◆ SENIOR PERSONNEL:

Robert J. Marks II, Principle Investigator
Harald Philipp, Research Associate

◆ OTHER PROFESSIONALS PERSONNEL:

Technician

◆ SECRETARIAL SUPPORT:

Secretary

TOTAL SALARIES

\$13108

EQUIPMENT

Xilinx Custom Chip maker~

\$ 8935

OTHER DIRECT COSTS

◆ CONSULTANTS:

Dr. Les E. Atlas
Dr. Dmitry Kaplan

1/2 person month \$ 3047
1 person month \$ 6094

◆ SUBCONTRACTS

The ISDL @ the University of Washington~

\$ 5400

TOTAL DIRECT COSTS

\$36584

INDIRECT COSTS

◆ OVERHEAD

35% of direct costs

\$12804

◆ WASHINGTON STATE BUSINESS AND OCCUPATION TAX
1.5% of direct costs

\$ 549

TOTAL DIRECT & INDIRECT COSTS

\$49937

~ see Section J for a description and purchase justification.
~ see Section J for this proposed subcontractor's role in the proposed research.

MULTIDIMENSIONAL SYSTEMS ASSOCIATES

13219 Northrup Way Suite #203
Bellevue, Washington 98003
Telephone: (206) 746-1642
FAX (206) 746-0566

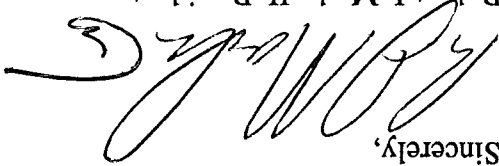
1/5/89

AFOSSR/XOT
SBIR Program Manager
Bldg 410, Rm A-113
Bolling AFB
Washington, D.C. 20332-5000
attn: Carmen Hernandez

Enclosed are five copies of a proposal Volumetric Architectures for Artificial Neural Networks that are in response to SBIR AF89-241: "Neurocomputers, New Architectures and Models of Computation".

We look forward to the reviews.

Sincerely,



Robert J. Marks II, President
Multidimensional Systems Associates

cc: H. Philipp
Enclosures

CONFIDENTIALITY AGREEMENT

PROGRAM/PROJECT:

te:

Whereas, The Washington Technology Center (WTC) is the owner of certain proprietary, confidential information relating to the above technology (TECHNOLOGY); and

Whereas, (COMPANY) wishes to receive the proprietary, confidential information to facilitate analysis and evaluation of the technology for commercial exploitation; and

Therefore, to assure WTC that all such proprietary information will be maintained by COMPANY under circumstances of strict confidentiality, COMPANY acknowledges and agrees as follows:

1. Proprietary information means any information relating directly or indirectly to the TECHNOLOGY not generally known to the public provided to COMPANY by WTC or its assignors/inventors. Proprietary information may be conveyed in written, graphic, aural or physical form and may include scientific knowledge, know-how, processes, inventions, techniques, formulas, products, business operations, customer requirements, data, plans or other records and information.

Proprietary information does not include information which COMPANY can demonstrate:

- (a) was in its knowledge or possession prior to disclosure by WTC or its assignors/inventors;
- (b) was public knowledge or has become public knowledge through no fault of COMPANY; or
- (c) was properly provided to COMPANY by an independent third party who has no obligation of secrecy to WTC or its assignors.

3. COMPANY agrees to maintain the disclosed proprietary information as confidential and agrees not to use this information for its own benefit or for the benefit of any other person or entity.

4. COMPANY may use the disclosed proprietary information only for the purposes of analyzing and evaluating the potential commercial uses of this information. The following restrictions apply:

- (a) COMPANY may duplicate or reproduce the disclosed proprietary information; if duplicated or reproduced in whole or in part, the disclosed information must carry a proprietary notice similar to that with which it was submitted to COMPANY.
- (b) COMPANY may not use, duplicate or disclose proprietary information for purposes of manufacture or procurement of the invention contained within the disclosed proprietary information.
- (c) COMPANY shall not use the disclosed proprietary information for research purposes nor to develop products or technologies for commercialization.

5. COMPANY agrees to protect WTC's proprietary information from further disclosure by taking equivalent precautions used to protect confidential information of COMPANY. In the event of unauthorized disclosure, COMPANY shall indemnify WTC for damages incurred as a result of the disclosure.

6. Upon completion of COMPANY's evaluation of proprietary information or at WTC's request, COMPANY will discontinue the use of and promptly return all proprietary information without retaining copies of that information and will promptly return samples or specimens embodying that information.

7. COMPANY agrees that violation of the Agreement will cause irreparable harm to WTC and that money damages will be inadequate to compensate WTC for its losses or damage. Therefore, COMPANY will stipulate to a motion for injunctive relief prohibiting violation or further violations of this Agreement should WTC desire such relief.

8. Any action arising out of this Agreement shall be decided in King County, Washington. This Agreement shall be construed under the laws of the State of Washington.

If COMPANY agrees to the foregoing, please indicate acceptance thereof by executing this Confidentiality Agreement.

Agreed to and Accepted this:

day of 19

Signature:

me:

Title:

Company:

CONFIDENTIALITY AGREEMENT

PROGRAM/PROJECT:

Date:

Whereas, The Washington Technology Center (WTC) is the owner of certain proprietary, confidential information relating to the above technology (TECHNOLOGY); and

Whereas, (COMPANY) wishes to receive the proprietary, confidential information to facilitate analysis and evaluation of the technology for commercial exploitation; and

Therefore, to assure WTC that all such proprietary information will be maintained by COMPANY under circumstances of strict confidentiality, COMPANY acknowledges and agrees as follows:

1. Proprietary information means any information relating directly or indirectly to the TECHNOLOGY not generally known to the public provided to COMPANY by WTC or its assignors/inventors. Proprietary information may be conveyed in written, graphic, aural or physical form and may include scientific knowledge, know-how, processes, inventions, techniques, formulae, products, business operations, customer requirements, data, plans or other records and information.

Proprietary information does not include information which COMPANY can demonstrate:

(a) was in its knowledge or possession prior to disclosure by WTC or its assignors/inventors;

(b) was public knowledge or has become public knowledge through no fault of COMPANY; or

(c) was properly provided to COMPANY by an independent third party who has no obligation of secrecy to WTC or its assignors.

3. COMPANY agrees to maintain the disclosed proprietary information as confidential and agrees not to use this information for its own benefit or for the benefit of any other person or entity.

4. COMPANY may use the disclosed proprietary information only for the purposes of analyzing and evaluating the potential commercial uses of this information. The following restrictions apply:

(a) COMPANY may duplicate or reproduce the disclosed proprietary information; if duplicated or reproduced in whole or in part, the disclosed information must carry a proprietary notice similar to that with which it was submitted to COMPANY.

(b) COMPANY may not use, duplicate or disclose proprietary information for purposes of manufacture or procurement of the invention contained within the disclosed proprietary information.

(c) COMPANY shall not use the disclosed proprietary information for research purposes nor to develop products or technologies for commercialization.

5. COMPANY agrees to protect WTC's proprietary information from further disclosure by taking equivalent precautions used to protect confidential information of COMPANY. In the event of unauthorized disclosure, COMPANY shall indemnify WTC for damages incurred as a result of the disclosure.

6. Upon completion of COMPANY's evaluation of proprietary information or at WTC's request, COMPANY will discontinue the use of and promptly return all proprietary information without retaining copies of that information and will promptly return samples or specimens embodying that information.

7. COMPANY agrees that violation of the Agreement will cause irreparable harm to WTC and that money damages will be inadequate to compensate WTC for its losses or damage. Therefore, COMPANY will stipulate to a motion for injunctive relief prohibiting violation or further violations of this Agreement should WTC desire such relief.

8. Any action arising out of this Agreement shall be decided in King County, Washington. This Agreement shall be construed under the laws of the State of Washington.

IF COMPANY agrees to the foregoing, please indicate acceptance thereof by executing this Confidentiality Agreement.

Agreed to and Accepted this:

_____ day of _____, 19____

Signature:

me:

Title:

Company:

CONFIDENTIALITY AGREEMENT

PROGRAM/PROJECT: _____

Date: _____

Whereas, The Washington Technology Center (WTC) is the owner of certain proprietary, confidential information relating to the above technology (TECHNOLOGY); and

Whereas, _____ (COMPANY) wishes to receive the proprietary, confidential information to facilitate analysis and evaluation of the technology for commercial exploitation; and

Therefore, to assure WTC that all such proprietary information will be maintained by COMPANY under circumstances of strict confidentiality, COMPANY acknowledges and agrees as follows:

1. Proprietary information means any information relating directly or indirectly to the TECHNOLOGY not generally known to the public provided to COMPANY by WTC or its assignors/inventors. Proprietary information may be conveyed in written, graphic, aural or physical form and may include scientific knowledge, know-how, processes, inventions, techniques, formulae, products, business operations, customer requirements, data, plans or other records and information.
2. Proprietary information does not include information which COMPANY can demonstrate:
 - (a) was in its knowledge or possession prior to disclosure by WTC or its assignors/inventors;
 - (b) was public knowledge or has become public knowledge through no fault of COMPANY; or
 - (c) was properly provided to COMPANY by an independent third party who has no obligation of secrecy to WTC or its assignors.
3. COMPANY agrees to maintain the disclosed proprietary information as confidential and agrees not to use this information for its own benefit or for the benefit of any other person or entity.
4. COMPANY may use the disclosed proprietary information only for the purposes of analyzing and evaluating the potential commercial uses of this information. The following restrictions apply:
 - (a) COMPANY may duplicate or reproduce the disclosed proprietary information; if duplicated or reproduced in whole or in part, the disclosed information must carry a proprietary notice similar to that with which it was submitted to COMPANY.
 - (b) COMPANY may not use, duplicate or disclose proprietary information for purposes of manufacture or procurement of the invention contained within the disclosed proprietary information.
 - (c) COMPANY shall not use the disclosed proprietary information for research purposes nor to develop products or technologies for commercialization.
5. COMPANY agrees to protect WTC's proprietary information from further disclosure by taking equivalent precautions used to protect confidential information of COMPANY. In the event of unauthorized disclosure, COMPANY shall indemnify WTC for damages incurred as a result of the disclosure.
6. Upon completion of COMPANY's evaluation of proprietary information or at WTC's request, COMPANY will discontinue the use of and promptly return all proprietary information without retaining copies of that information and will promptly return samples or specimens embodying that information.
7. COMPANY agrees that violation of the Agreement will cause irreparable harm to WTC and that money damages will be inadequate to compensate WTC for its losses or damage. Therefore, COMPANY will stipulate to a motion for injunctive relief prohibiting violation or further violations of this Agreement should WTC desire such relief.
8. Any action arising out of this Agreement shall be decided in King County, Washington. This Agreement shall be construed under the laws of the State of Washington.

If COMPANY agrees to the foregoing, please indicate acceptance thereof by executing this Confidentiality Agreement.

Agreed to and Accepted this:

_____ day of _____, 19_____

Signature: _____

Name: _____

Title: _____

Company: _____

the sides may be so used. Both data input and output may be so facilitated. It is also possible to focus an image of data on one or more sides of the array by incorporating photodetectors and appropriate detection electronics into neurons on each such side. Alternatively, special cubes may be affixed to each such side with photoreceptive properties, and little or no neural simulation ability. Energy fields other than light may also be used, such as microwave, sound, radiation, etc.

INVENTIVE ASPECTS

The inventive aspects of the proposed neural network we believe include but are not limited to the following:

1. A design for a neural network comprising a plurality of three dimensional structures or cells, each such cell having an ability to electrically or optically interconnect on a plurality of sides or edges of each such cell and each having an ability to simulate the characteristics of a neuron to varying degrees of modification in programming, learning and operational modes.
2. An ability to construct an arbitrary stacking of such cells into an array essentially without restriction or limit except for a requirement of physical contact with adjacent cells of similar type.
3. An ability of each cell within the array to electrically or optically communicate one or more of programs, data, or commands, the cells in general having an ability to originate, retransmit, receive and reconfigure as a function of such communications.
4. Several electro-mechanical means for interconnecting cells by stacking, involving one or more of: compression mated contacts, plug-together mechanisms, adhesive mating methods, or magnetic attraction.
5. A communications interconnection among cells which permits global or large-subset transmissions among cells, without requiring the retransmission function among cells.
6. An ability of each cell to perform computations on data received from other cells within the array or external to the array. A further ability of each cell to originate communications to one or more other similar cells, the communicated data or programming being dependent on an algorithm and on the nature of communications from other cells prior to the communication.
7. An ability for cells to self-determine their locations within an array by an algorithm and the communications means.
8. An ability for such an array and its component cells to propagate programs and data from an external source, either to all cells in an array or to a subset thereof.
9. An ability for such an array and its component cells to have programs and data extracted from it via an external computer or controller, either for storage, analysis, or duplication purposes.
10. The use of specially designed or programmed interface cells on one or more faces of the array, engineered to permit communications to and/or from external sources. The further use of light or other radiative means to couple either into or out of such cells in

order to simplify the task of connection, and the use of radiatively active transducers such as phototransistors and light emitting diodes to facilitate such external interface coupling.

11. An ability of functional cells to ignore malfunctioning cells via communications methods and algorithms governing the communications paths. A further ability of other cells to simulate the functions of malfunctioning cells if required.

12. An ability of a cell to simulate more than one neuron via computational algorithms, and to communicate information from such simulations to other cells in the array via similar communications means.

FIGURE 1: A SINGLE NEURON CUBE - EDGES AND FACES MAY BE USED FOR INTERCONNECTS. COOLING CHANNELS ARE CONSTRUCTED FOR MODULAR CONNECTION. SPRINGY INTERCONNECTS, SHOWN HERE, ARE ONE OF A NUMBER OF AVAILABLE TECHNIQUES FOR MECHANICAL COUPLING.

FIGURE 2: OTHER GEOMETRICAL SHAPES SUCH AS THE HEXAGONAL ONE SHOWN HERE CAN BE USED AS A NEURAL ELEMENT.

FIGURE 3: AN 8X5X4 ARRAY OF NEURAL CUBES. POSSIBLE GEOMETRIES ARE DICTATED ONLY BY THE SHAPE OF THE NEURAL UNIT.

FIGURE 4: (LEFT) AN EXAMPLE OF A CONSTRUCTION ELEMENT THAT ALLOWS AMPLE COOLANT FLOW. (RIGHT) A 2X2 ARRAY OF THESE ELEMENTS.

FIGURE 5: (TOP) PHASING THE LAYERS OF A CUBIC NEURON UNIT ALLOWS EACH NEURAL UNIT TO INTERACT WITH THE 12 OTHER NEURAL CUBES THAT IT TOUCHES. (BOTTOM) ADDITIONAL PHASING OF ADJACENT ROWS ALLOWS A CUBE TO DIRECTLY CONNECT TO 14 OTHER CUBES.

FIGURE 6: ILLUSTRATION OF THE MANNER THAT ADJACENT CONSTRUCTION ELEMENTS CAN BE OPTICALLY COUPLED

FIGURE 7: ILLUSTRATION OF CYCLICALLY CHANGING SIGNAL FLOW DIRECTIONS. THE TECHNIQUE IS USED TO REDUCE COLLISIONS OF TRAVELING INFORMATION PACKETS. (ALL REQUIRED DIRECTION FLOWS FOR INTENSE INTERCONNECTION ARE NOT SHOWN.) ALTERNATELY, THE DIRECTION OF FLOW IN ADJACENT LAYERS CAN BE DIFFERENT AT DIFFERENT POINTS OF TIME.

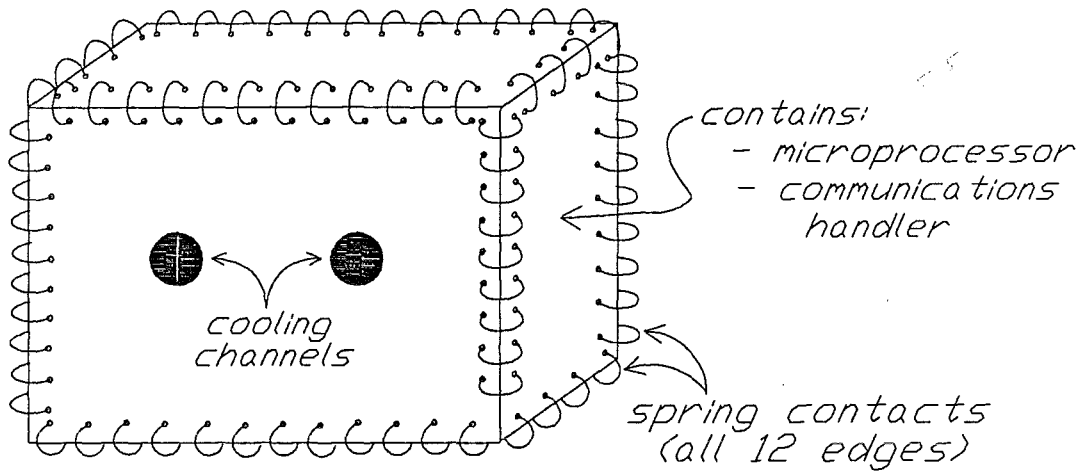


figure 1

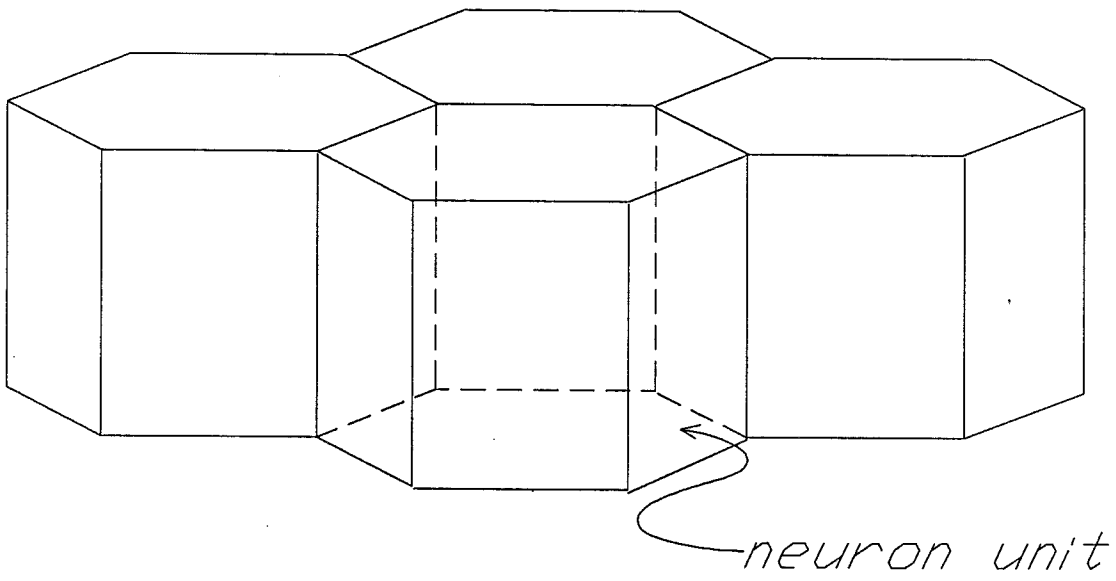


figure 2

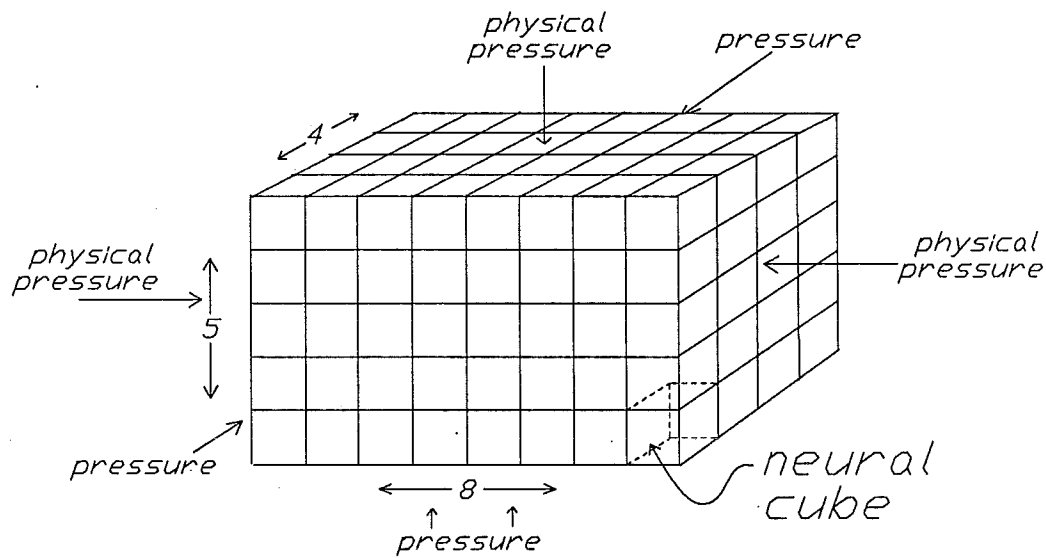


figure 3

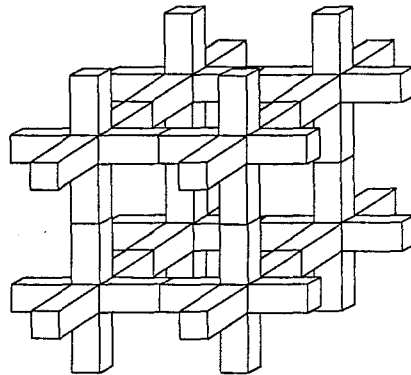
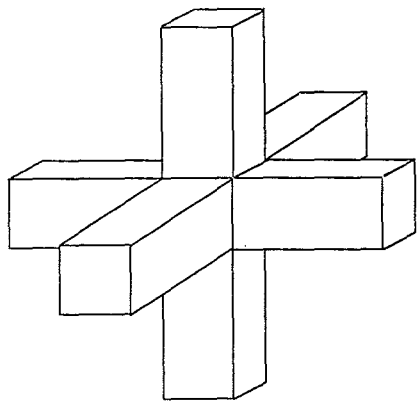


figure 4

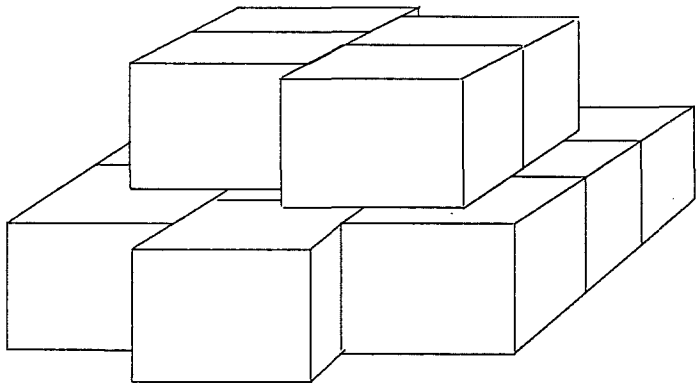
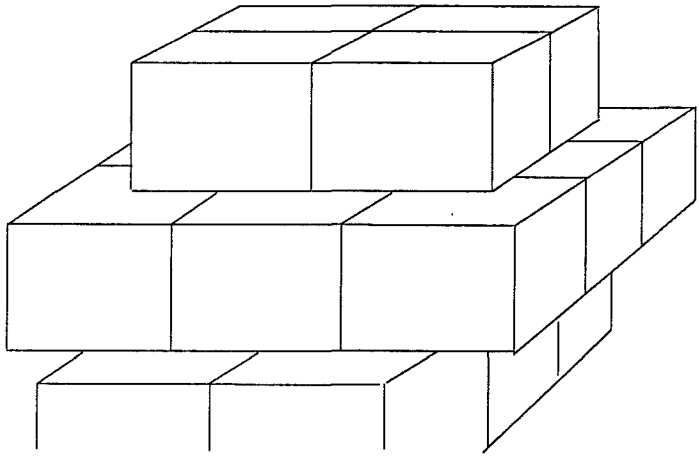


figure 5

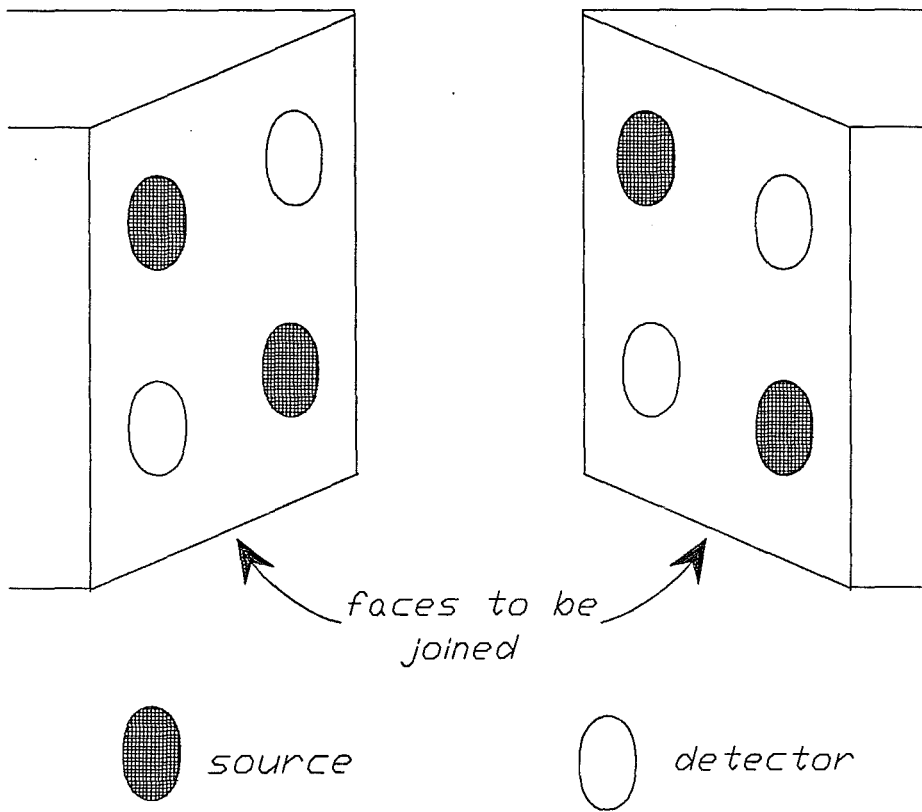


figure 6

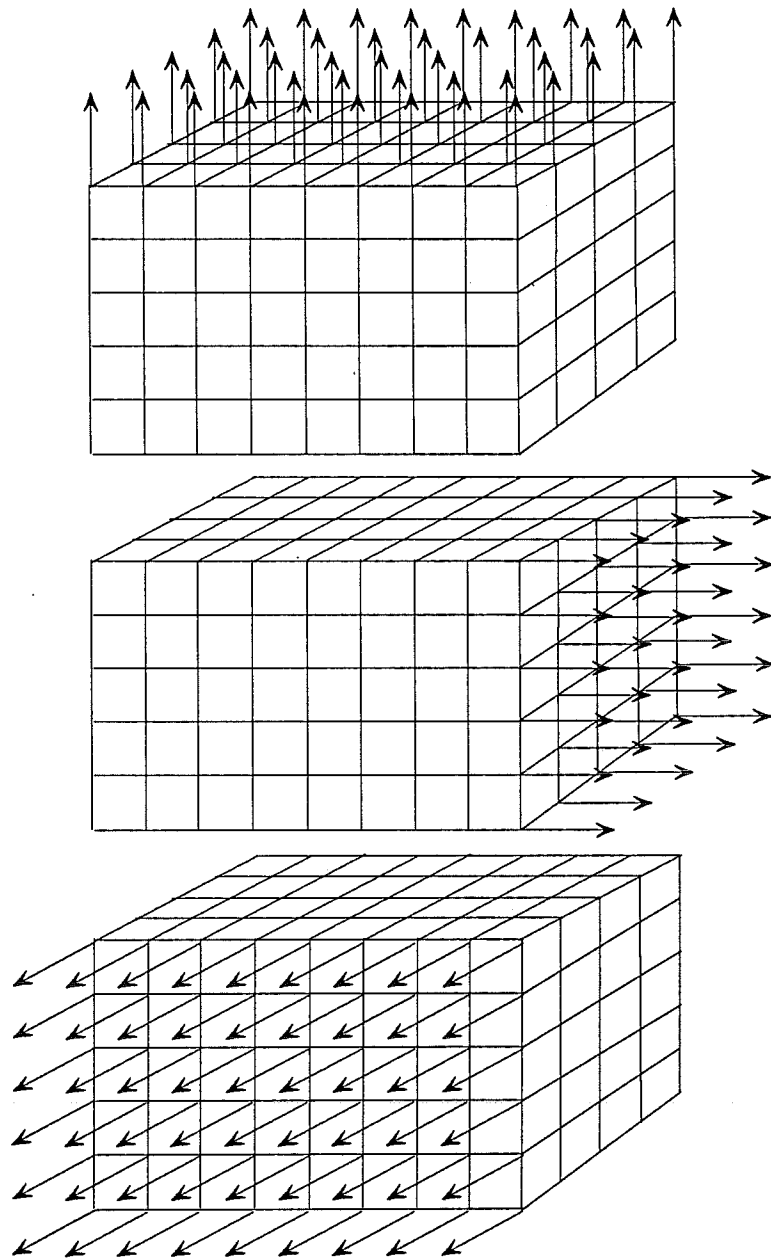


figure 7

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON

*Interactive Systems Design Laboratory
Department of Electrical Engineering, FT-10
Telephone: (206) 543-6990 or 543-6061*

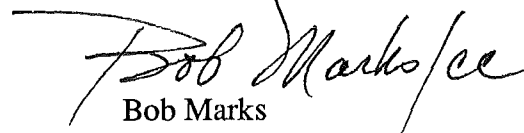
September 28, 1988

Ron Melton
Batelle Northwest
FAX: 509-376-3876
VERI: 509-375-2580

Dear Ron:

Here is a draft of the nondisclosure agreement we have been using. See you Monday!

Best regards,


Bob Marks

ROBERT J. MARKS II
Neural Processing, Optical Computers & Signal Analysis

16515 Ashworth Ave. N.
Seattle, Washington 98133
Telephone: (206) 542-0828

September 29, 1988

Eugene V. Ochs
President
Electronic Systems Inc.
7720 Woods Creek Road
Monroe, WA 98272

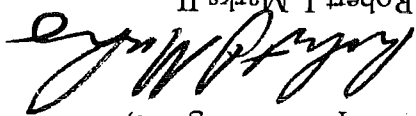
Dear Gene:

Here's a copy of the nondisclosure agreement. I hope some good things happen as a result. I'll keep you up to date on any developments.

From my notes, a contribution you made beyond the patent description was use of the lithium batteries in each cube. Your suggestion concerning bathing the unit in RF is assumed, I believe, in the last paragraph on page 3.

Please let me know if I've missed anything. Thanks for spending the time with me. I hope some exciting things result.

Best personal regards,



Robert J. Marks II

RJM:cc

cc: H. Philipp

Enclosure

NON-DISCLOSURE AGREEMENT

In connection with planned discussions and exchange of information between representatives of Ochs Electronic Systems, Inc. (hereinafter, Ochs Electronics) and Philipp Technologies Corp., Bellevue, Washington, (hereinafter, PTC) it is understood that certain information concerning neural networks, and considered confidential by PTC, will be disclosed to Ochs Electronics' representatives. Ochs Electronics and PTC wish to avoid any possible misunderstanding with respect to the disclosure of confidential information and, accordingly, agree as follows:

1. All disclosures of confidential information will be in writing and marked "confidential" at the time such writings are first furnished to the other party.
2. Ochs Electronics and its representative(s) shall maintain the identified confidential information in confidence for a period of three (3) years after receipt. During this period, Ochs Electronics shall not divulge such information to any third party or use such information for its own benefit without the prior written consent of PTC. Ochs Electronics shall treat such information with the same degree of care as it accords to its own confidential information.
3. It is understood by the parties hereto that this obligation of confidentiality shall not apply to information which is or becomes published or otherwise becomes generally available to the public through no breach of this Agreement by Ochs Electronics, or information which Ochs Electronics can show was properly in its possession prior to receipt of the disclosure from PTC, or becomes available to Ochs Electronics from an independent source without breach of agreement or violation of law.

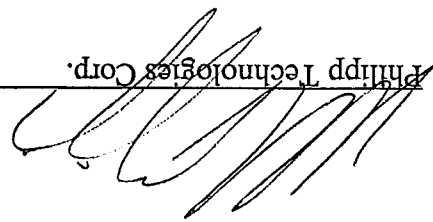
4. Confidential information regarding the technology disclosed hereunder shall remain the property of PTC. No license under any patent, copyright, trademark or trade secret is granted or implied.
5. Promptly after a receipt of a written request from PTC, and in the absence of such a request no later than thirty (30) days prior to the date of termination of this agreement as set forth below, Ochs Electronics shall return all documents concerning the confidential information to the party who furnished such items and all copies of any such documents, subject to Ochs Electronics' right to retain one copy of each such document in the files of its law department or outside legal counsel for record purposes only.

6. This agreement shall be governed by and construed in accordance with the laws of the State of Washington.
7. Any controversy or claim arising out of relating to this contract, or the breach thereof, shall be settled by arbitration in accordance with the Commercial Arbitration Rules of the American Arbitration Association, and judgment or decree upon any award or decision rendered by the arbitrator in such proceeding may be entered in any court having jurisdiction thereof. The prevailing party in any such proceeding shall be entitled to receive from the other party all attorneys' fees incurred by such prevailing party and all costs incurred in connection therewith. The locale of the arbitration shall be Seattle, Washington.

This Agreement shall remain in force and effect for one (1) year from the effective date hereof, except to the extent provided in Paragraph (2) above. The effective date shall be determined by the date affixed hereto by the party last signing this Agreement.

Pertaining to the Three Dimensional Artificial Neural Network Array as disclosed in the document dated 9/24/88, PTC

Philip Technologies Corp.

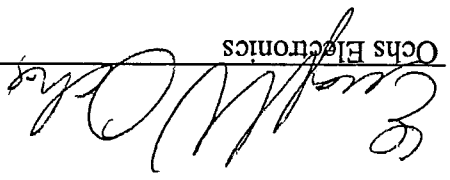


By Harold Pimm

Title President

Date 9/26/88

Ochs Electronics



By Eugene V. Ochs

Title President

Date September 26, 1988

PATENT DISCLOSURE
THREE DIMENSIONAL ARTIFICIAL NEURAL NETWORK ARRAY

Harald Philipp
Dr. Robert J. Marks II

This disclosure covers a new method of constructing electronic neural networks that permits modular fabrication. Artificial neural networks (ANN's) attempt to simulate the construction and operation of their biological counterparts. While considerable effort has been made to create such electronics, most efforts to date have concentrated on using conventional high speed serial computers designed on a highly planar structure. This is in contrast to the parallel three dimensional structures found in many biological neural systems. As a result, a primary obstacle to manufacturing more complex electronic ANN's is the degree of interconnectivity required by a large number of neurons. This disclosure describes a method for overcoming these problems.

In this disclosure, a three dimensional ANN architecture is described which is based on a building block approach. The basic construction element is three-dimensional. For sake of discussion we will use a cube but spheres, polyhedron, or other arbitrary three-dimensional shapes can also be used. As is illustrated in the cube example in Figure 1, each construction element contains a processing element such as a microcomputer. This cube has at each of its edges or sides or both (for sake of discussion edges will be assumed) a series of electrical connections which are used to communicate with adjacent neurons. Such connectors carry information relating to the state of one or more neurons, plus electrical power to permit the neurons to function.

These cubes may be stacked in volumetric fashion, e.g. the 5x5x8 cubic array as shown in Figure 2. Other arbitrary stackings may be obtained by simply ordering cubes differently. Nor is it necessary to have three stacking dimensions; an array could be laid out as a planar geometry, for example as simply 5x5x1, or as a linear array, for example 5x1x1. Neither do we require the same number of neurons in each layer. The resulting dimensions of the ANN is dictated only by the geometry of the basic construction element.

It may be seen that as each neuron cube consumes power, the power is converted to heat which must be dissipated in some manner. The neuron cubes may be modified to permit air or coolant channels (Fig. 1) when stacked. As shown, these channels would be designed to automatically couple when the units are connected.

A stack of neurons with springy interconnections must be somehow made to compress together to make good electrical contacts through-out. This can be accomplished with external pressure plates from all sides of the array (Fig. 1). Dummy construction elements containing no electronics can be used to fill out the geometry to a rectangular box to allow for better pressurized mechanical coupling.

)

Another mechanical method of interconnecting such arrays is to have each cube snap together with adjacent cubes, obviating the need for external pressure plates. Cubes may also be simply cemented together or adhered via any of a number of commercially available means, or through the attraction of magnets imbedded in each cube.

The ANN will operate in three modes: programming, learning and operational:

(1) The type of ANN architecture to be used is established in the programming mode. The operations here include establishment of the set of neurons to which a given neuron is (directly or indirectly) connected and the (sigmoidal) nonlinearity to be used by the neuron.

(2) In the learning mode, the interconnect weights among neurons are established using training data or, in certain applications such as combinatorial search problems, some training algorithm. When training data are used, some or all of the neurons are assigned certain states. The interconnect weights are then determined internal to the ANN by algorithms both known and yet to be discovered. In certain training algorithms, the initial interconnect weights are algorithmically specified by, say, a random number generator.

(3) In the operational mode, the neuron cubes perform three primary functions: a) computation of the neuron state which is a function of the neurons to which it is connected, b) conversion of the neuron's state into an electrical signal, c) retransmission of neuron states from other adjacent neurons to yet other neurons in a message passing type of procedure.

)

The interconnects from a neuron to the set of neurons with which it communicates are stored within the neuron cube with the corresponding cube addresses. In the learning process, these values are established algorithmically (possibly iteratively) as a function of the states desired in the operational mode. This is done internally to the ANN, for example, by imposing desired states on a class of neuron cubes, letting the ANN compute the states at some other group of neuron cubes, and computing the difference of this value and the states desired. This error is then used to alter the interconnect weights to reduce or compensate for this error.

A neuron state is typically computed as the (interconnect) weighted sum of connected neuron states nonlinearly altered using some memoryless nonlinearity such as a sign function or a (biologically motivated) sigmoid. The conversion to an electrical signal of the state possibly involves scaling of the state value and generation of a destination address (each neuron contains within it an address locator number which may be used to designate its position within the neuron array) if required. Retransmission of adjacent state signals is done using a messenger function. They are employed to distribute state signals from a first neuron which generates the signal to another neuron (or a plurality of neurons) not adjacent to the first neuron.

The function of retransmission is employed to simulate the action of biological neurons which have a high degree of connectivity to numerous other neurons, some at great distance from the source neuron.

In any physical geometry of electronic neurons, this connectivity aspect represents a real problem. Allowing autoconnects, for example, in a 10x10x10 neuron array, it is possible to require up to one million interconnection paths in some algorithms. Wiring such a set of interconnections is clearly extremely difficult physically.

In the structure outlined here, all interconnects among non-adjacent neurons are performed by having other neurons retransmit the sending state signal until the signal reaches its destination. Additionally, it is possible for a signal to be broadcast to a defined subset of all neurons, or even all neurons, via specially encoded messages. This is taken care of in the address portion of the signal. As a simple example, one neuron may transmit a signal to one full layer of the array with a single transmission properly encoded with address information. Or, it could address all elements of the array at once.

In cases where a neuron typically communicates with a very large number of other neurons, the interconnects may also provide for a global communications path. Such a path would consist of an electrical interconnection common to all neurons (or perhaps a large subset of all neurons), which would facilitate the transmission of a signal from any one neuron so connected to all other neurons on the common connection, simultaneously. The design would require fault tolerance to any failure of a neuron on the interconnect which might 'hog' or clamp the global interconnect, rendering it useless. Fortunately, as with biological neural networks, such fault tolerance is characteristic in many ANN algorithms.

Algorithms for inter-neuron communication need to be designed to facilitate such relayed state information. Alternatively, each neuron could also contain a separate communications processor, perhaps hard wired in silicon (i.e. not implemented in software) for higher speed. The microcomputer would then be free to compute its new state from its existing state and new transmissions received from other neurons.

Each neuron must thus contain a communications handler whose purpose is to receive, redirect, and generate state signals. Each neuron must also contain a computational element for computing state changes, and for applying weights to signals received from other neurons and also perhaps to weight its own outgoing signal. It must contain memory for program storage, which may be in the form of read-write, read-only, or read-mostly memory. It must contain read-write memory for storing parameters associated with changes in state and state weighting functions.

Neuron addresses may be either programmed permanently into each neuron prior to assembly of the array, or, preferably, would be self-programmed on power-up of the array. For example, a neuron cube in the top left corner could through internal software ascertain its position simply via the fact that certain of its sides are not connected to other cubes. It could then communicate to adjacent cubes its position, allowing adjacent neurons to determine their locations and hence addresses. The process can propagate automatically through the entire array until completed and all neurons have assigned

themselves addresses; the addresses would be stored in read-write memory or read-mostly memory in each neuron.

The interconnects may be simple mechanical contacts, perhaps spring loaded, which touch and make contact with adjacent neurons. If each neuron is a cube having 12 edges and 6 faces, then each neuron may communicate with up to 18 adjacent neurons. A neuron cube with corners modified and connectors placed on the corners may communicate with up to 26 adjacent neurons (Fig. 3). Power may be obtained from these connectors as well. External power applied to the sides of the array would flow through these interconnects.

One primary characteristic of a neuron is its reprogrammability, in the sense that the other neurons it communicates with may be reprogrammed to be more or less restrictive. A neuron may "grow" communications paths to other neurons during a learn cycle, or similarly destroy such paths. It may also modify state weights on its own. Also, it may be desirable to modify the actual structure of the microcomputer program, either on its own through a learning process or through external intervention. For example, during development of a neural network computer the cubes may require program modification. A human programmer may then create a new microcomputer program and load this program into the array. Since neurons imbedded deeply in the array are unreachable by direct electrical contact, the program may be 'downloaded' into each neuron via the retransmission process, or into just a subset of the array. A single neuron may be used as an entry node to facilitate the downloading. The programs may be loaded into the array via a conventional computer. Weights and communications paths may also be loaded into the array on a neuron by neuron basis if required by a similar process.

The ability to download neural information may be complemented by an 'upload' feature used to extract all neuron state and program information, especially information and programming of a variable nature. This is a critical feature for saving neural state information permanently onto hard media, such as a magnetic or optical disk. On power down of the network, all such information may be otherwise lost. Also, if a neural network is to be replicated in mass production with specific programming, such uploads are crucial to extracting the information required for duplication. Only then can the extracted information be reprogrammed into one or more other similar neural networks which, for example, may utilize a higher speed operational mode dedicated architecture. If this process cannot be performed, it may be required to unnecessarily teach each network individually, a process which can be tedious and impractical. The upload/download techniques are a form of cloning akin to software duplication of a conventional computer's programs and information.

Another related issue is fault tolerance. If thousands of neurons are employed in a network, failures of neurons are inevitable. The software in each neuron must be designed to tolerate failures. For example, a communications failure of a single neuron may block transmission of messages among many other neurons. Considerable thought must be given to making communications automatically

reroutable if such failures occur. It is possible to design a neuron algorithm such that an adjacent neuron could 'take over' the functioning of a bad neuron.

Since each neuron contains a digital computing element, it is possible for each neuron to simulate a number of neurons at once. The 5x5x8 array shown may actually be made to simulate not 200 neurons but 800 neurons if each neuron cube simulates the action of four neurons. Communications among such 'internal' neurons may be facilitated with appropriate software. Communications among neurons would be quite similar except that additional burden would be placed on the inter-cube electrical connections.

Signals external to the array must be interfaced in such a manner as to permit large amounts of data throughput. The sides of the array and the open connections found on the sides may be so used. Both data input and output may be so facilitated. It is also possible to focus an image of data on one or more sides of the array by incorporating photodetectors and appropriate detection electronics into neurons on each such side. Alternatively, special cubes may be affixed to each such side with photoreceptive properties, and little or no neural simulation ability. Energy fields other than light may also be used, such as microwave, sound, radiation, etc.

INVENTIVE ASPECTS

The inventive aspects of the proposed neural network we believe include but are not limited to the following:

1. A design for a neural network comprising a plurality of three dimensional structures or cells, each such cell having an ability to electrically interconnect on a plurality of sides or edges of each such cell and each having an ability to simulate the characteristics of a neuron to varying degrees of modification.
2. An ability to construct an arbitrary stacking of such cells into an array essentially without restriction or limit except for a requirement of physical contact with adjacent cells of similar type.
3. An ability of each cell within the array to electrically communicate one or more of programs, data, or commands, the cells in general having an ability to originate, retransmit, receive and reconfigure as a function of such communications.
4. Several electro-mechanical means for interconnecting cells by stacking, involving one or more of: compression mated contacts, plug-together mechanisms, adhesive mating methods, or magnetic attraction.
5. A communications interconnection among cells which permits global or large-subset transmissions among cells, without

PHYSIO CONTROL

Corporate Headquarters
11811 Willows Road Northeast
Post Office Box 97006
Redmond, WA 98073-9706 USA

Telephone: 206/867-4000
Telex: 990211 D PHYSIO RDMD
Telefax: 206/881-2405

November 24, 1987

Les E. Atlas
Assistant Professor, Electrical Engineering
401 Electrical Engineering Building, FT-10
University of Washington
Seattle, WA 98195

Dear Les:

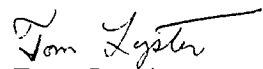
Physio-Control Corporation is interested in supporting your work in artificial neural networks. We recognize that your research in this area is part of an ongoing project within the Washington Technology Center, and we plan to give an unrestricted gift of \$10,000 to the Washington Technology Center at the University of Washington to supplement this research.

We approve of matching funds from the National Science Foundation Presidential Young Investigator Award and we approve of the publication of our participation in funding your work.

We are looking forward to many potential applications of artificial neural networks in solving challenging problems in science and industry.

Sincerely yours,

PHYSIO-CONTROL CORPORATION


Tom Lyster
Senior Research Engineer

TDL/msm

cc: Clif Alferness
John Adams

) requiring the retransmission function among cells.

6. An ability of each cell to perform computations on data received from other cells within the array or external to the array. A further ability of each cell to originate communications to one or more other similar cells, the communicated data or programming being dependent on an algorithm and on the nature of communications from other cells prior to the communication.
7. An ability for cells to self-determine their locations within an array by an algorithm and the communications means.
8. An ability for such an array and its component cells to propagate programs and data from an external source, either to all cells in an array or to a subset thereof.
9. An ability for such an array and its component cells to have programs and data extracted from it via an external computer or controller, either for storage, analysis, or duplication purposes.
10. The use of specially designed or programmed interface cells on one or more faces of the array, engineered to permit communications to and/or from external sources. The further use of light or other radiative means to couple either into or out of such cells in order to simplify the task of connection, and the use of radiatively active transducers such as phototransistors and light emitting diodes to facilitate such external interface coupling.
11. An ability of functional cells to ignore malfunctioning cells via communications methods and algorithms governing the communications paths. A further ability of other cells to simulate the functions of malfunctioning cells if required.
12. An ability of a cell to simulate more than one neuron via computational algorithms, and to communicate information from such simulations to other cells in the array via similar communications means.

Disclosed by the undersigned this day, _____, 1988.

Harold Phillipp

)

Dr. Robert J. Marks II

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

Department of Electrical Engineering, FT-10
Telephone: (206) 543-2150

December 11, 1987

Tom Lyster
PHYSIO-CONTROL
11811 Willows Road Northeast
Post Office Box 97006
Redmond, WA 98073-9706

Dear Tom,

I would like to thank you and the Physio-Control Corporation for the \$10,000 gift to help support my research at the Washington Technology Center. Our work in artificial neural networks will greatly benefit from the needed help and, I hope, future collaboration in problems of mutual interest. I am now in the state of research where the identification of important applications of this new neural network technology is crucial. The fast and accurate automatic identification of temporal patterns such as ECG signals is one of these applications. I look forward to speaking with you about this application in the future and would be willing, of course, to present more talks on artificial neural networks at your company.

I have put you on the mailing list for weekly seminars which our research group (the Interactive Systems Design Lab) hosts. These seminars are held at 3:30 PM Wednesdays when classes are in session. As you will see in the upcoming announcements, many of these seminars relate to artificial neural networks. Please let me know if anyone else at Physio-Control would like to be on this mailing list.

Sincerely,

Les Atlas
Les Atlas, Asst. Professor

Phone:(206)545-1315

LA/la-unix

cc: Prof. Ed Stear, Director, Washington Technology Center
Prof. Robert Marks, Director, Interactive Systems Design Lab

University of Washington Correspondence

INTERDEPARTMENTAL

DEPARTMENT OF ELECTRICAL ENGINEERING, FT-10

Date: December 11, 1987

To: Prof. Stear, Washington Technology Center, FH-10

From: Prof. Atlas, Electrical Engineering, FT-10

Les Atlas

Subject: Gift from Physio-Control

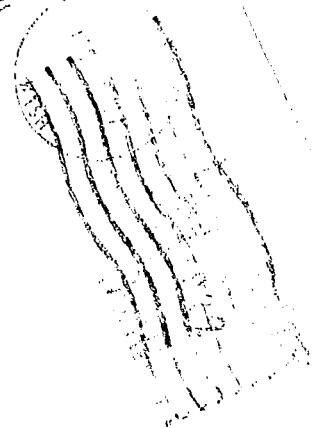
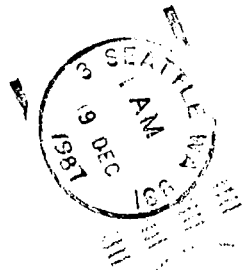
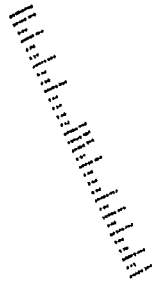
I have received a \$10,000 check from Physio-Control for my unrestricted use in artificial neural networks research within the Washington Technology Center. This form of funding is very appropriate for our current research direction. While it would be hard for us to offer short-term deliverables to industrial sponsors, the potential for industrial support is very high. Many other companies have recently expressed an interest in gift support to maintain and enhance our research program in order to "keep a foot in the door" of artificial neural networks. I therefore intend to pursue (with the help of Bob Marks) putting together a Neural Network Research Consortium to formalize this gift program. If it is possible, I would like the gift account which is established by this check to be general enough to incorporate future gifts without new budget numbers.

cc: Prof. Robert Marks
Prof. Robert Porter

**PHYSIO
CONTROL**

Corporate Headquarters
11811 Willows Road Northeast
Post Office Box 97006
Redmond, WA 98073-9706 USA

Les E. Atlas
Assistant Professor, Electrical Engineering
401 Electrical Engineering Building, FT-10
University of Washington
Seattle, WA 98195



Phillipp Technologies Corporation

13219 Northup Way, Suite 203
Bellevue, Washington 98005
(206) 746-1642
FAX: (206) 746-0566

September 22, 1988

Dr. Robert J. Marks II
University of Washington
Department of Electrical Engineering
Seattle, WA

Bob,

Attached is the nondisclosure agreement. A disk with the info is also on its way in Microsoft format (hopefully). You will need to modify it depending on who you are talking to (private individual, corporation, etc.)

I've also asked the mechanical design group (Stratos on Capitol Hill) to forward one of their brochures to you. As I mentioned, they have a high level contact at Microsoft, access to funding, an incredible CAD system, and some excellent design experience. They come highly recommended.

I will still make a first pass at the patent, and then hopefully have Tom Noe at Fluke to help touch it up and so forth.

Regards,

Hal

NON-DISCLOSURE AGREEMENT

In connection with planned discussions and exchange of information between representatives of Bellevue, Washington, (hereinafter, "PTC") it is understood that certain information concerning neural networks and considered confidential by PTC, will be disclosed to " " representatives. " " and PTC wish to avoid any possible misunderstanding with respect to the disclosure of confidential information and, accordingly, agree as follows:

1. All disclosures of confidential information will be in writing and marked "confidential" at the time such writings are first furnished to the other party.

2. " " and its representative(s) shall maintain the identified confidential information in confidence for a period of three (3) years after receipt. During this period, " " shall not divulge such information to any third party or use such information for its own benefit without the prior written consent of PTC. " " shall treat such information with the same degree of care as it accords to its own confidential information.

3. It is understood by the parties hereto that this obligation of confidentiality shall not apply to information which is or becomes published or otherwise becomes generally available to the public through no breach of this Agreement by " ", or information which " " can show was properly in its possession prior to receipt of the disclosure from PTC, or becomes available to " " from an independent source without breach of agreement or violation of law.

4. Confidential information regarding the technology disclosed hereunder shall remain the property of PTC. No license under any patent, copyright, trademark or trade secret is granted or implied.

5. Promptly after a receipt of a written request from PTC, and in the absence of such a request no later than thirty (30) days prior to the date of termination of this agreement as set forth below, " " shall return all documents concerning the confidential information to the party who furnished such items and all copies of any such documents, subject to " "s right to retain one copy of each such document in the files of its law department or outside legal counsel for record purposes only.

6. This agreement shall be governed by and construed in accordance with the laws of the State of Washington.

7. Any controversy or claim arising out of or relating to this contract, or the breach thereof, shall be settled by arbitration in accordance with the Commercial Arbitration Rules of the American Arbitration Association, and judgment or decree upon any award or decision rendered by the arbitrator in such proceeding may be entered in any court having jurisdiction thereof. The prevailing party in any such proceeding shall be entitled to receive from the other party all attorneys' fees incurred by such prevailing party and all costs incurred in connection therewith. The locale of the arbitration shall be Seattle, Washington.

This Agreement shall remain in force and effect for one (1) year from the effective date hereof, except to the extent provided in Paragraph (2) above. The effective date shall be determined by the date affixed hereto by the party last signing this Agreement.

 Philipp Technologies Corp.

 " "

 By

 Title

 Date

UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

*Interactive Systems Design Laboratory
Department of Electrical Engineering, FT-10
Telephone: (206) 543-6990, 543-6061 or 543-2150*

November 4, 1988

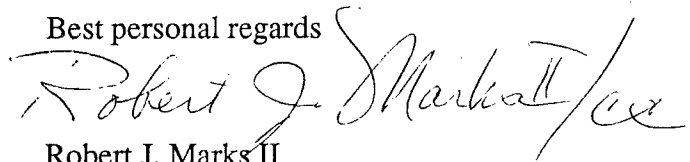
Dr. Dmitry Kaplan
208 Mountain Park Boulevard
Apt. E302
Issaquah, WA 98027

Dear Dmitry:

Here's the BAA from China Lake and the SDI effort. If you do call either Swenson (China Lake) or Bromley (SDI), and you actually talk to them, please mention that you're talking about the project that I called them about so they know that they don't have to call me back.

Let's get some contracts, have fun and get rich!

Best personal regards

A handwritten signature in cursive script that reads "Robert J. Marks II". The signature is written in dark ink and is positioned to the right of the typed name.

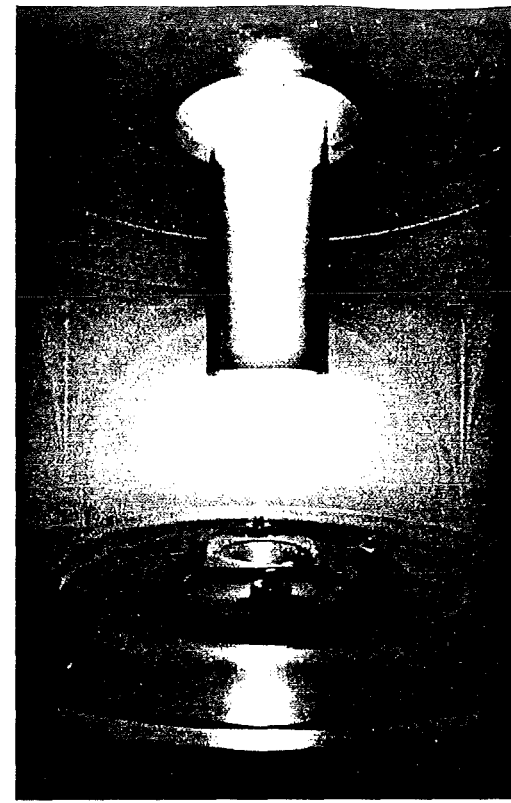
Robert J. Marks II
Professor & all round swell guy

OVER THE COURSE OF THE TWENTIETH CENTURY, A

few ambitious initiatives have captured the imagination and intellect of the nation's leading scientists and engineers: the Manhattan project, Apollo moon missions, and, now, the Strategic Defense Initiative (SDI).

SDI's goal—to eliminate the nuclear threat—demands the best and brightest. Its enabling technologies, spanning advanced computing, materials, propulsion and energy sources, create exciting opportunities for researchers. These include:

- The opportunity to contribute to a critically important defense science initiative.
- The opportunity to work with leaders in academia, government, and industry on next-generation technologies.
- The opportunity to pursue promising innovations.

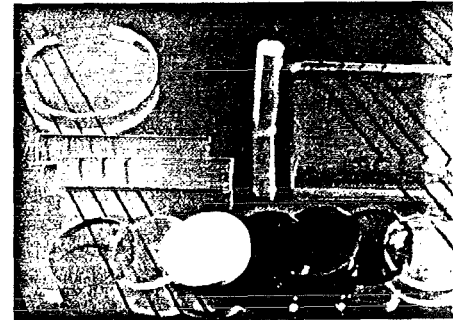


■ Chemical vapor deposition system builds large diamond films for advanced electronics.

IST nurtures and supports programs related to the SDI mission from fundamental research into scientific feasibility of concept, to exploration of engineering feasibility, to demonstrating practicability.

By providing a responsive, flexible, and stable management structure, IST has fostered innovation. It provides the direction, coordination, and funding necessary to carry out a large-scale diversified research effort.

We welcome the interest and involvement of scientists and researchers. Specific program administration information, as well as a list of Science and Technology Agents, begins on p.4. The program summaries and case examples provide an in-depth look at the types of innovation sought by IST.



■ Unique sol-gel approach casts optical glasses of unequalled size and purity.

THE STRATEGIC DEFENSE INITIATIVE ORGANIZATION

was created to explore the development of a defense system envisioned by President Reagan in his address to the nation on March 23, 1983.

To fulfill its mission, SDIO is organized into two primary areas, "technology" programs and "systems" programs. The Innovative Science and Technology Office is the technical directorate within SDIO tasked with seeking out innovative approaches to all aspects of ballistic missile defense. It funds research in these approaches and assures that the other technical directorates within SDIO are apprised of new results and breakthroughs from IST programs.

The IST office has several roles. First, it establishes a technology base for strategic defense via fundamental research conducted in universities, government and national laboratories, small businesses, and large industries. Second, it brings infant technologies to a stage where they can be validated. These technologies either transition into applications or go on the shelf for future exploitation. Third, the IST Office administers the SDIO Small Business Innovation Research Program.

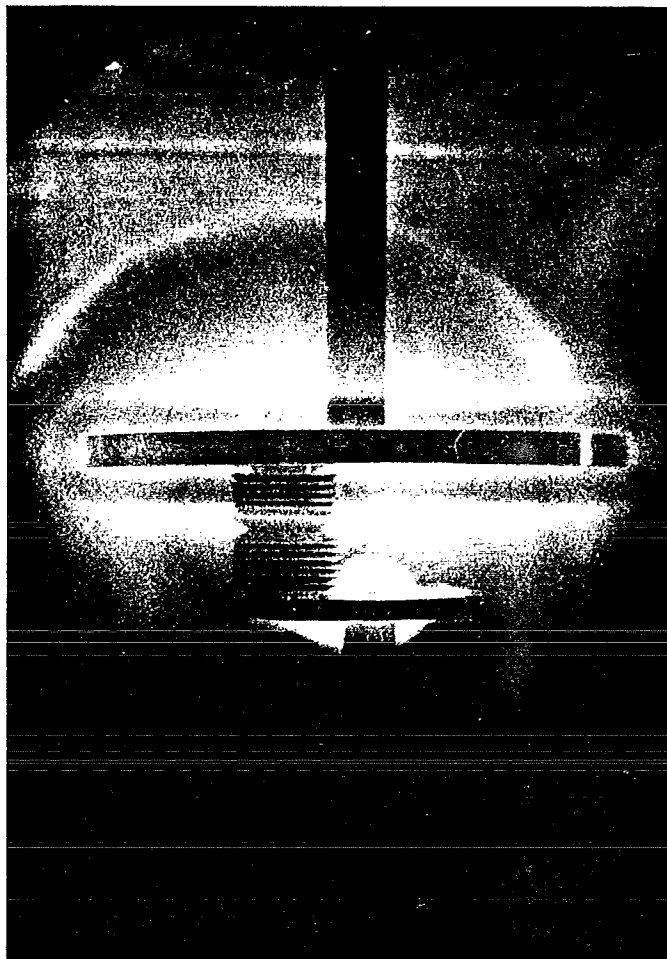
THE CURRENT RESEARCH PROGRAM SUPPORTED BY

IST focuses on six general areas:

- High speed computing
- Sensing, discrimination and signal processing
- Space power and power conditioning
- Directed and kinetic energy concepts
- Materials and structures
- Propulsion and propellants

Other areas may be added in the future.

While the range of programs is broad, they all fulfill the IST criteria: each is directed toward revolutionary (not evolutionary) advances; each relates to some aspect of the strategic defense system — its architecture, weapon system or sensing components, or command and control; each blends the best thinking in academia, government and industry.



■ Vacuum outgassing experiments help scientists understand electricity behavior in the ionosphere.



Looking Ahead

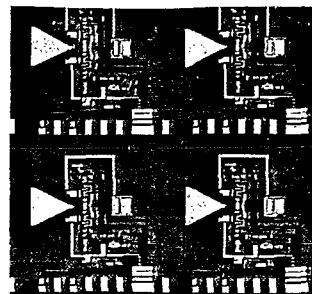
Scientists at the Naval Research Laboratory are coordinating their studies of ultra-short wavelength lasers with research being conducted at the University of Rochester, University of Texas, Stanford, and Physics International Company—as well as x-ray and gamma-ray laser applications being demonstrated by innovative small businesses.

At the Jet Propulsion Laboratory (JPL), researchers have achieved major gains in computing speed using a “hypercube” network of multiple computers working simultaneously on different pieces of a problem. JPL researchers are also studying neural networks as a faster, fault-tolerant alternative to conventional numerical processing.

The Innovative Nuclear Space Power Institute (INSPI), a consortium of universities and small businesses, promotes new technologies to meet the needs of SDI space platforms for efficient, lightweight, high-power energy sources. INSPI is concentrating its efforts in two most promising areas: the gas core reactor and TRICE (thermionic reactor inductively coupled elements).

These examples of innovation and multidisciplinary teamwork are the norm, not the exception, at IST. They exemplify the importance IST places on its mission of disseminating technical knowledge.

■ Dense array of superconducting circuits typifies the state-of-the-art in electronics.



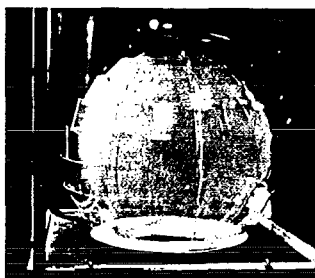
IST's EFFORTS AIM AT THE DEVELOPMENT OF AN

effective strategic defense system. As the architecture of the United States' SDI system emerges and moves toward deployment, IST's role will adjust accordingly. Some research programs will become more narrowly focused on key enabling technologies to support SDI system specifications.

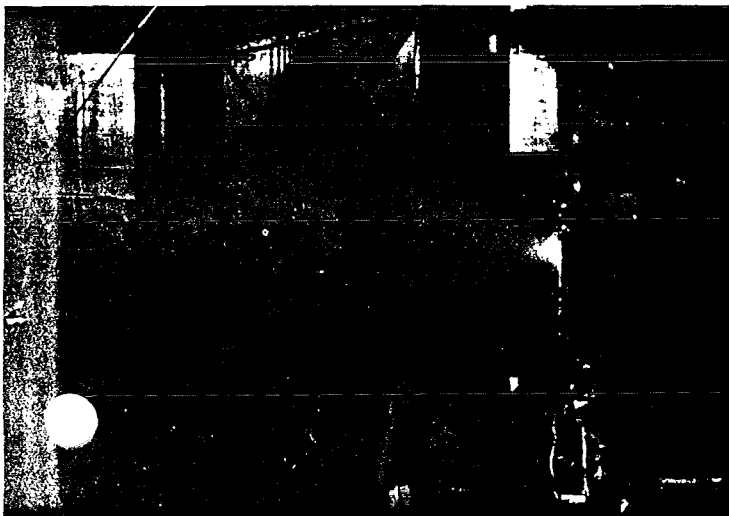
IST's research findings will also extend beyond strategic defense. Advances in computing, sensing, and electronic materials will contribute to the nation's entire defense effort, as well as to civilian applications in industries like electronics, telecommunications, and automotive. Likewise, specialized lasers being developed under IST sponsorship will find medical diagnosis and treatment applications. The benefits of advanced energy and propulsion technologies will flow to NASA, commercial space ventures, and consumer products.

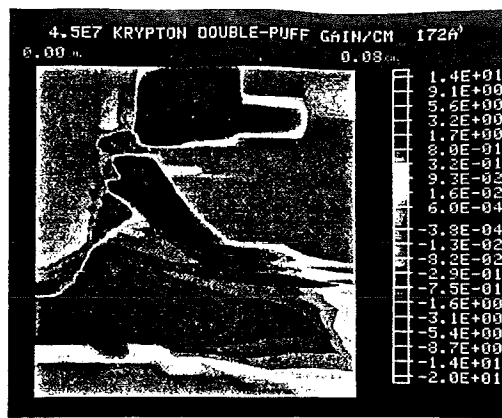
In a very direct way, IST's efforts are key to the nation's technology future—as a catalyst for scientific and technical achievements of the 21st century.

■ Mock-up of spherical gas core reactor models power density, mass flow and heat transfer.



■ Electron-beam experiments reveal mysteries of x-ray lasing.





■ Numerical simulation of x-ray laser gain as a function of radius (horizontal) and time (vertical).

TECHNICAL MANAGEMENT OF THE WIDE DIVERSITY OF

IST-sponsored research is conducted by the directorate's Science and Technology Agents (STAs). The STAs are affiliated with such defense research agencies as the Office of Naval Research, the Air Force Office of Scientific Research, and the Army Research Office. Administration, procurement, and reporting are generally carried out by the parent agencies of the cognizant STAs.

The STAs are the official representatives of SDIO and IST. Generally research will be funded by IST only with STA review and recommendation. Thus, proposals and inquiries should be sent to the appropriate STA, not to the IST directorate office. Each program description includes the name, address and phone number of the responsible STA. The information is also summarized in Appendix A of the brochure.

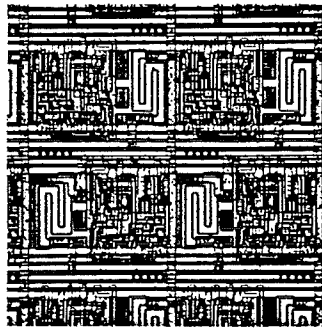
Researchers wishing to propose IST-sponsored projects should identify a program component. Then, they should directly contact the appropriate STA to initiate a dialogue regarding IST support. If, as a result of this

dialogue, an STA determines that a proposed effort is of interest to SDIO, the researcher will be encouraged to follow up with additional documentation—for example, a white paper or a formal proposal.

When an investigator wishes to propose a program in an area for which no STA is identified, a brief two-page summary of the program should be sent to IST. The summary should stress the innovative nature of the proposed work, its relationship to perceived SDIO needs, and potential results. Appendix B of the brochure contains a sample format.

IST encourages early contact with STAs regarding novel and innovative concepts or approaches in any scientific or technology discipline applicable to the Strategic Defense Initiative. We are seeking revolutionary advances that can have high payoffs in enhancing strategic defense—and, we are seeking broad participation from the entire research community.

The following pages describe IST program components and the crucial technology challenges they address—challenges that point the way to tomorrow's technology frontiers.



■ Neural networks like this promise dramatic increases in computing speed.

Contents

HIGH SPEED COMPUTING

- 6 Optical Computing and Optical Signal Processing
- 7 Parallel Processing
- 7 Mathematical Methods and Algorithms
- 8 Self Adaptive Processing and Simulation

SENSING, DISCRIMINATION AND SIGNAL PROCESSING

- 9 Detectors for Sensing and Discrimination
- 10 Optical Sensors
- 10 Reliable Advanced Electronics
- 11 Integrated Detection Estimation and Communication Theory
- 12 Laser Satellite Networking
- 13 Boost Phase Detection
- 14 Terahertz Technology
- 15 Interactive Discrimination

SPACE POWER AND POWER CONDITIONING

- 16 Non-Nuclear Space Power and Power Conditioning
- 17 Advanced Pulse Power Physics
- 18 Nuclear Space Power
- 19 Advanced Electro-Chemical Prime Power

DIRECTED AND KINETIC ENERGY CONCEPTS

- 20 Electromagnetic Propagation and Directed Energy Concepts
- 21 Short-Wavelength Chemical Lasers
- 22 High-Power Microwave Sources
- 23 Advanced Beam Combining Concepts
- 23 Advanced Accelerators
- 24 Particle Beams
- 25 KE Interceptor Integration
- 26 Ultra Short Wavelength Lasers
- 26 Propagation Through Disturbed Environments
- 27 Mid-Atmospheric Effects

MATERIALS AND STRUCTURES

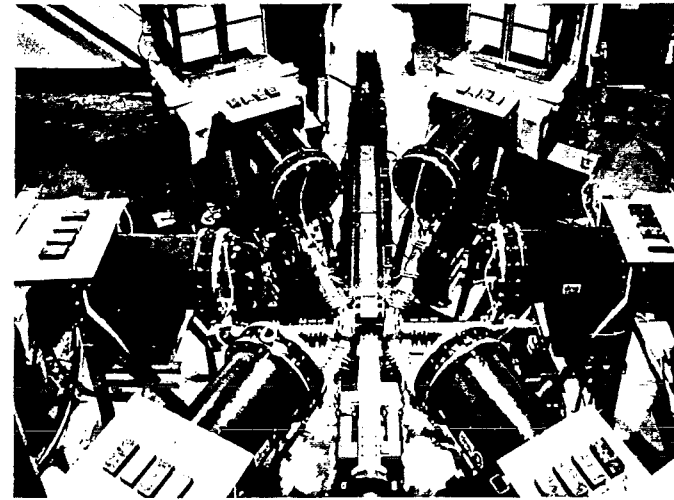
- 28 Advanced Composite Materials
- 29 Electronic and Optical Materials
- 30 Diamond Technology
- 30 Electronic-Materials Interfacing
- 31 Optical Glass and Macromolecular Materials
- 32 Space Structures and Dynamics
- 33 High Pressure Metastable Materials
- 33 Optical Sensor Survivability
- 34 Superconducting Materials
- 35 Interactive Space Technologies

PROPULSION AND PROPELLANTS

- 36 Electric Propulsion
- 37 Advanced Propellants
- 38 Low Emission Propellants

39 APPENDIX A

41 APPENDIX B



■ CHECMATE electromagnetic launcher impels projectiles to record speeds.

HIGH SPEED COMPUTING


**Optical Computing and
Optical Signal Processing**

 STA:
Dr. William Miceli

 Office of Naval Research
495 Summer Street
Boston, MA 02210-2109
(617) 451-3172

Objective:
Optical computing refers to the exploitation and application of suitable optical technology within a computational environment. This program is predicated upon the inherent parallelism of optical systems and addresses all computational aspects associated with the SDI, particularly sensor signal processing, target/decoy discrimination, and the data management functions associated with BM/C³. This research addresses all aspects of optics, opto-electronics, and acousto-optics applicable to analog signal processing, digital computing, and biologically inspired neuromorphic

computing—commonly referred to as neural networks.

Program Description:
The program consists of research efforts in the following areas:

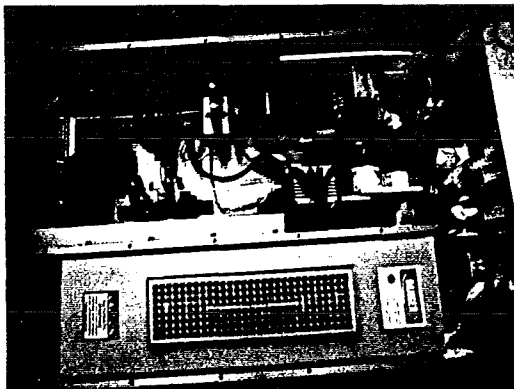
- Optical (Analog) Signal Processing
- Optical Digital Computing
- Optical Neural Networks

Emphasis is placed upon devising suitable processing architectures in each of these areas, as well as developing the technology base necessary to optically implement these architectures.

Opportunities:
The following topics have

been identified as particular areas of interest:

- 2-dimensional arrays of bistable optical devices with acceptable power dissipation, packing densities, speed, signal/noise, fabrication costs, etc.
- Spatial light modulators with suitable data formatting, speed, modulation depth, power consumption, optical quality, etc.
- Dynamically reconfigurable optical interconnections between cascaded 2-dimensional arrays of optical devices. The need for real-time holographic elements, preferably with gain, is anticipated.



Prototype optical CPU

OPTICAL TECHNOLOGY FOR COMPUTING ■ Benefits flowing from the use of optical technology for communications (like fiber optic telecommunications) are well-established, and range from wider bandwidth and higher speed transmission to non-interference of lightwaves and low power consumption. Optical technology promises similar benefits for computing. As trends in computer architecture evolve from conventional sequential processors to multi-processor systems, optics can dramatically improve communications between processors. Also, recent developments in low-power nonlinear optics can trick photons into interacting with one another, thus facilitating 1) switching networks, 2) logic gate arrays and 3) other devices that form

the basis of novel computational approaches known as optical neural nets. Such neural networks mimic the human brain. ■ Leading universities and government laboratories, directed by Dr. William Miceli at the Office of Naval Research, collaborate on this research

C A S E E X A M P L E

in optical (analog) signal processing, optical digital computing and optically-based neural networks. Current research focuses on building the technology base and devising processor architectures that compete with conventional computing methods.

Build a Brain

Parallel Processing

STA:
Dr. Keith Bromley

Naval Ocean Systems Center
Code 47 1T
271 Catalina Blvd.
San Diego, CA 92152
(619) 553-2535

Objective:

The proposed SDI mission requires improvement in the performance of current sensor technology and signal processing. This research intends to provide extremely high throughput computing techniques for use in SDI signal processing.

Program Description:

The program's primary research centers on the investigation of algorithmically specialized systolic processors. Unlike in

past systolic work, study is being conducted on system level issues. Efforts to expand the systolic style of design into the middle ground between algorithmically specialized devices and general purpose parallel machines are being considered. One technique already in this middle ground is the programmable systolic array.

Opportunities:

Future studies will continue to exploit systolic

processor technology for advancement in signal processing. Approaches which will be investigated will include SATCOM and control computations. Because of long SDI mission times and the likelihood for serious battle damage, fault tolerance techniques will be pursued. Approaches include reconfiguration in a wafer scale integration context and behavior-based fault detection at the system level.

Mathematical Methods and Algorithms

STA:
Dr. Jagdish Chandra

Army Research Office
P.O. Box 12211
Research Triangle Park, NC
27709-2211
(919) 549-0641

Objective:

The command, control, and data manipulation phases of proposed SDI systems demand state-of-the-art parallel supercomputers, algorithmic processes and technologies. This research intends to explore applicability of related fields in large scale scientific computing.

Program Description:

The program concerns

itself with studies in the following areas:

- Parallel methods in high-speed computing
- Systolic algorithms for signal processing
- High resolution imaging
- 3-dimensional robotic vision and shape recognition
- Segmentation and image detection in natural images.

Current investigations are restricted to parallel computer hardware development.

Opportunities:

SDI requires the use of supercomputers and parallel computers in the design, testing, and implementation phases. Software for these systems needs extensive algorithmic development and should be able to run on a

variety of parallel architectures. Accomplishments in signal and image processing center on algorithmic improvements.

Special interest lies in the interaction between mathematical methods and algorithms, and systolic array architectures (in reference to large-scale parallel computation). Specific targets for the investigation of large-scale optimization prob-

lems with parallel computers will include:

- Understanding the problem
- Formulation of mathematical models appropriate to parallelism
- Development of appropriate computer languages, data structures and implementations for parallelism and adaptivity on various parallel architectures
- Development and implementation of systolic algorithms for linear algebra
- Signal processing
- High resolution imaging and switching applications
- Development and implementation of parallel algorithms for graphical display and animation of solutions.

Self Adaptive Processing And Simulation

STA:
Mr. Doyce Satterfield

Army Strategic Defense
Command
Attn: CSSD-H-VP
Processing Technology Division
P.O. Box 1500
Huntsville, AL 35807-3801
(205) 895-4819

Objective:

This research intends to accomplish a two-fold task. It will devote part of its efforts to the study of innovative concepts for advancing processing technology for system self modification while deployed in a dynamically evolving threat environment. In addition, research will be conducted to facilitate the simulation of the SDI battle management and C³ network. The effort will demonstrate the advanced features of:

- Automated model generation
- Automated analysis of simulation results
- Goal-directed instrumentation

- Integration of adaptive hardware and software with the simulation(s).

Program Description:

The program focuses on the self-adaptive simulation research. Modeling of the entire BM/C³ system will require the correct modeling of subsystems, the various threats, and the multidimensional environment. This effort will entail millions of lines of code and necessitate advancement of the state-of-the-art in computer-aided model generation, instrumentation of models, and simulation analysis.

Opportunities:

Future study will deal

with the extensive SDI C² network with its many local nodes. Advancements in intelligent communicating agents that would reside at these local nodes are sought, since having the agents reside at the nodes will allow for adaptation to changing environments. Other required technologies and approaches include:

- Search, acquisition, and track networks using self-adaptive processing for in-circuit reconfigurability of logic and architecture
- Adaptive hardware and software for self-diagnosis, self-repair, and self-modification during operation.

PATENT DESCRIPTION:
THREE DIMENSIONAL ARTIFICIAL NEURAL NETWORK ARRAY

(Confidential & Proprietary)
(contains 6 pages plus seven figures)

9-22-88

Harald Philipp
Robert J. Marks II

In this description, we present a new method of constructing electronic neural networks that permits modular three dimensional fabrication. Artificial neural networks (ANN's) attempt to simulate the construction and operation of their biological counterparts. While considerable effort has been made to create such electronics, most efforts to date have concentrated on using conventional high speed serial computers designed on a highly planar structure. This is in contrast to the parallel three dimensional structures found in many biological neural systems. As a result, a primary obstacle to manufacturing more complex electronic ANN's is the degree of interconnectivity required by a large number of neurons. This disclosure describes a method for overcoming these problems.

In this disclosure, a three dimensional ANN architecture is described which is based on a building block approach. The basic construction element is three-dimensional. For sake of discussion we will use a cube (Figure 1) but spheres, polyhedron, or other arbitrary three-dimensional shapes can also be used. A hexagonal neural construction unit is shown in Figure 2. As is illustrated in the cube example in Figure 1, each such construction element contains a processing element such as a microcomputer. This cube has at each of its edges or sides or both a series of electrical connections which are used to communicate with adjacent neurons. Such connectors carry information relating to the state of one or more neurons, plus electrical power to permit the neurons to function.

These cubes may be stacked in volumetric fashion, e.g. the $8 \times 5 \times 4$ cubic array as shown in Figure 3. Other arbitrary stackings may be obtained by simply ordering cubes differently. Nor is it necessary to have three stacking dimensions; an array could be laid out as a planar geometry, for example as simply $5 \times 5 \times 1$, or as a linear array, for example $5 \times 1 \times 1$. Neither do we require the same number of neurons in each layer. The resulting dimensions of the ANN is dictated only by the geometry of the basic construction element.

It may be seen that as each neuron cube consumes power, the power is converted to heat which must be dissipated in some manner. The neuron cubes may be modified to permit air or coolant channels (Figure 1) when stacked. As shown, these channels would be designed to automatically couple when the units are connected. Alternatively, the geometry of the basic construction element can be modified to commit a large percentage of the volume to coolant flow. An example that can be used in lieu of the cube is shown in Figure 4. A single construction element is shown of top. A 2×2 array of these elements is shown on the bottom.

A stack of neurons with springy interconnections must be somehow made to compress together to make good electrical contacts through-out. This can be accomplished with external pressure plates from all sides of the array (Figure 1). Dummy construction elements containing no electronics can be used to fill out the geometry to a rectangular box to allow for better pressurized mechanical coupling.

Another mechanical method of interconnecting such arrays is to have each cube snap together with adjacent cubes, obviating the need for external pressure plates. Cubes may

also be simply cemented together or adhered via any of a number of commercially available means, or through the attraction of magnets imbedded in each cube.

The ANN will operate in three modes: programming, learning and operational:

(1) The type of ANN architecture to be used is established in the programming mode. The operations here include establishment of the set of neurons to which a given neuron is (directly or indirectly) connected and the (sigmoidal) nonlinearity to be used by the neuron.

(2) In the learning mode, the interconnect weights among neurons are established using training data or, in certain applications such as combinatorial search problems, some training algorithm. When training data are used, some or all of the neurons are assigned certain states. The interconnect weights are then determined internal to the ANN by algorithms both known and yet to be discovered. In certain training algorithms, the initial interconnect weights are algorithmically specified by, say, a random number generator.

(3) In the operational mode, the neuron cubes perform three primary functions: a) computation of the neuron state which is a function of the neurons to which it is connected, b) conversion of the neuron's state into an electrical signal, c) retransmission of neuron states from other adjacent neurons to yet other neurons in a message passing type of procedure.

The interconnects from a neuron to the set of neurons with which it communicates are stored within the neuron cube with the corresponding cube addresses. In the learning process, these values are established algorithmically (possibly iteratively) as a function of the states desired in the operational mode. This is done internally to the ANN, for example, by imposing desired states on a class of neuron cubes, letting the ANN compute the states at some other group of neuron cubes, and computing the difference of this value and the states desired. This error is then used to alter the interconnect weights to reduce or compensate for this error.

A neuron state is typically computed as the (interconnect) weighted sum of connected neuron states nonlinearly altered using some memoryless nonlinearity such as a sign function or a (biologically motivated) sigmoid. The conversion to an electrical signal of the state possibly involves scaling of the state value and generation of a destination address (each neuron contains within it an address locator number which may be used to designate its position within the neuron array) if required. Retransmission of adjacent state signals is done using a messenger function. They are employed to distribute state signals from a first neuron which generates the signal to another neuron (or a plurality of neurons) not adjacent to the first neuron.

The function of retransmission is employed to simulate the action of biological neurons which have a high degree of connectivity to numerous other neurons, some at great distance from the source neuron. In any physical geometry of electronic neurons, this connectivity aspect represents a real problem. Allowing autoconnects, for example, in a 10x10x10 neuron array, it is possible to require up to one million interconnection paths in some algorithms. Wiring such a set of interconnections is clearly extremely difficult physically.

In the structure outlined here, all interconnects among non-adjacent neurons are performed by having other neurons retransmit the sending state signal until the signal reaches its destination. Additionally, it is possible for a signal to be broadcast to a defined subset of all neurons, or even all neurons, via specially encoded messages. This is taken care of in the address portion of the signal. As a simple example, one neuron

may transmit a signal to one full layer of the array with a single transmission properly encoded with address information. Or, it could address all elements of the array at once.

In cases where a neuron typically communicates with a very large number of other neurons, the interconnects may also provide for a global communications path. Such a path would consist of an electrical interconnection common to all neurons (or perhaps a large subset of all neurons), which would facilitate the transmission of a signal from any one neuron so connected to all other neurons on the common connection, simultaneously. The design would require fault tolerance to any failure of a neuron on the interconnect which might 'hog' or clamp the global interconnect, rendering it useless. Such fault tolerance is characteristic with biological neural networks.

Algorithms for inter-neuron communication need to be designed to facilitate such relayed state information. Alternatively, each neuron could also contain a separate communications processor, perhaps hard wired in silicon (i.e. not implemented in software) for higher speed. The microcomputer would then be free to compute its new state from its existing state and new transmissions received from other neurons.

Each neuron must thus contain a communications handler whose purpose is to receive, redirect, and generate state signals. Each neuron must also contain a computational element for computing state changes, and for applying weights to signals received from other neurons and also perhaps to weight its own outgoing signal. It must contain memory for program storage, which may be in the form of read-write, read-only, or read-mostly memory. It must contain read-write memory for storing parameters associated with changes in state and state weighting functions.

Neuron addresses may be either programmed permanently into each neuron prior to assembly of the array, or, preferably, would be self-programmed on power-up of the array. For example, a neuron cube in the top left corner could through internal software ascertain its position simply via the fact that certain of its sides are not connected to other cubes. It could then communicate to adjacent cubes its position, allowing adjacent neurons to determine their locations and hence addresses. The process can propagate automatically through the entire array until completed and all neurons have assigned themselves addresses; the addresses would be stored in read-write memory or read-mostly memory in each neuron.

The interconnects may be simple mechanical contacts, perhaps spring loaded, which touch and make contact with adjacent neurons. If, for example, every other layer in the cube structure was phased as illustrated in the top of Figure 5, then each cube makes physical contact with 12 adjacent cubes. Sides of 14 adjacent cubes can be made to have physical contact if adjacent rows in a layer are phased as is illustrated at the bottom of Figure 5. If similar phasing is applied to the hexagonal structure in Figure 2, then each unit will also make contact with 14 other units.

Alternately, communication among construction elements can be done optically thereby eliminating the need for transmitting signals through mechanically coupled interconnects. (Note that, however, unless power can be provided internal to the construction element or through some other externally applied field, mechanical interconnects would still be required to provide power.) As is shown in Figure 6, optical sources, such as LED's, would be aligned to optical detectors at the construction element's surface through a skin of optically transparent material. Inter-element communication could be established by any one of a number of commonly used modulation techniques.

The flow of signals must be organized in such a fashion as to avoid collision of moving packets of information. For artificial neural network algorithms that require each neuron to communicate with every other neuron, this can be achieved by alternating signal flow directions as is illustrated in Figure 7. At one instance, communication can be with neuron elements in a specified direction. In the next communication cycle, this direction would change. The technique can also be modified for the less severe case to algorithms where a neuron is only required to be connected to each neuron in an adjacent layer.

One primary characteristic of a neuron is its reprogrammability, in the sense that the other neurons it communicates with may be reprogrammed to be more or less restrictive. A neuron may "grow" communications paths to other neurons during a learn cycle, or similarly destroy such paths. It may also modify state weights on its own. Also, it may be desirable to modify the actual structure of the microcomputer program, either on its own through a learning process or through external intervention. For example, during development of a neural network computer the cubes may require program modification. A human programmer may then create a new microcomputer program and load this program into the array. Since neurons imbedded deeply in the array are unreachable by direct electrical contact, the program may be 'downloaded' into each neuron via the retransmission process, or into just a subset of the array. A single neuron may be used as an entry node to facilitate the downloading. The programs may be loaded into the array via a conventional computer. Weights and communications paths may also be loaded into the array on a neuron by neuron basis if required by a similar process.

The ability to download neural information may be complemented by an 'upload' feature used to extract all neuron state and program information, especially information and programming of a variable nature. This is a critical feature for saving neural state information permanently onto hard media, such as a magnetic or optical disk. On power down of the network, all such information may be otherwise lost. Also, if a neural network is to be replicated in mass production with specific programming, such uploads are crucial to extracting the information required for duplication. Only then can the extracted information be reprogrammed into one or more other similar neural networks which, for example, may utilize a higher speed operational mode dedicated architecture. If this process cannot be performed, it may be required to unnecessarily teach each network individually, a process which can be tedious and impractical. The upload/download techniques are a form of cloning akin to software duplication of a conventional computer's programs and information.

Another related issue is fault tolerance. If thousands of neurons are employed in a network, failures of neurons are inevitable. The software in each neuron must be designed to tolerate failures. For example, a communications failure of a single neuron may block transmission of messages among many other neurons. Considerable thought must be given to making communications automatically reroutable if such failures occur. It is possible to design a neuron algorithm such that an adjacent neuron could 'take over' the functioning of a bad neuron.

Since each neuron contains a digital computing element, it is possible for each neuron to simulate a number of neurons at once. The $8 \times 5 \times 4$ array shown may actually be made to simulate not 160 neurons but 640 neurons if each neuron cube simulates the action of four neurons. Communications among such 'internal' neurons may be facilitated with appropriate software. Communications among neurons would be quite similar except that additional burden would be placed on the inter-cube electrical connections.

Signals external to the array must be interfaced in such a manner as to permit large amounts of data throughput. The sides of the array and the open connections found on