

RL-TR-95-123
Final Technical Report
July 1995



OPTO-ELECTRONIC ASSOCIATIVE MEMORIES BASED ON A MOTIONLESS- HEAD PARALLEL READOUT OPTICAL DISK

University of California, San Diego

Philippe J. Marchand and Sadik C. Esener

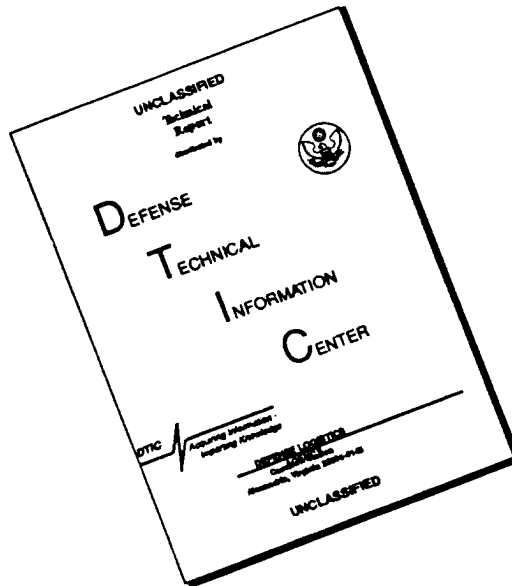
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

19960415 144

**Rome Laboratory
Air Force Materiel Command
Griffiss Air Force Base, New York**

FORM 87-1 (REV. 10-1-85)

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

This report has been reviewed by the Rome Laboratory Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RL-TR-95-123 has been reviewed and is approved for publication.

APPROVED: *Bernard J. Clarke*

BERNARD J. CLARKE, Captain, USAF
Project Engineer

FOR THE COMMANDER:



DELBERT B. ATKINSON, Colonel, USAF
Director of Intelligence & Reconnaissance

If your address has changed or if you wish to be removed from the Rome Laboratory mailing list, or if the addressee is no longer employed by your organization, please notify RL (IRAP) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)		2. REPORT DATE July 1995	3. REPORT TYPE AND DATES COVERED Final Feb 93 - Aug 94	
4. TITLE AND SUBTITLE OPTO-ELECTRONIC ASSOCIATIVE MEMORIES BASED ON A MOTIONLESS-HEAD PARALLEL READOUT OPTICAL DISK			5. FUNDING NUMBERS C - F30602-93-C-0013 PE - 62702F PR - 4594 TA - 15 WU - K1	
6. AUTHOR(S) Philippe J. Marchand and Sadik C. Esener			8. PERFORMING ORGANIZATION REPORT NUMBER N/A	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of California, San Diego Office of Contract & Grant Adm. 9500 Gilman Drive La Jolla CA 92093-0934			10. SPONSORING/MONITORING AGENCY REPORT NUMBER RL-TR-95-123	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Rome Laboratory (IRAP) 32 Hangar Rd Griffiss AFB NY 13441-4114			11. SUPPLEMENTARY NOTES Rome Laboratory Project Engineer: Bernard J. Clarke, Captain, USAF/IRAP/ (315) 330-4581	
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Current secondary storage systems have low transfer rates relative to CPU processor needs. For memory intensive applications, such as data base machines with large on-line storage, this creates a performance bottleneck since the I/O subsystem forces the CPU to wait for data. Optical disks, which are now a commercially available medium, offer high storage densities (100 Mbits per square centimeter), low cost (\$0.1/Mbyte), robustness (no head crash) and, as for other optical memories, the possibility for parallel readout. Still, although high storage densities are possible with optical disk systems, high throughput has not yet been achieved. One solution to processing bottleneck is the Motionless-head Parallel Readout Optical Disk System. This is a hybrid system combining holographic and imaging optical techniques. The underlying principle is that all mechanical motions of the head above the disk surface have been eliminated for addressing focusing and tracking. The Motionless-head Parallel Readout system is designed to generate 2-D binary bit plane outputs, while the recording of the data on the disk remains sequential and can be done using a commercially available drive.				
14. SUBJECT TERMS Associative memory, Optical disk, Parallel addressing, Motionless optical head			15. NUMBER OF PAGES 72	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

Table of Contents

1. Introduction.....	4
2. Motionless-head parallel readout optical disk system.....	5
2.1. Generalities	5
2.2. Disk data encoding.....	6
2.3. Holographic encoding.....	7
2.3.1. Gray-level encoding.....	8
2.3.2. Direct Binary Search algorithm.....	8
2.3.3. New algorithms.....	10
2.3.4. Comparison criteria	11
2.3.5. Reconstruction example	13
2.3.6. Comparisons.....	14
2.4. Optical readout system.....	17
2.5. Illumination element	18
2.6. Disk readout	27
2.7. System modifications for 1-D data patterns.....	28
2.8. System performance and choice of approach (1-D vs. 2-D).....	30
2.8.1. 1-D case	30
2.8.2. 2-D case	30
2.8.3. Choice of approach and comparison	31
3. Disk system experimental results	31
4. Error coding studies	34
4.1. System Model	34
4.2. Parallel partial response precoding.....	35
4.2.1. 1D Parallel partial response	36
4.2.1.1. Theory	36
4.2.1.2. 1D (1+D) PR precoding: an example.....	38
4.2.2. Extension of PR precoding to multiple dimensions	40
4.2.2.1. Theory	40
4.2.2.2. 2D (1+D) PR precoding: an example of multi-dimensional PR precoding ...	42

4.3. Equalizers	43
4.4. Simulation results	47
4.5. Experimental results	48
4.6. Summary.....	49
5. Optoelectronic associative memory	50
6. Optoelectronic associative memory integrated circuit.....	52
6.1. Optoelectronic test chip.....	54
6.2. 8x8 optoelectronic associative memory chip.....	56
6.2.1. Optoelectronic chip functionality	56
6.2.2. Logic level designs and simulations.....	58
6.2.2.1. XNOR level (Level #1).....	58
6.2.2.2. Summation levels.....	58
6.2.2.3. Summation and thresholding level.....	60
6.2.2.4. Generation of the control signals	60
6.2.3. Chip layout and fabrication.....	60
6.2.4. Chip testing	60
7. Summary.....	62
8. References.....	62

1. Introduction

Current secondary storage systems have low transfer rates relative to CPU processing speeds ¹. For memory intensive applications such as data base machines with large on-line storage, this creates a performance bottleneck since the I/O subsystem forces the CPU to wait for data. Solid state disks (such as RAM boards) now reach capacities of up to 1 GByte but can provide sustained throughputs no better than 10 MBytes/sec ². Although projected developments in main memory technologies such as SRAM and DRAM could provide higher throughputs (100 MBytes/sec and higher), their capacity will remain limited for use as secondary storage systems. Large magnetic disk based storage systems such as the Redundant Arrays of Inexpensive Disks (RAID) offer very large storage capacities (100 GBytes and up) but their throughput remains limited to about 50 MBytes/sec ³. Alternatively, optical storage technologies, which combine very high storage densities with a potential for very high throughputs through parallel access, are good candidates for secondary storage systems.

Optical disks, which are now a commercially available medium, offer high storage densities (10^8 bits/cm²), low cost (\$0.1/MByte), robustness (no head crash) and, as for other optical memories, the possibility for parallel readout. Several parallel readout optical disk systems have been proposed in the past. A first approach consists in using a laser diode array in conjunction with a photo-detector array, instead of a single laser diode and a single detector as in commercially available systems, in order to read multiple tracks in parallel ⁴. The main limitations of these systems are their cost and hardware complexity which increase as the number of channels being read is increased. Another type of system stores the desired data plane as a real image on the disk surface. In this case, the data is read using an interferometric imaging optical system and gray level images can be reconstructed ⁵. However, such features as tracking and focusing large images on a rotating disk remain a problem. Another approach stores the 2-D binary data as a Fourier or Fresnel computer generated hologram on the disk surface ⁶. In this case the 2-D bit plane retrieved from the disk is coherent and can be fed directly into a coherent optical processor. Furthermore, due to the properties of Fourier transform holograms, tracking and focusing servo requirements can be eliminated. However, the readout head or the illuminating beam has to be moved or deflected by mechanical or electro-optical means to address different locations on the disk surface. This increases dramatically the hardware requirements and the system cost, and slows down the system speed.

2. Motionless-head parallel readout optical disk system

Another approach to parallel readout optical disk systems, which can be considered hybrid since it combines both holographic and imaging optical techniques, has been developed and its complete design presented in 7. The underlying principle is that all mechanical motions of the head above the disk surface are eliminated for addressing, focusing, and tracking. Therefore, the appropriate coding of the recorded data and the appropriate optical readout system must be developed. The motionless-head parallel readout system is designed to generate 2-D binary bit plane outputs while the recording of the data on the disk remains sequential and can be done using a commercially available drive.

2.1. Generalities

In the following, it is assumed that 5.25" (130 mm) WORM disks with a $1.5 \mu\text{m}$ track pitch and a $1 \mu\text{m}$ pit size are used (see Figure 1). The disk active surface, where the data is actually recorded, is the area between an inner radius of 30 mm and an outer radius of 60 mm, and therefore contains 20,000 concentric tracks. The disk total capacity is then 7.54 Gbits (942 MBytes) per side. Note that the system described in this paper can be modified to match any existing disk size, from 3.5" diameter to 12" and 14" diameter without changing any of the operating principles.

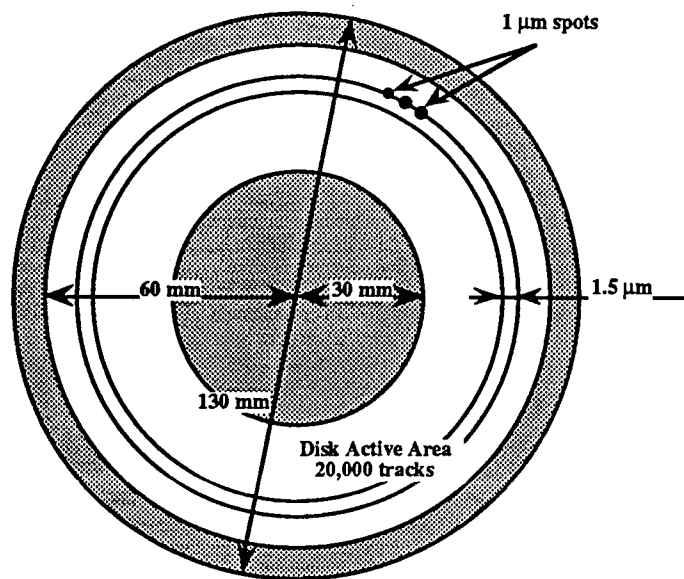


Figure 1: Characteristics of a standard 5.25" optical disk

2.2. Disk data encoding

In order to suppress all the motions of the readout mechanism above the disk for addressing, the data of one given 2-D bit plane is recorded on an area whose length is equal to the radius of the disk active surface. Therefore, data addressing is achieved solely through the disk rotation and the average random access time of the system is half a rotation. The entire memory disk is read in one rotation which yields potential data rates of over 10 Gbits/sec for a 2400 rpm rotation speed. In this case, the data rate (or throughput) is defined as the number of bit planes read-out per second multiplied by the size of these bit planes. As shown in Figure 2, there is one data block for each column of the 2-D bit plane to be stored. Each data block is a 1-D Fourier-transform Computer Generated Hologram calculated to reconstruct one column of the desired 2-D output bit plane. All the data blocks corresponding to one bit plane, are recorded on the disk so that their Fourier direction (i.e. the direction along which the Fourier transform operation will be performed to reconstruct the hologram) is the radial direction. They are also radially distributed to span the entire radius of the disk active surface.

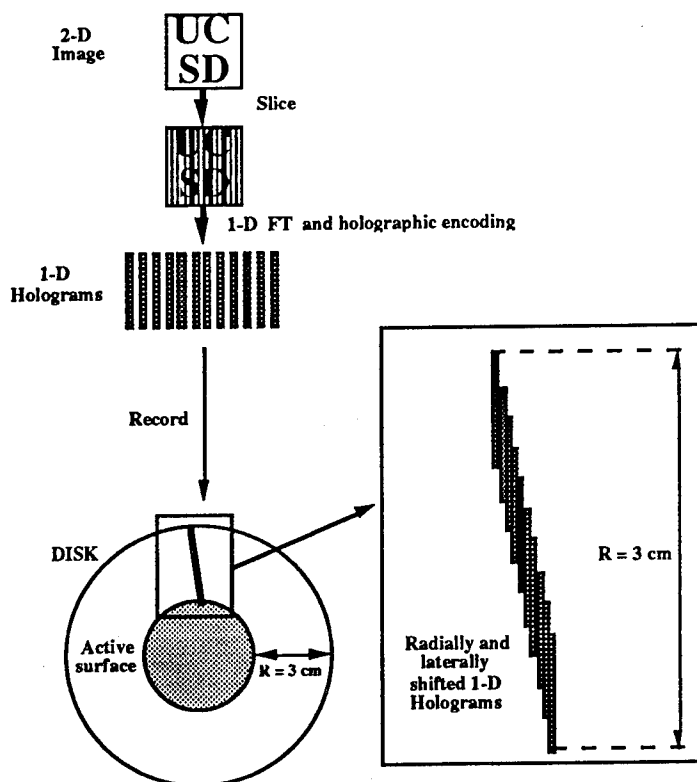


Figure 2: Disk data encoding

The encoding of the disk holograms is a key feature for the good operation of the system. The choice of the right encoding is based on finding the best compromise between different criteria: a short computation time, a high light efficiency to minimize the system power requirements, a high Contrast Ratio to minimize the Bit Error Rate of the system, and a space bandwidth product of the holograms as low as possible to maximize the disk capacity. To satisfy these requirements, a special encoding method based on an iterative optimization algorithm has been developed for the disk holograms.

2.3. Holographic encoding

The data encoding on the disk is a key factor for a good operation of the parallel readout system. The quality of the reconstruction and also the size of the hologram, therefore the capacity of the disk will both depend on the holographic encoding. Thus, the best compromise between the quality of the reconstruction and the disk capacity must be found. Due to the nature of the data recording on a disk, the holograms must be binary. In the application we consider, the reconstructed images have also to be binary. Taking into account all these requirements, a CGH encoding method has been developed specifically for the disk holograms. A comparative study between several encoding methods (1) has shown that the Direct Binary Search algorithm adapted to a method based on a gray level encoding scheme is best adapted to the optical disk application due to its Signal to Noise Ratio and diffraction efficiency.

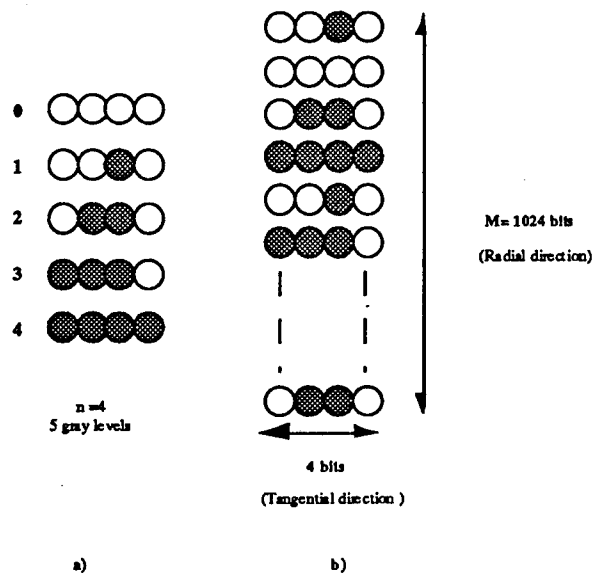


Figure 3: a) hologram gray level encoding method b) example of one data block.

2.3.1. Gray-level encoding

Since the $N \times N$ pixel image to be stored on the disk is sliced in N columns, the study will be focused on the encoding of 1-D images. The hologram is 1-D along the Y direction, therefore the X direction is used to encode $n+1$ gray levels on a n bit pattern using a density modulation algorithm (figure 3 a). In order to reduce the speckle the hologram is replicated once along the Y direction. In this study 128×128 pixel images to be stored are used, the 128 bits of each column are encoded in a 512 cell hologram with $n+1$ gray levels (2 to 5 gray levels). Therefore the hologram after replication is a $n \times 1024$ bit datum (figure 3 b).

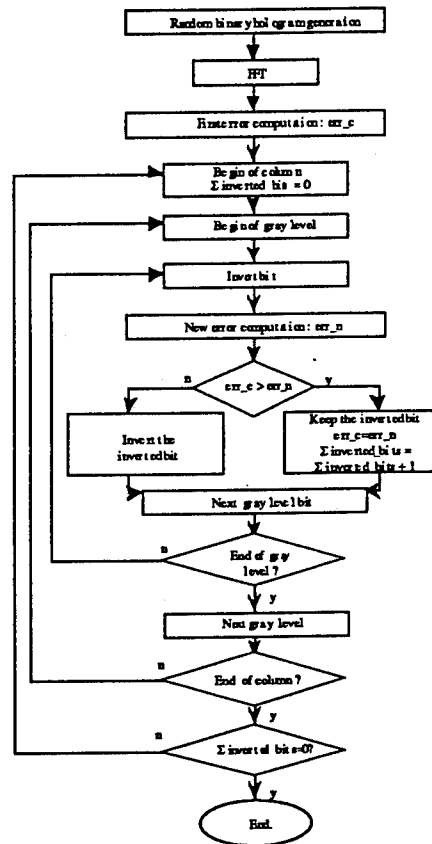


Figure 4: Flow chart of the iterative holographic gray-level encoding method.

2.3.2. Direct Binary Search algorithm

The flow chart of this algorithm is given on figure 4. A random gray level hologram is first generated. The reconstruction of this hologram is then computed by FFT. An error function is calculated by comparing the reconstructed image and the original image to be reconstructed. The

bits of each cell of the hologram are then inverted one after another and a new error is computed each time between the new reconstruction and the original image. If the new error is smaller than the previous one, then the change of the bit is maintained and the new error is memorized; if not the change is ignored. A loop is completed when all the gray level bits of the hologram cells have been tested. The iterative process continues until all the changes are ignored during one complete iteration.

Different types of error between the reconstructed signal and the desired signal can be calculated.

- Amplitude Error

$$\text{ampl.err} = \frac{1}{N} \sum_{k \in \mathbb{R}} (|f(k)| - \lambda |h(k)|)^2$$

with

$$\lambda = \frac{\sum_{k \in \mathbb{R}} |f(k)| |h(k)|}{\sum_{k \in \mathbb{R}} |h(k)|^2}$$

$f(k)$ the sampled desired signal, $h(k)$ the sampled reconstruction of the hologram and \mathbb{R} the region of observation of the reconstruction which is the first diffraction order.

- Intensity error

$$\text{int.err} = \frac{1}{N} \sum_{k \in \mathbb{R}} (|f^2(k)| - \lambda |h^2(k)|)^2$$

with

$$\lambda = \frac{\sum_{k \in \mathbb{R}} |f^2(k)| |h^2(k)|}{\sum_{k \in \mathbb{R}} |h^2(k)|^2}$$

As for the amplitude error, the phase parameter remains free. The convergence is done on an intensity wavefront.

- Complex error

$$\text{cplx.err} = \frac{1}{N} \sum_{k \in R} (f(k) - \lambda \cdot h(k))^2$$

with

$$\lambda = \frac{\sum_{k \in R} f(k) \cdot h^*(k)}{\sum_{k \in R} |h(k)|^2}$$

where * represents the complex conjugate

In this case the convergence is done on a wavefront where both amplitude and phase are respected. The only interesting constraint in the optical disk application is to reconstruct a signal with the desired intensity or amplitude, the phase is a free parameter.

2.3.3. New algorithms

For each of the three error types presented above, one algorithm has been developed. These algorithms offer different mathematical simplifications. The first algorithm (AMPL) is using the error in amplitude (ampl.err) evaluated from the reconstruction which is calculated at each test. But it is not necessary to use an FFT to calculate each reconstruction, since changing one bit of the hologram is equivalent to adding (bit changed from 0 to 1) or subtracting (bit changed from 1 to 0) a plane wave to the previous reconstruction. The second algorithm (INT) is based on the intensity error as presented by Jennison et al (3). With this algorithm, it is possible to evaluate the error without calculating directly the reconstruction. Therefore a gain in computation time is obtained, although the mathematical development of the error is more complicated. The algorithm based on the complex error has the same advantage as the algorithm based on the intensity error but the mathematical development is relatively simple as demonstrated by Jennison et al (4) and moreover the optimization of some variables is possible by using Markov chains (3). Since the convergence is done in amplitude and phase, a diffuser is added to the input signal. The diffuser reduces the spectrum dynamic, easing the encoding and increasing the diffraction efficiency. The nature of the diffuser determines two implementations of the complex error algorithm, one algorithm with a random diffuser (CPLXR) and one with a calculated iterative diffuser (CPLXI).

The calculated diffuser is based on the iterative algorithm of Gerchberg and Saxon (5) (figure 5). The purpose of this algorithm is to obtain a spectrum as flat as possible, this means that the

spectrum intensity Peak Mean Ratio (PRM) is as low as possible. If $F(u)$ is the Fourier transform of $f(k)$, the signal to be recorded, the constraint in the spectrum domain will be:

$$F'(u)^{(i)} = \begin{cases} F(u)^{(i)} & \text{if } |F(u)^{(i)}| < C; \\ C \cdot \exp(j \arg(F(u)^{(i)})) & \text{if } |F(u)^{(i)}| > C \end{cases}$$

i is the iteration index, $j = \sqrt{-1}$, C is the clipping level smaller than $|F(u)|_{\max}$ (positive number). C is fixed in order to obtain a spectral intensity PMR of approximately two.

The constraint in the spectral domain is:

$$f(k)^{(i+1)} = |f(k)| \exp(i \arg(f(k)^{(i)})).$$

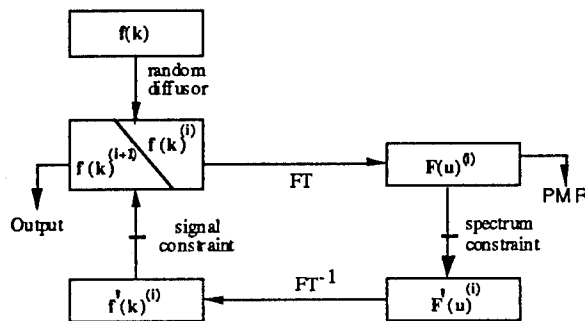


Figure 5: Iterative diffuser algorithm

2.3.4. Comparison criteria

In order to evaluate the hologram reconstruction quality, different criteria have been chosen to compare the reconstruction to the desired image:

- The diffraction efficiency.

It is the ratio between the total intensity of the reconstruction first order and the intensity of the wave illuminating the hologram surface:

$$\eta = \frac{\sum 1. \text{ order reconstruction intensity}}{\sum \text{ total intensity illuminating the hologram}}$$

- The mean-squared errors.

These are the errors presented in the previous section but the signals total intensity is normalized to 1. In this case the complex error is not offering an interesting criterion. The fidelity of the

reconstructed image compared to the original is inversely proportional to the value of the different errors.

- The signal to noise ratio for binary images.

Instead of the mean-squared errors, we have defined an other ratio more suited for binary images (without gray levels). To calculate this criterion, we consider only the samples of the reconstruction +1 order. It is the ratio between the mean of the total intensity of the reconstructed samples corresponding to a "1" in the desired signal, and the mean of the total intensity of the reconstructed samples corresponding to a "0" in the desired signal. This ratio is noted here SNR.

We also calculate the SNR in the worst case (WSNR). It is the ratio between the lowest intensity of a sample corresponding to a "1" of the desired signal and the highest intensity of a sample corresponding to a "0" of the desired signal.

The higher these ratios, the better the reconstructions since in the optical disk system, the signals are received by an array of photodetectors (PN diodes). Therefore the WSNR is the primary criterion, it will determine the system capability to discriminate a "0" from a "1". This, in turn, will drive the Bit Error Rate of the system.

- Ratio standard deviation over the mean of "1"

A fourth criterion can be defined, it is not very important for the system but can bring a better comprehension of the results. It is the ratio between the standard deviation of the "1" and the mean of the "1" of the reconstructed image. It is represented by the following equation:

$$RDM = \frac{\sqrt{\frac{\sum_i^{N1} |h1^2(i)|^2}{N1} - \left(\frac{\sum_i^{N1} |h1^2(i)|}{N1}\right)^2}}{\left(\frac{\sum_i^{N1} |h1^2(i)|}{N1}\right)} \cdot 100 \quad (\text{in } \%)$$

where $h1(i)$ are the "1" samples of the reconstructed signal and $N1$ is their number. In display holography, the reconstructed image will be better if this ratio is low.

2.3.5. Reconstruction example

Figure 6 represents the total field of the simulated reconstruction of a 128 by 128 pixel image. The 128 1-D reconstructions are placed side by side forming a pattern symmetrical with a line consisting of the 128 central orders. The simulation is calculated using the algorithm in amplitude (AMPL) with five gray levels.

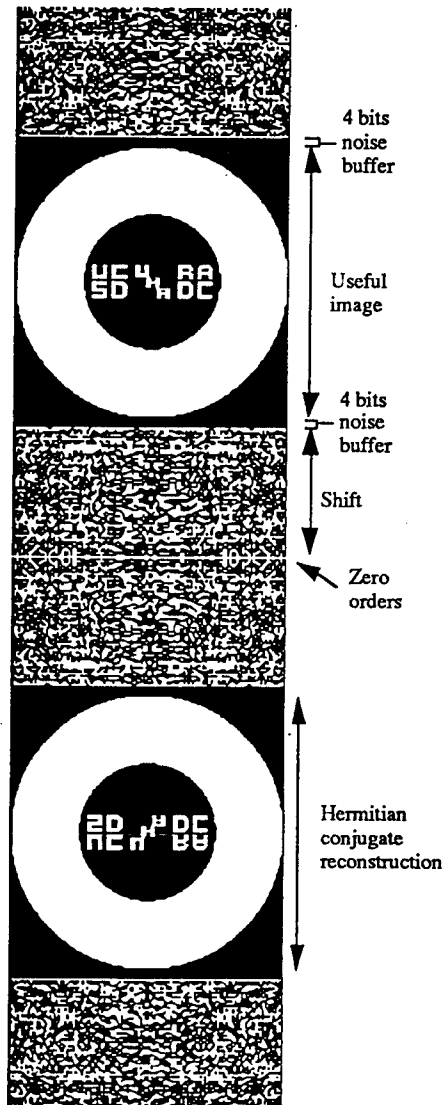


Figure 6: Example of the simulation of a reconstruction (AMPL with 5 gray levels)

The reconstruction region \mathbb{R} of the first diffraction order is defined with a shift of 64 pixels from the central order. Moreover, the top and on the bottom of the desired image are padded with four "0", to keep away the reconstruction noise.

2.3.6. Comparisons

The different tests were processed on a Work Station Sparc Station 1 from Sun. Until now, only holograms calculated with the algorithms AMPL and CPLXR were recorded on a real WORM disk. The results shown in the figures 7, 8, 9, 10, 11, are means computed from 640 different holograms whose reconstruction patterns were randomly defined.

Figure 7 shows that the diffraction efficiencies for the algorithm AMPL and INT are about the same (6.5%). CPLXR has the lowest diffraction efficiency about 5.5% but it is still reasonable. Two reasons can explain the small diffraction efficiency decrease with the increase of gray levels. First, the energy proportion in the central order of the reconstruction is increasing with the number of gray levels. Secondly, the SNR is increasing with the number of gray levels (figure 8), therefore the noise energy on the "0" is lower. CPLXI gives a much better result (7.5 to 8.5%). This is the result of the iterative diffuser which produces a flatter spectrum of the desired signal, and therefore the algorithm is converging easily.

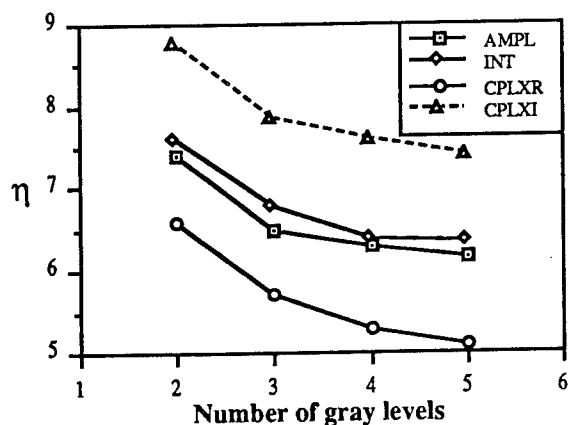


Figure 7: Diffraction efficiency with the different methods

The SNR and WSNR represented on figures 8 and 9 respectively are lower for CPLXR and INT than for AMPL. These differences come from the type of error used for the convergence. With INT the low amplitude noise on a "0" is not eliminated since it is squared and becomes negligible. In the contrary with AMPL the noise is not negligible and it will be eliminated.

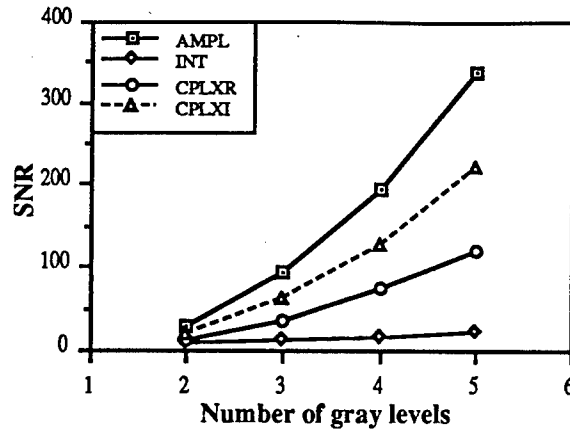


Figure 8: Signal to Noise ratio versus number of gray levels and used method.

The weakness of the SNR and WSNR of CPLXR compared to AMPL can be explained by the fact that with CPLXR there is an additional constraint, the convergence is achieved on both amplitude and random phase. CPLXI gives the best SNR and WSNR after AMPL, due to the low dynamic of the spectrum generated by the iterative diffuser.

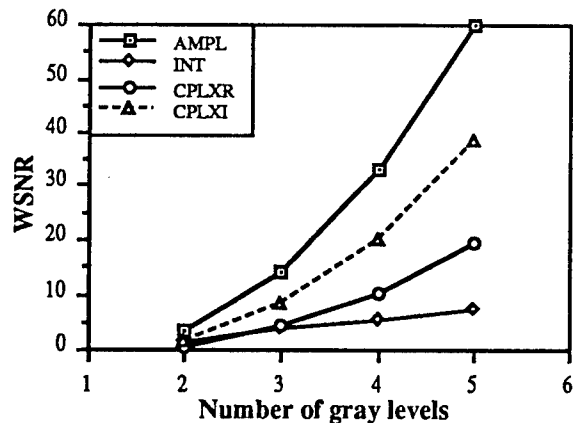


Figure 9: Worst Signal to Noise ratio versus number of gray levels and used method.

Figure 10 represents the different computing times for one 512 x n pixel hologram reconstructing 128 pixels. The computation time for AMPL is ranging from 25 sec. to 55 sec. per hologram, this represents 1 to 3 years computation for an entire disk. Due to the time scale, the CPLXR and CPLXI curves seem flat. In fact for CPLXR the computation time is about 1 sec. for 2 gray levels and 1.8 sec for 5 levels. The computing time for CPLXI is the computing time of the iterative diffuser (independent of the number of gray levels since it is calculated with the input signal) plus

the computing time of the complex error algorithm which is approximately the computing time of CPLXR (random diffuser calculation negligible).

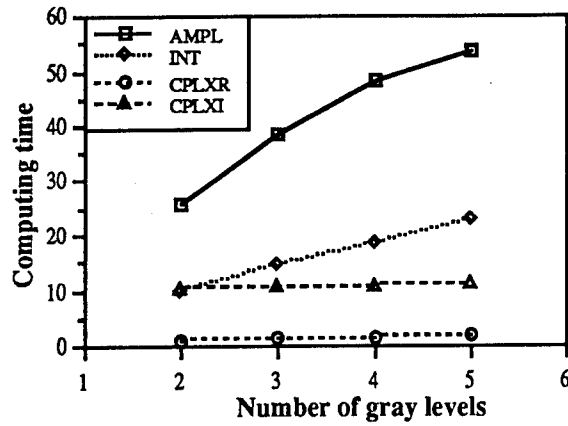


Figure 10: Computing time in seconds by hologram versus gray levels and used method

The RDM (figure 11) gives the "1" intensity variation in the output plane. A small variation is important for good detector array adjustment and operation. For more than three gray levels, all methods give a reasonable RDM lower than 25%. The RDM decrease with the number of gray levels can be explained by a SNR increase. The low standard deviation of the "1" obtained with INT, makes this method best suited for display holography when the image has no dark areas.

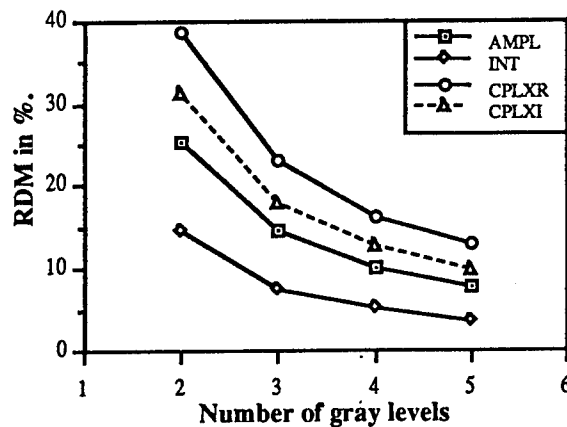


Figure 11: Ratio standard Deviation / Mean of "1"s versus gray levels and the used method.

We have studied the best ways for encoding the binary amplitude holograms which are the information support of our parallel readout optical disk system. We have compared four iterative algorithms based on DBS and adapted for gray level encoding. The first, based on a comparison between the amplitude of the reconstruction and the desired signal, exhibits good performances,

specially a high SNR, but its computing time is prohibitive. A second, for which the signal intensity is the convergence criterion, has a computing time two times shorter, but its SNR is too low to be used in our application. The third algorithm is based on a complex value comparison, a random diffuser being added to the signal. It gives satisfying results and the shortest computation time (25 times less than the amplitude method). Finally, the fourth method is also based on complex value comparison but in this case an iterative computed diffuser is added to the signal. This method gives the best compromise: a relatively short computation time (2.5 to 5 times shorter than the amplitude method), a good SNR and moreover an excellent diffraction efficiency, 1% to 2% higher than the other methods.

This comparative study shows clearly that this last method offers the best results for our application. The different simulations will allow to choose the number of gray levels which will be needed in order to satisfy the SNR required by the detector sensitivity. Note that the most interesting result from this study is clearly the fact that both the capacity of the disk and the quality of the reconstructions can be improved by using these optimized encoding techniques for the disk holograms. In our system that has been built, the actual user's capacity on a 5.25" disk is approximately 30 MBytes or 15,000 planes of 128x128 bits due to the CGH coding overhead. However, as shown in this section, there are certain tradeoffs and compromises that can be found in order to reduce the loss factor and maximize the disk user capacity. In fact, it can be shown that the capacity can be increased by a factor 2.6 for a given SNR by using the above described encoding schemes.

2.4. Optical readout system

The optical readout system described in Figure 12 maps the disk data blocks distribution to a regular 2-D output bit-plane, usually on a photo-detector array. It consists of only three optical elements. First, a collimating lens transforms the laser beam into a circular uniform plane wave. This plane wave is further transformed by a cylindrical lens into a line which illuminates the area on the disk surface that contains all the data blocks of a given bit-plane. As described in the next section, this cylindrical lens can be replaced by a Fresnel hologram for better intensity and phase uniformity as well as reduced aberrations in the illuminated line on the disk surface. As the disk rotates, successive data blocks will come under the illumination line and therefore successive bit planes will be readout. A single custom-designed optical element (a hybrid diffractive-refractive lens ⁷) performs the required Fourier-transform operation along the radial direction. Therefore, all the columns of the output 2D bit plane, which are the reconstructions of the different data blocks,

are reported to the same axis. This is due to the shift invariance property of FT hologram. The same optical element also images the disk surface along the tangential direction so that the columns of the output bit plane are spatially separated. This way, a square 2-D bit plane can be reconstructed on a photo-detector array.

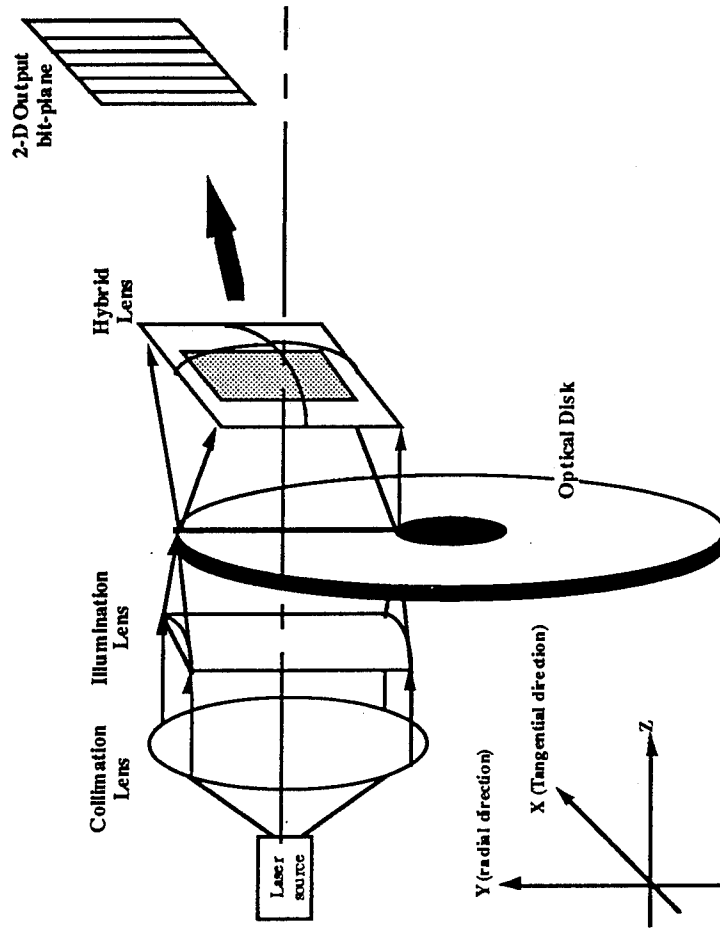


Figure 12: Optical Readout system

2.5. Illumination element

As was described in the final report for contract F30602-91-C-0087, section 2.2, the illumination of the correct data of one given bit pattern on the optical disk is currently achieved by means of a cylindrical lens used slightly out of focus. This technique actually presents severe drawbacks in terms of optical performance due to the inherent poor quality of cylindrical lenses. It also is not very controllable and the positioning of the illumination lens is very difficult. To solve this problem we have proposed to use either a diffractive or a hybrid refractive/diffractive illumination element.

The illumination element has to be used in the motionless-head parallel readout optical disk system that will be later interfaced to the associative memory chip. It has to create a line of 3 cm (radius of the disk active surface) by $30\ \mu\text{m}$ (width of the illumination line) with constant intensity and phase. The constant phase in the illumination line is required in order that the Fourier transform computer generated holograms stored on the disk reconstruct properly. The constant intensity in the illumination line is required so that all the columns of the output image that come from holograms stored in different radial positions reconstruct with a constant intensity.

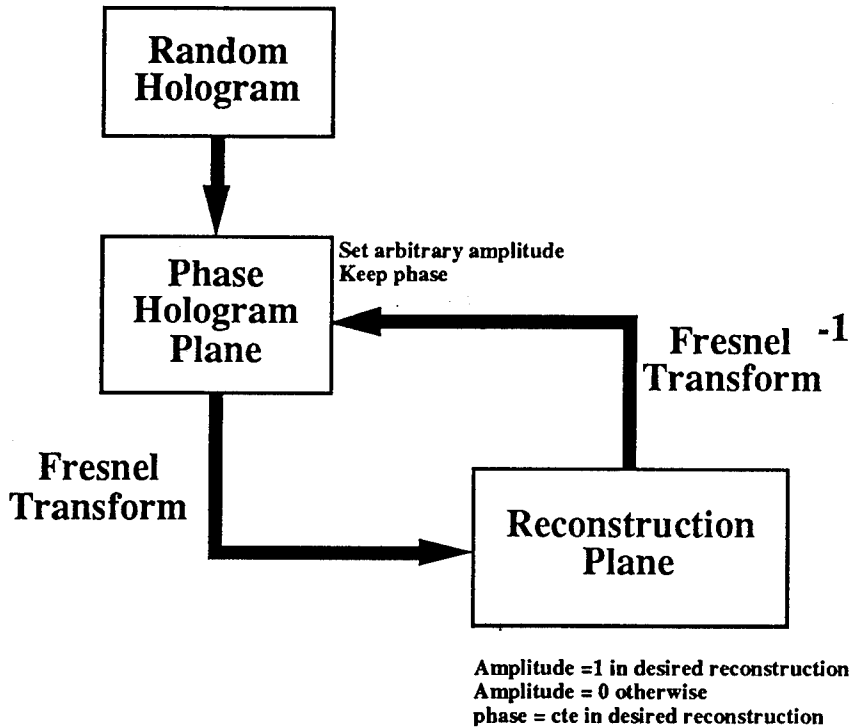


Figure 13: Flow chart of the iterative algorithm

The approach we had initially chosen proved impractical: we had intended to use Code V^[8] (optical system design software) and take advantage of the binary optics option (which allows the design of any arbitrary aspheric diffractive element described by the coefficients of a polynomial), as we had done for the hybrid lens already being used in the system, to generate and compare several designs. However, there is no possibility, in Code V, to control the phase front of the propagating light at a given location. Therefore there is no way to give a phase constraint to the optimization procedure and produce the required illumination element.

To solve the problem we chose a numerical computation approach based on a iterative design known as the Gerchberg Saxon algorithm [9]. The algorithm generates a phase only hologram given a set of input and output phase and amplitude distributions; namely the laser beam profile in the input plane and the desired phase and amplitude distribution in the output plane. As can be seen on figure 13, the algorithm starts by the generation of a random phase mask. This initial phase function is then multiplied by the amplitude and phase distribution of the laser beam to yield a complex wave front in the hologram plane. A Fresnel Transform propagates this wave front to the output plane where the resulting amplitude and/or phase are replaced by the desired ones: a beam front (line) of given length and width with constant phase and intensity. An inverse Fresnel Transform propagates then this new wave front back to the hologram plane. This time, the phase distribution is kept but the amplitude is replaced by that of the laser beam, and another iteration begins. The computation stops after a given number of iterations or after the improvement on the root mean square error between two successive iterations in the reconstruction is smaller than a preset threshold.

Figure 14 through 19 show the simulation results for the illumination element that has been designed using the method described above. The element was designed to create a line of uniform phase and intensity. Figure 14 shows the resulting 16 level phase hologram Figures 15 and 16 show the output intensity and phase. As expected, the intensity and phase are uniform over the desired area. It can be seen that in the X scan (Figure 17), the average intensity reconstructed is 0.95 with a variation of $\pm .05$ meeting the design goals. In the Y direction scan (Figure 18), it can be observed that between pixels 200 and 312 (the reconstruction being centered at pixel 256) the background noise is very low compared to outside these pixels. When a horizontal scan of the phase output (Figure 19) is being taken along pixel #256 it can be observed that the measured phase is constant and centered around 0 with a $\pm \pi/10$ margin meeting the design goals.

This 16 phase level element has been fabricated in our laboratories, however fabrication errors (alignment errors between the successive masks of the multilevel phase hologram) made it so that the element did not perform as expected.

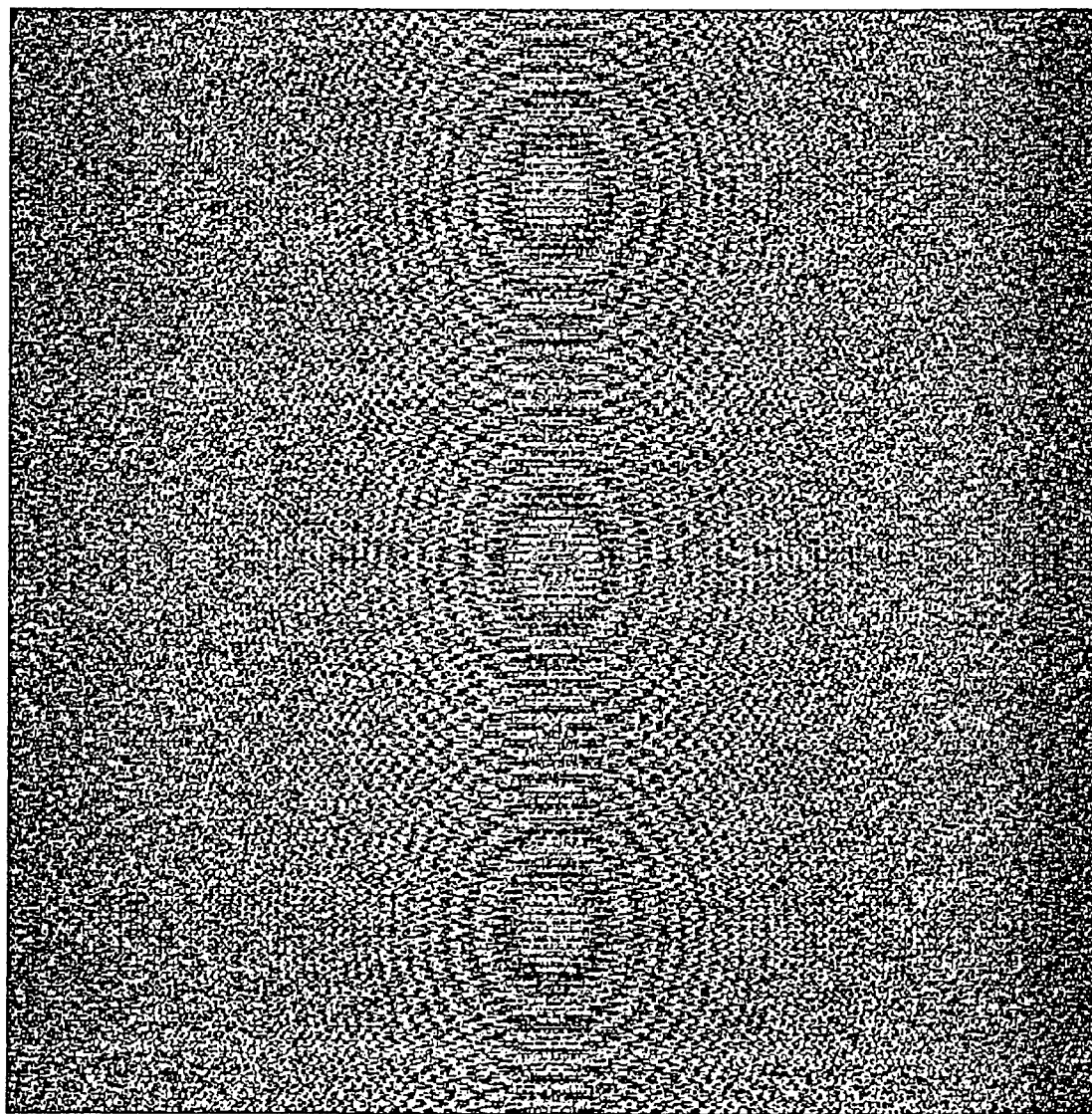
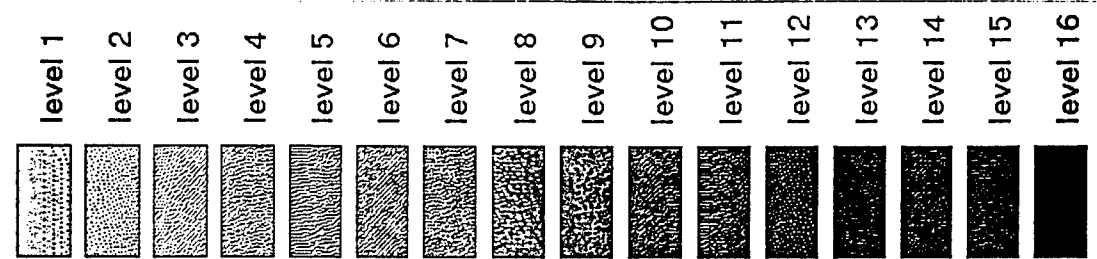


Figure 14: 16 level phase illumination element

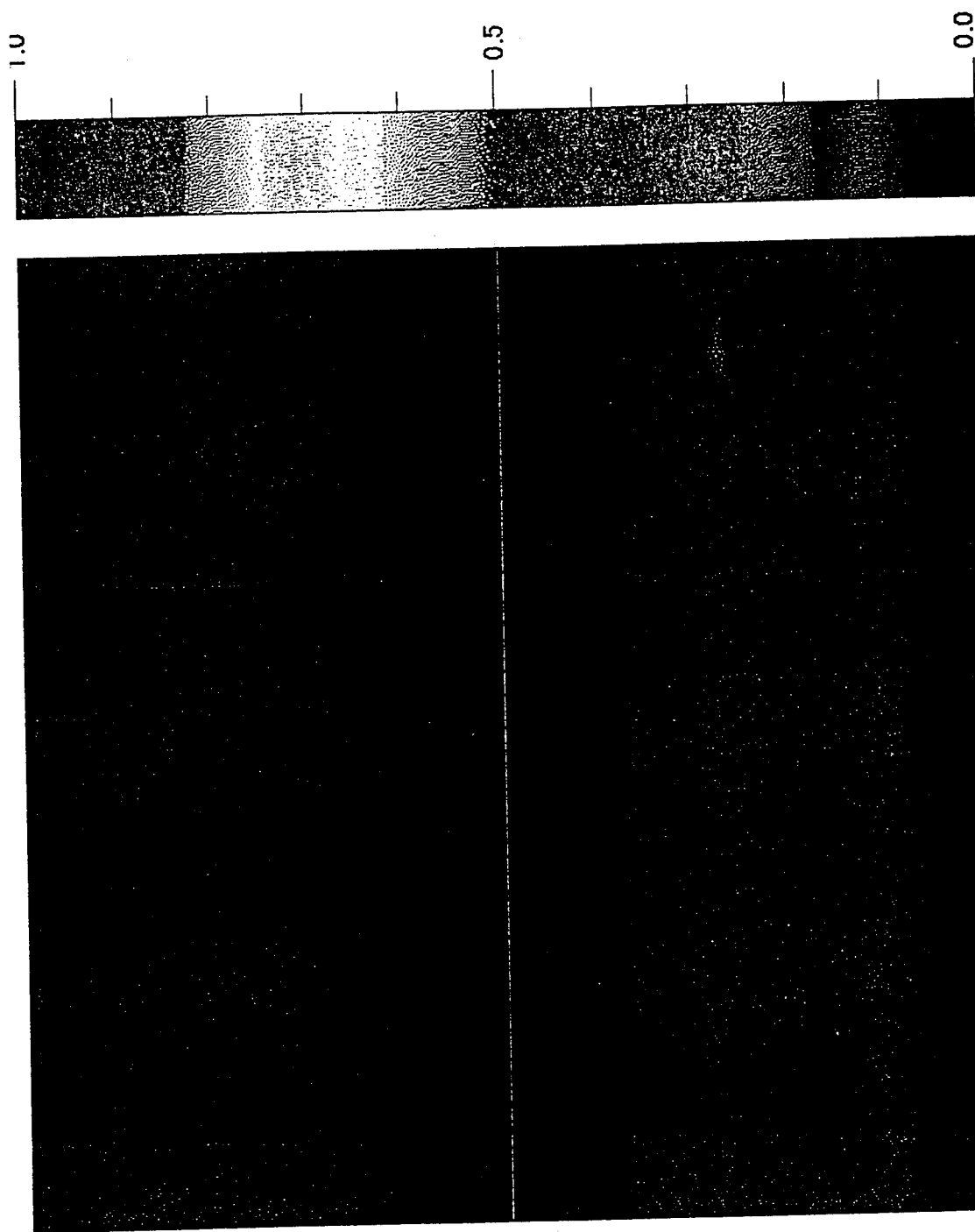


Figure 15: Simulated intensity in the output plane

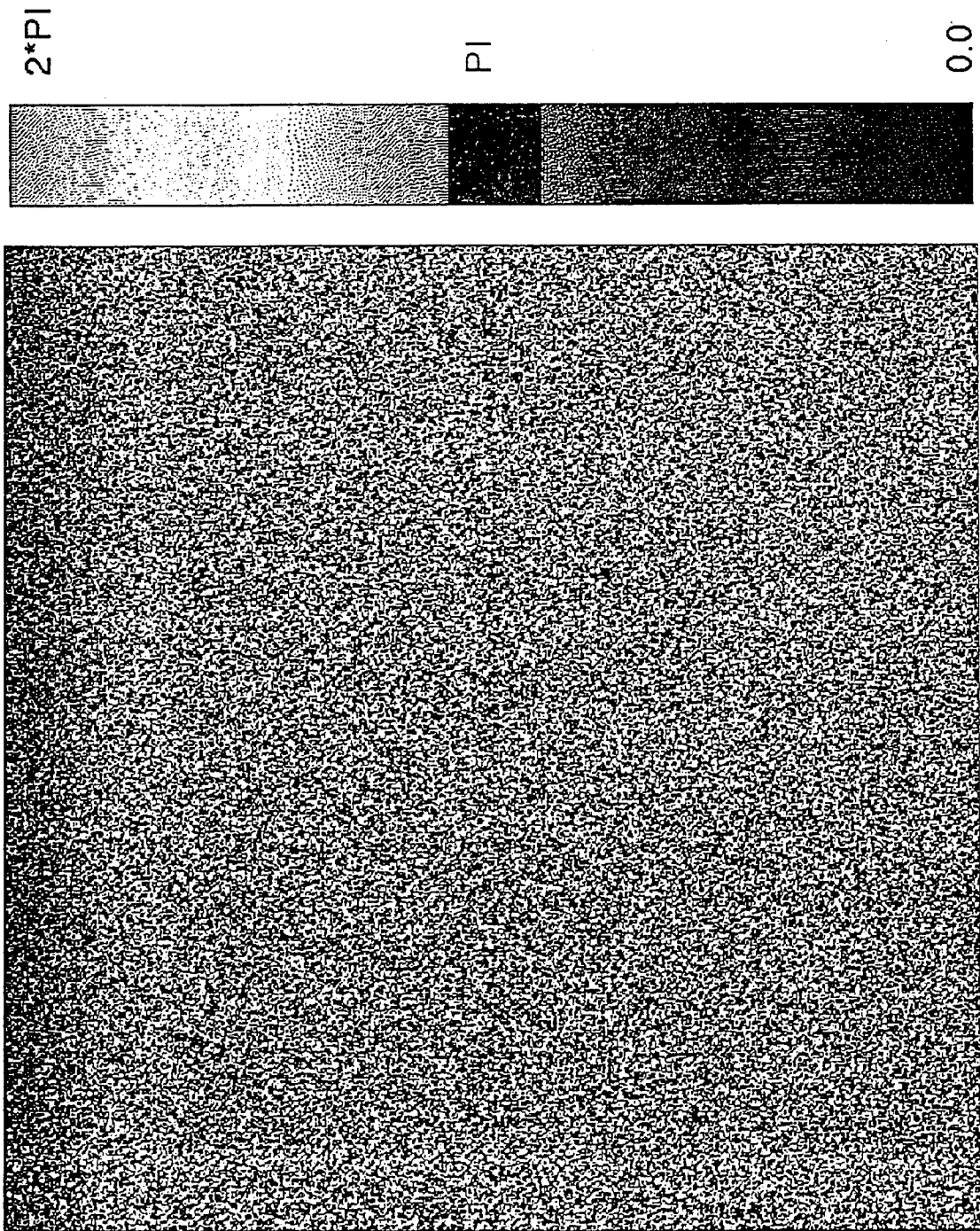


Figure 16: Simulated phase in the output plane

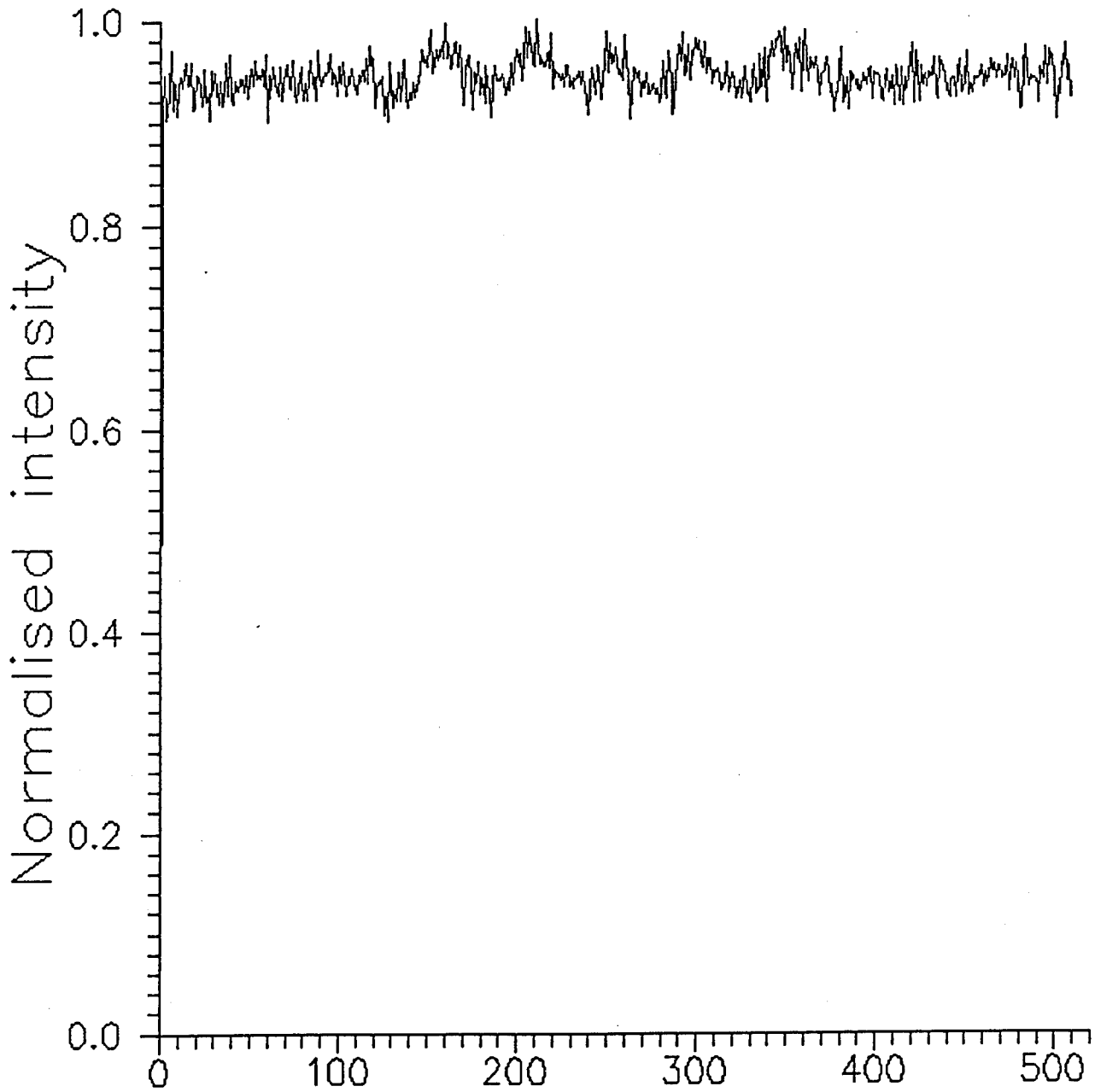


Figure 17: Intensity scan along the reconstructed line

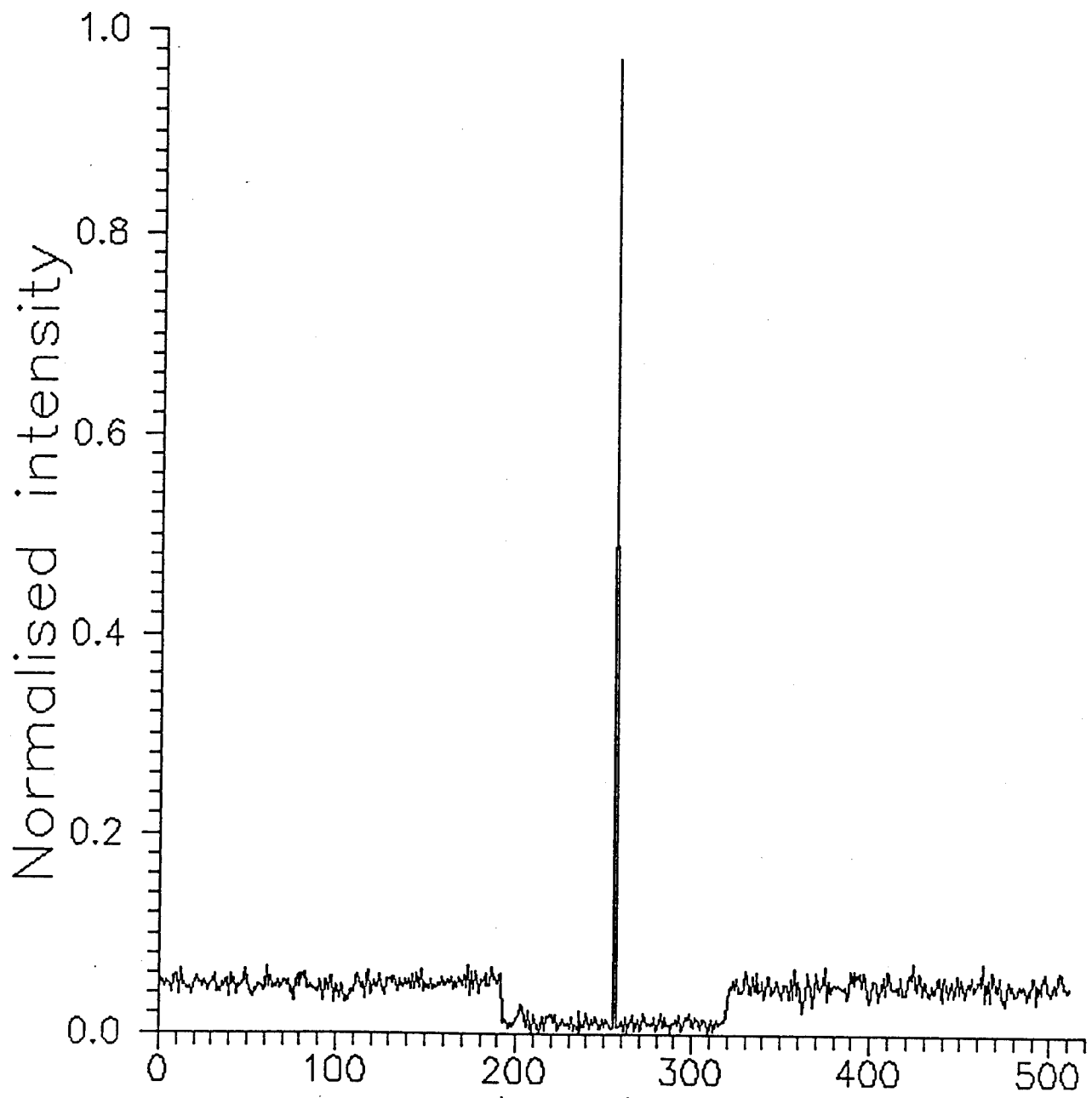


Figure 18: Intensity scan across the reconstructed line

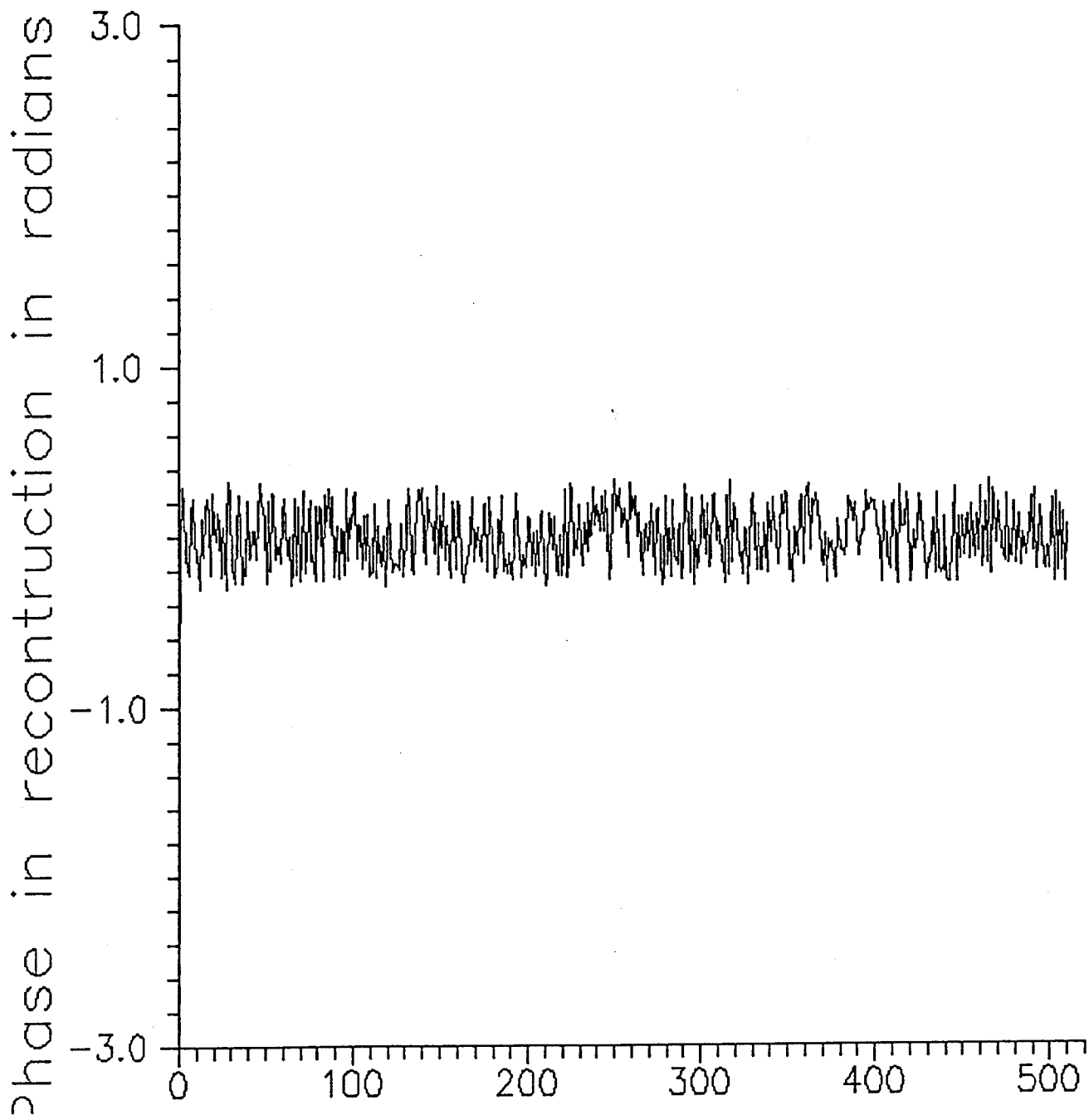


Figure 19: Phase scan along the reconstructed line

2.6. Disk readout

The problems usually encountered with the dynamic operation of disk drives must be solved to ensure the good operation of the readout system. Due to the eccentricity (runout) of the spinning disk and to the surface flatness variations of the disk surface, several parasitic mechanical motions occur. These motions can be decomposed into three independent components: a radial shift of the disk surface, a vertical shift of the disk surface (wobble), and a tilt of the disk surface. In current optical disk readout systems, these motions are corrected by electro-mechanical tracking and focusing servo mechanisms. In our system, the properties of FT holograms prove very useful for overcoming or suppressing them. First, radial shifts of the disk surface occur because of the eccentricity of the rotating disk. However, since FT holograms are shift invariant and the Fourier direction of the disk holograms is radial, the radial shifts will not affect the position or intensity of the reconstructed 2-D bit plane. Second, an optical disk is not a perfectly flat surface. Thus, as the disk rotates, a vertical wobble of the disk surface happens. This will cause the width of the illumination line on the disk surface to vary. This might create crosstalk if the width of the line increases such that the holograms of neighboring bit planes also become illuminated. By taking advantage of the information redundancy property of FT holograms and under-illuminating the area that contains the data blocks of one given bit plane on the disk surface, no crosstalk is created and the holograms are still properly reconstructed, although with less signal. Third, as a disk rotates at high speeds (2400 rpm and above) its edges tend to lift up inducing a tilt of its surface. In the case of a transmissive system like ours, the effect is simply a scaling of the reconstructed data. This effect can be neglected since, for practical systems, the amount of scaling will always be less than $1 \mu\text{m}$ for spot sizes of 30 to 60 μm .

The main advantages of this system are that it suppresses the need for any mechanical motions of the readout head and allows parallel readout of 2-D bit planes. Accessing any bit-plane stored on the disk is achieved solely through the disk rotation and the average random access time to any bit-plane is half a rotation time. In addition, the content of the whole memory disk can be retrieved in only one rotation. Using the same set of numbers as previously (15,000 planes of 128x128 bits per disk) and assuming the disk can rotate at 2400 rpm which is a typical rotation speed for optical disk drives, the average access time is 12.5 msec, the time to read the whole disk is 25 msec, and the data rate is 1.2 GBytes/sec or 600,000 bit-planes/sec. This last figure is particularly important in view of the application in which the disk system will be used where optoelectronic correlations are computed on the disk bit-planes.

2.7. System modifications for 1-D data patterns

The OptoElectronic Associative Memory (OEAM) system as described in this project assumes that the parallel readout optical disk system outputs 2-D bit planes. However, the entire OEAM system design is still valid if the disk output consists of 1-D data patterns. In the case of 1-D outputs, the design and layout of the OEIC is simplified and becomes much cheaper than the 2-D case (see below). Conceptually, there is no difference between the 1-D and 2-D cases for the optical disk readout system, for the OEIC functionality, and for the overall system operation. However, there are a few modifications to be made to the layout of the disk holograms, to the optics of the disk system, and to the layout of the OEIC.

As illustrated in figure 20, the layout of the holograms will be different between the 1-D and 2-D case. In the 2-D output bit-plane case, the N 1-D holograms of a NxN output 2-D bit-plane are laid out radially and shifted from one another until they span the entire radius of the disk active surface. In the 1-D case, a single hologram is recorded for a given output 1-D bit-vector and the radial extent of this hologram is the entire radius of the disk active surface.

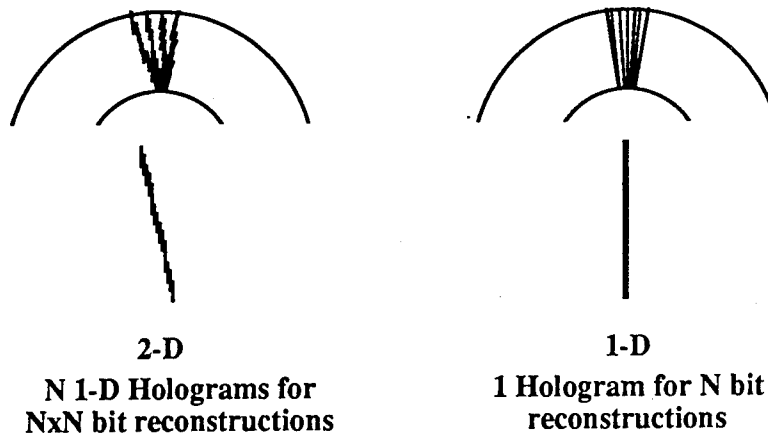


Figure 20: Layout of the 1-D holograms on the optical disk for the 1-D bit-vector and the 2-D bit-planes reconstructions.

In the 2-D output case, the user's capacity (C) of a disk, i.e. the number of NxN bit-planes that can be stored on one disk surface, is:

$$C = \frac{C_T}{KN^2}$$

where C_T is the disk's total raw capacity and where K is the ratio between the space bandwidth product of one 1-D hologram and the space bandwidth product of its corresponding 1-D reconstruction (e.g. one column of the 2-D output bit-plane). If, for example, a hologram space-bandwidth product of 4096 bits is used to reconstruct a 128 bit column ($K=32$), the user's capacity of a 5.25" disk is approximately 14,300 bit-planes of 128 x 128 bits each.

In the 1-D output case, the user's capacity is expressed as:

$$C = \frac{C_T}{KN}$$

where C_T is the disk's total raw capacity and where K is the ratio between the space bandwidth product of the 1-D hologram on the disk and the space bandwidth product of its 1-D bit-vector reconstruction. If a hologram of 35,000 bits ($2 \times 17,500$, where 17,500 is the number of tracks on the disk active surface) is used to reconstruct a 1024 bit-vector ($K=34$), the user's capacity of a 5.25" disk is approximately 190,000 bit-vectors of 1024 bits each. Note that in the two numeric cases given above the values of K are similar; both will lead to the exact same contrast ratios in the output reconstructions. This particular value of K ensures that the system (and more specifically the OEIC detectors) will be able to function at 500 KHz which corresponds to a rotation speed of the disk of 2000 rpms.

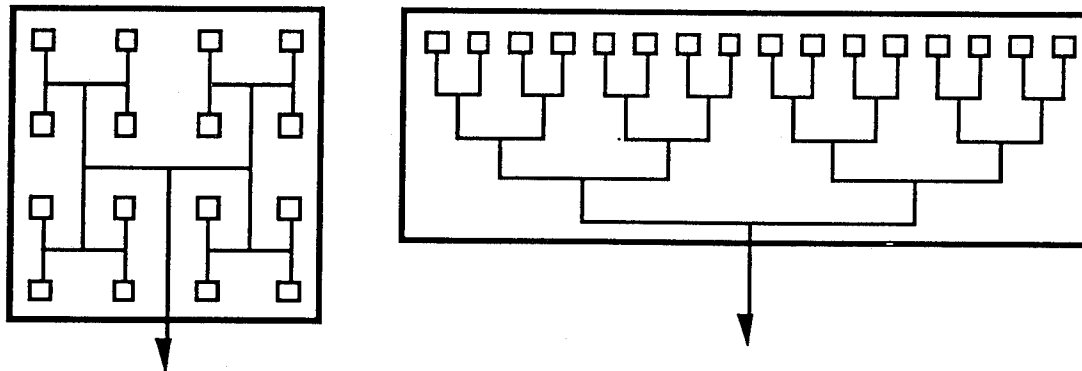


Figure 21: schematics of the OEIC layout in the 2-D H-tree case on the left vs. the 1-D linear tree case on the right. The squares at the leaves of the trees represent the detector inputs to the chip and the arrows represent the numeric output (Hamming distance) of the chip.

The most significant difference in the OEAM is that, in the 1-D case, the OEIC layout will not use a fan in H-Tree structure, but it will use instead a linear fan-in tree with detectors at each leaf of

the tree, as illustrated in figure 21. In both cases, the detector units, XNOR units, fan-in and comparator units will be exactly the same; only their location on the silicon die will be different. Note that in the 1-D case, the VLSI layout operation will be simplified by the fact that all the automatic tools (auto-router, interconnect checking,...) can be used (which is not the case for the 2-D OEIC). Also, the 1-D OEIC fabrication should lead better yields. For all these reasons, the fabrication of 1-D OEICs should be less expensive than 2-D chips.

2.8. System performance and choice of approach (1-D vs. 2-D)

The decision to adopt a 1-D or a 2-D system will be taken depending on what effects the actual sizes of the database, of each record in the database, and of the queries have on the layout of the optoelectronic circuit and the OEAM performance and cost.

For example let us assume the following numbers:

- Query size: 128 bits (note: $128=16 \times 8$)
- Record size: 1024 bits (note: $1024=32 \times 32$)
- The database to be stored is mainly textual (ASCII characters) although some of the data can be arbitrary (digital) if known ahead of time since this influences the query management in the system. The database total size does not exceed 300 Mbits.

2.8.1. 1-D case

For bit-vector reconstructions of 1024 bits and assuming a factor $K=34$, a hologram of 35,000 bits ($2 \times 17,500$, where 17,500 is the number of tracks on the disk active surface) is needed; the user's capacity of a 5.25" disk is then approximately 190,000 bit-vectors of 1024 bits each. This corresponds to a capacity of 24 MBytes. Assuming a rotation speed of 2400 rpms, this leads to a data rate of almost 1 Gbyte/sec (960 Mbytes/sec) for the disk system. In this case, the detector units of the OEIC will have to operate at 10 MHz since successive bit-vectors will be read every 0.13 μ sec.

2.8.2. 2-D case

For bit-plane reconstructions of 32×32 bits and assuming a factor $K=32$, a hologram of 1024 bits (512×2) bits is needed for each column of the bit-plane. The user's capacity is then approximately 200,000 bit-planes of 32×32 bits which corresponds to 26 MBytes. Assuming a

rotation speed of 2400 rpms, this leads to a data rate of a little over 1 Gbyte/sec (1.04 GBytes/sec) for the disk system. In this case, the detector units of the OEIC will have to operate at 10 MHz since successive bit-planes will be read every 0.12 μ sec.

2.8.3. Choice of approach and comparison

According to the number from the 2 previous sections, both 1-D bit-vector and 2-D bit-plane systems are viable. A detailed analysis of the trade-offs will determine which system is the best suited (including system cost factors). However, a first level of analysis leads to the preliminary conclusion that the most economical system is the 1-D system while the most performant (although very slightly) in terms of system speed and capacity, is the 2-D version.

3. Disk system experimental results

A scaled-down version of the system utilizing commercially available plastic substrate 5.25" WORM disks has been implemented. It reads out 2-D bit-planes at a maximum rotation speed of 30 rpm. The disks complete characteristics can be found in ¹⁰. Holograms were recorded on the disks using an OHMT-300 optical disk recorder ¹¹. Figure 22 shows the actual experimental system.

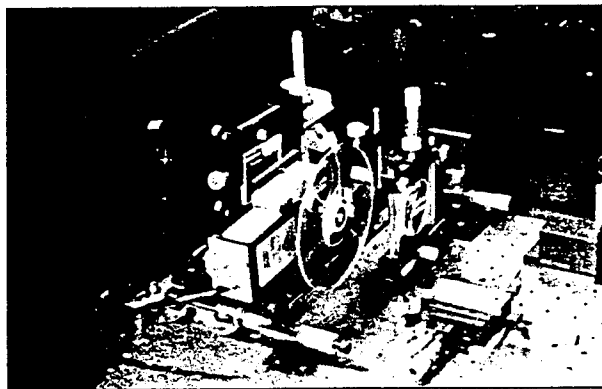


Figure 22: Experimental set-up

The different components of this experimental system include (from left to right on figure 22): the collimation optics, illumination cylindrical lens, the disk and its rotation stage, the hybrid lens, and, in the output plane, a CCD camera. Figure 23 pictures some of the holograms recorded at the

center of the disk active area where the radial shift of adjacent holograms of the same bit-plane can be observed.

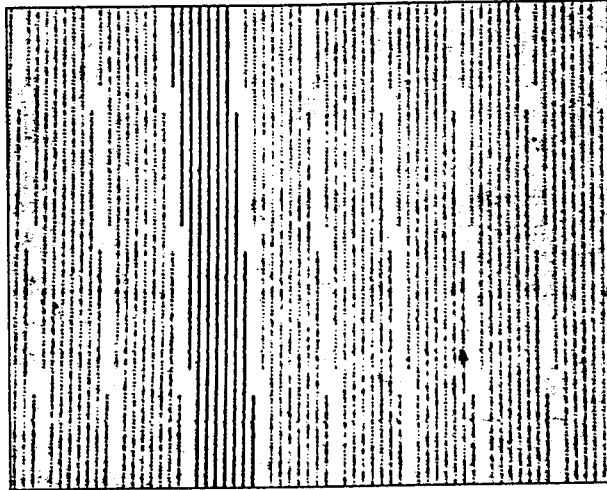


Figure 23: The holograms stored on the disk

Typical output bit-planes can be seen in figure 24 for both 16x16 and 128x128 bit-planes. For all the bit-planes recorded on the disk there is actually an additional bit that is reconstructed outside the 16x16 or 128x128 area. It is used as a clock bit and will allow synchronization of the successive bit plane read from the disk with the system that will process the data in the output. In order to improve the quality of the reconstructed bit-planes, another component was added to this experimental system. A slit, with an estimated aperture of about 100 μm , was placed 3 to 4 mm above the disk surface. This was used to eliminate the optical aberrations and intensity non-uniformities of the illumination line on the disk due to the illumination cylindrical lens. The disk was then rotated at low speeds (up to 30 rpm) and experimental measurements were performed (see figure 25). These measurements were effected by replacing the CCD camera in the output plane with a single photodetector whose spatial location can be varied to look at various bits in the output bit plane. The upper part of the plot in figure 25 shows the optical power vs. time measurements as the detector is located at the clock bit, outside the data field. The lower part of the plot shows the recorded data as the detector is located at the (5,5) bit in a 16x16 bit-plane. It can be seen that with the appropriate threshold (T), the expected 10110001 bit sequence can be detected as the data is synchronized with the clock. At such speeds the behavior of the system was exactly as expected and no measurable effects due to eccentricity or wobble could be detected.

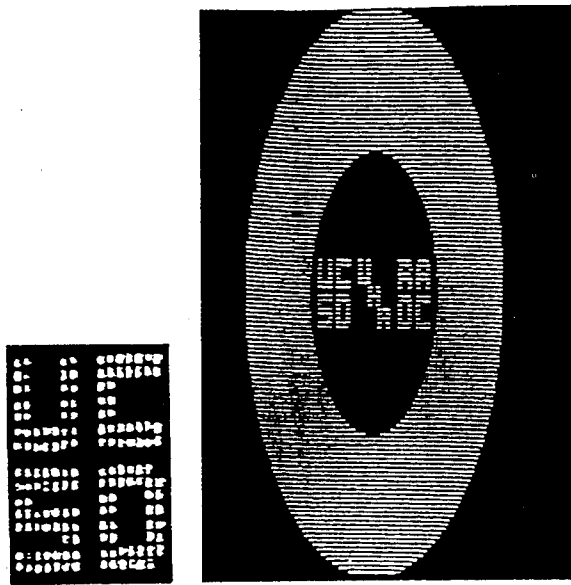


Figure 24: A typical 16x16 and 128x128 output bit-plane

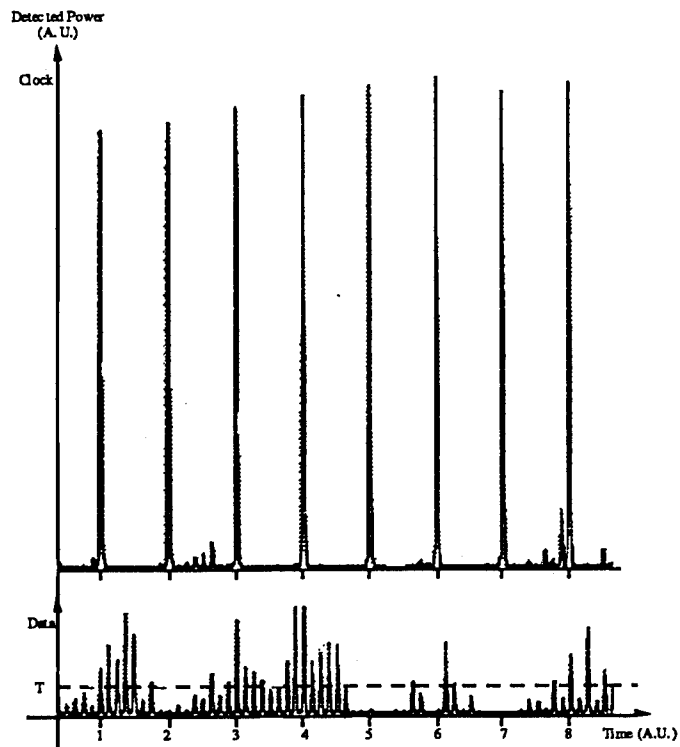


Figure 25: Data signals are detected at the times indicated by the tick marks which correspond to the detected clock signals. A possible binary data threshold value (T) is shown and in this case the detected bit string is 10110001. Note that data and clock measurements are at the same scale.

4. Error coding studies

Pixels retrieved from optical memories are broadened both spatially and temporally in the readout process. In high data density systems with high transfer rates, this broadening causes reconstructed spots to overlap and become difficult to resolve. The concept behind PR precoding is that the amount of broadening that a system causes is, for the most part, known a priori and can be used to determine quite accurately how output spots will interfere in the reconstruction plane. The PR algorithm uses the known broadening factor to precode the data in such a way that when the reconstructed spots overlap in the detection plane, the consequent summing process creates the desired data values. More precisely, it applies $1/P(f)$ (with respect to a particular residue class ring, i.e., using modulo arithmetic) to the desired data, where $P(f)$ is an approximation of the transfer function of the system. Since data values in PR based systems are formed by overlapping spots, these systems can tolerate reconstructed spots with wider main lobes than systems without precoding. This allows them to use optical equalizers, i.e., filters that reduce crosstalk by minimizing the tails of the reconstructed spots at the expense of creating wider main lobes.

4.1. System Model

Figure 26 depicts two equivalent models of the coherent imaging readout system upon which this study is based. In figure 26a the behavior of the optical system is described by $H_o(f_x, f_y)$ which we assume to be $rect[f_x] rect[f_y]$ in the spatial frequency domain. The point-spread-function, $h_o(x, y)$, is the Fourier transform of this, $sinc(x)sinc(y)$. We call the distance from the origin to the first zero of this particular function the half-width of a diffraction limited spot. The input data to the system is described as a two-dimensional (2D) array of abutted rectangular pixels of size $m_x \times m_y$, in units of the half-width of a diffraction limited spot. The reconstruction of such an array would be:

$$\sum_i \sum_j (a_{ij} * \delta(x - im_x) \delta(y - jm_y) * rect[x / m_x] rect[y / m_y]) * h_o(x, y),$$

where $\{a_{ij}\}$ are the values of the individual pixels. In the mathematical derivations that follow, we use $H(f_x, f_y)$, shown in figure 1b. In contrast to $H_o(f_x, f_y)$, $H(f_x, f_y)$ includes a frequency description of the retrieved data pixels, $m_x m_y sinc(m_x f_x) sinc(m_y f_y)$. With this notation, the expression for a single reconstructed pixel (the impulse response of the system), which we call $h(x, y)$, is obtained by taking the inverse Fourier transform of $H(f_x, f_y)$.

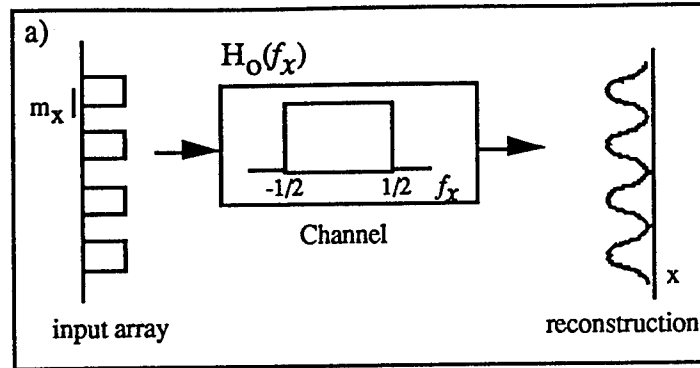


Figure 26 a

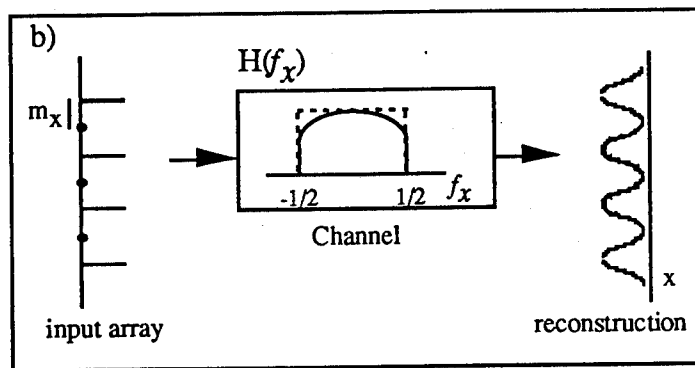


Figure 26 b

Figure 26: Equivalent system models of a coherent imaging readout system. $H(f)$ (figure 26b) includes a frequency description of the storage pixels, whereas $H_o(f)$ (figure 26a) does not. In our analysis, we use $H(f)$ to describe the behavior of the optical system.

4.2. Parallel partial response precoding

For clarity, we first provide a theoretical description of PR precoding applied to one-dimensional arrays, or strips, of data read in parallel from an optical memory, which we term 1D PR precoding. We follow this with a description of the specific form of 1D PR precoding that was applied to the rows of a 2D array of pixels. We then demonstrate how this theory can be extended to multiple dimensions to compensate for the 2D-spatial and temporal broadening that occurs during readout.

4.2.1. 1D Parallel partial response

4.2.1.1. Theory

Figure 27 depicts a hypothetical reconstructed pixel that has been broadened in the readout process. The energy from this spot that falls on neighboring detectors causes intersymbol interference (ISI). If not controlled, it will reduce the signal separation (the difference between a one and a zero) for many bit patterns and hence increase the overall bit error rate of the system.

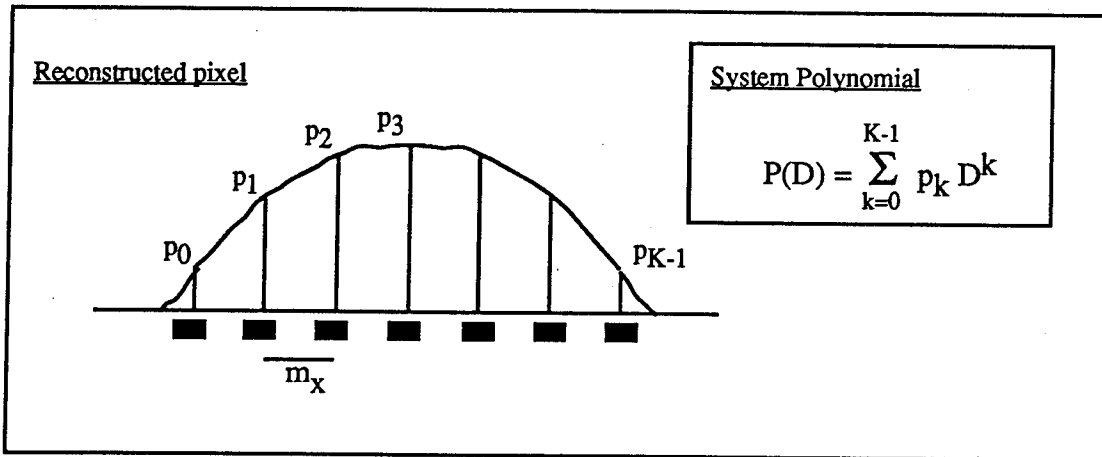


Figure 27: Reconstructed pixel that has been broadened in the readout process such that it has appreciable value over K detectors. This can be thought of as the impulse response of the readout system. The values $\{p_k\}$ are the values that the reconstructed pixel has over each of the K detectors. With PR precoding, the behavior of the readout system is approximated by these sampled values in the system polynomial.

With PR signaling, the behavior of a system is approximated with a series of delta functions whose amplitudes, $\{p_{k\ell}\}$, correspond to the value that a single reconstructed spot, $h(x, y)$, has at the center of each of the detectors that it spans. This is referred to as a system polynomial. With 1D PR signaling, this is described in one-dimension. The system polynomial of the reconstructed spot in figure 27 would be

$$P(D) = \sum_{k=0}^{K-1} p_k D^k,$$

where D is a spatial shift operator. With non-generalized PR precoding, these sampled values are approximated by integers with a greatest common divisor of one.¹² By substituting $exp(-$

$j2\pi f_x m_x$) (the Fourier transform of a shifted delta function) for D in $P(D)$, an approximation of the transfer function of the system, $\hat{H}(f_x)$, can be obtained:

$$\hat{H}(f_x) = \sum_{k=0}^{K-1} p_k \exp(-j2\pi f_x m_x k).$$

m_x is the center-to-center separation between pixels (and likewise detectors). The same D operator can also be used to describe a data pattern, $A(D)$, comprised of input pixels with values $\{a_i\}$.

$$A(D) = \sum_i a_i D^i$$

Using $P(D)$ to model the behavior of the readout process, the reconstruction of $A(D)$ can be approximated in the frequency domain by

$$A(D)P(D) \Big|_{D=\exp(-j2\pi f_x m_x)}.$$

In PR precoding, the operator $1/P(D)$ is applied to the recorded data so that when it is transformed by $P(D)$ in the readout process, the desired data sequence is obtained. More precisely, the inverse filter is applied with respect to a particular residue class ring, i.e., using modulo arithmetic, with a few restrictions (detailed in reference 12). Mathematically, the residue arithmetic makes the realization of the inverse filter, $1/P(D)$, stable. From a systems point of view, applying $[1/P(D)]_{\text{mod}N}$ to a data sequence, as opposed to $1/P(D)$, ensures that the elements in the precoded data sequence (to be stored in the memory device) range from 0 to $N-1$ in value. For example, using modulo 2 arithmetic guarantees a binary data sequence. It also prevents bit errors from propagating.¹³ Because we precode with $[1/P(D)]_{\text{mod}N}$, we need $[P(D)]_{\text{mod}N}$ as opposed to $P(D)$ to be consistent with the modulo N math. The mod N element depicted in figure 28 performs this function. In our implementation, this operation occurs naturally because the input pixels are binary phase valued.

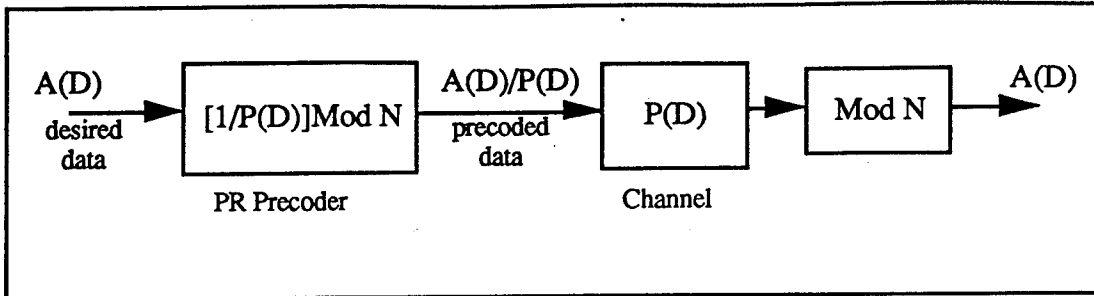


Figure 28: With PR precoding, $[1/P(D)] \text{Mod } N$ is applied to a desired data sequence, $A(D)$, where $P(D)$ is an approximation of the transfer function of the system. Upon readout the desired data is obtained.

4.2.1.2. 1D (1+D) PR precoding: an example

In this study we consider a particular form of parallel 1D PR precoding that we term 1D (1+D) PR precoding. It is used in systems that have impulse responses that can be modeled with a $(1+D_x)$ system polynomial, where reconstructed pixels span approximately two detectors in the x-direction and one in the y direction. Here we apply it to the rows of a 2D array. It has previously been applied to 1D arrays read out in isolation.¹⁴ We further assume that input pixels are abutted and are +1 or -1 in value (which could be accomplished with phase modulation).

When abutted data pixels are broadened by a factor of two, each reconstructed spot overlaps two other spots, one on either side. With 1D (1+D) PR precoding, a data value is formed at the location of each of these overlaps. Figure 29 shows the two different ways that ONE and ZERO values can be created at the detection plane by the coherent interference of two unit amplitude pixels. A ZERO value is formed when a +1 and a -1 output spot overlap and cancel. Similarly, a ONE value is formed by the overlap of two +1 valued or two -1 valued output spots, the two being equivalent in intensity. Detection of these PR precoded signals takes place halfway between the peaks of the two reconstructed spots used to create each data value, where the overlap is most significant.

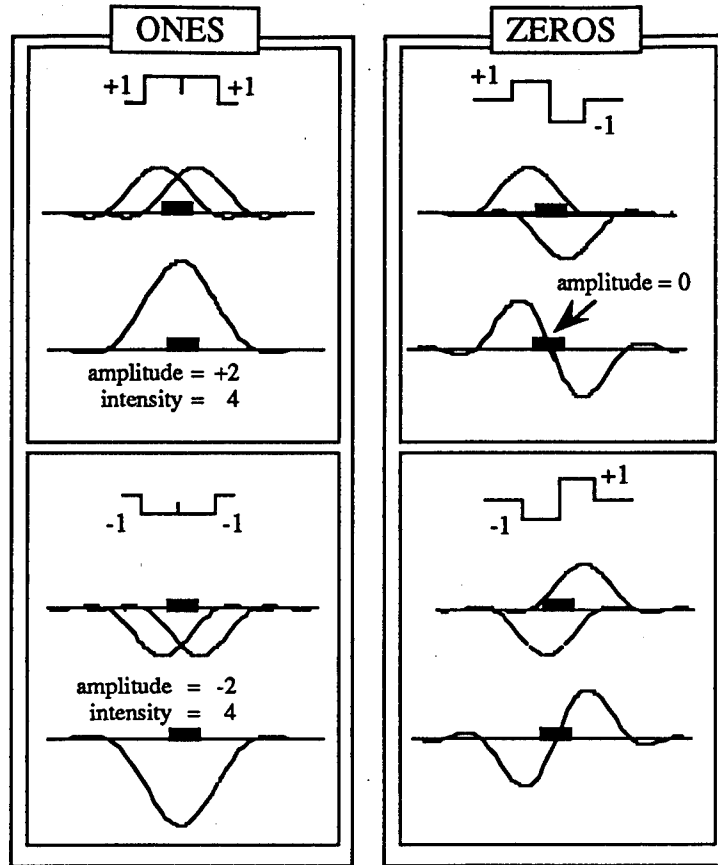


Figure 29: Formation of ONE and ZERO values with 1D (1+D) PR precoding assuming +1 and -1 valued input pixels. Detected values are created by the coherent summing of adjacent overlapping reconstructed pixels. Note that the two complex amplitude values that a ONE can have are equivalent in intensity.

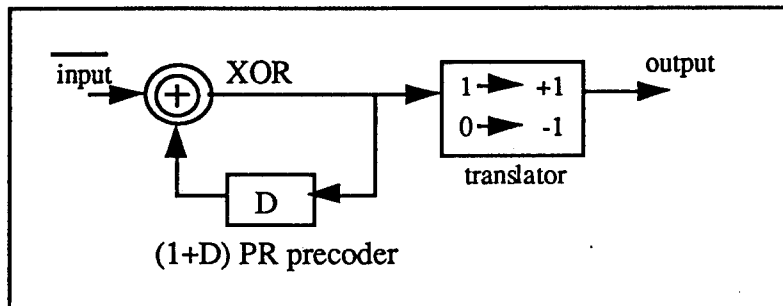


Figure 30: 1D (1+D) PR precoder and translator. The precoder recursively performs $[1/P(D)]_{Mod2}$. The translator eliminates the need for multi-valued detection.

The 1D (1+D) PR precoder (depicted in figure 30) recursively performs $[1/(1+D)]_{Mod2}$. In our implementation, we translate the 0s generated into -1s. A step-by-step description of the overall precoding process is as follows (see figure 31): The first bit in the precoded word is arbitrarily

chosen to be a +1 or -1. This bit is the "extra" bit. Each following bit in the precoded word is determined by looking at the desired output value and the most recently encoded bit. If an output value of ONE is desired, the most recent bit in the precoded sequence is repeated. If the desired output value had instead been ZERO, the precoded value would be opposite in sign to the most recently encoded one, ensuring that the proper cancellation takes place. Note that after the first bit, there is a one-to-one correspondence between desired bits and precoded bits, thus only one additional bit is needed for each row of a 2D array.

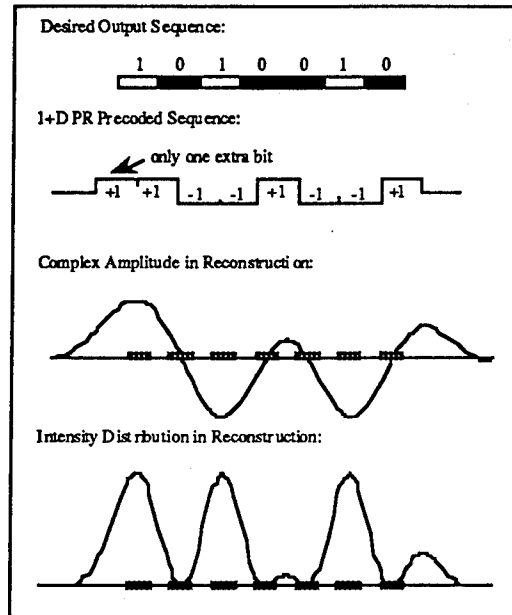


Figure 31: 1D (1+D) PR precoding example. Note that only one extra bit is needed for an arbitrarily long data word.

4.2.2. Extension of PR precoding to multiple dimensions

4.2.2.1. Theory

PR precoding can be extended to multiple dimensions to compensate for 2D spatial and temporal data broadening. Figure 32 shows a reconstructed spot that has been broadened in two-dimensions together with the positions on the spot that would be sampled in the detection process. This broadening might occur in a 2D spatially bandlimited system where 2D arrays of information are read out in parallel. It could also occur in a system where 1D arrays are read out in rapid succession and overlap in time. We introduce two shift operators, D_x and D_y , to describe this 2D

broadening. With this notation, the system polynomial for the reconstructed spot in figure 7 can be written as $(1+D_x)(1+D_y)$.

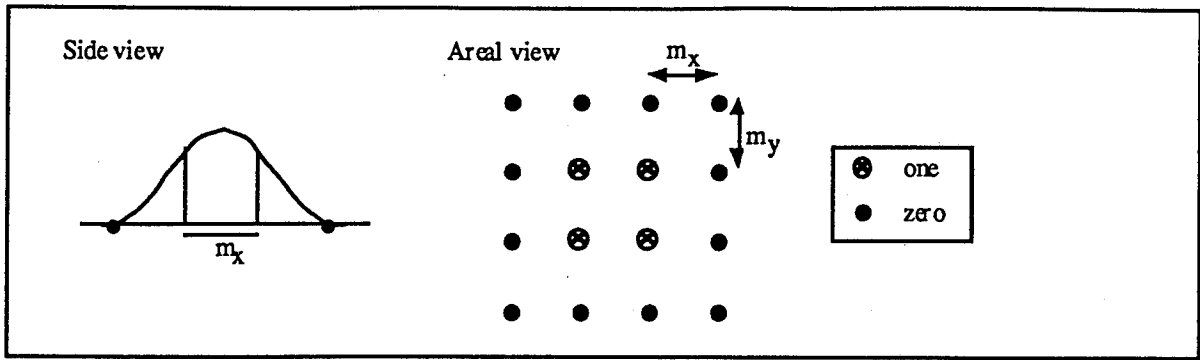


Figure 32: A reconstructed spot that has been broadened in two dimensions. This may be described by a $(1+D_x)(1+D_y)$ system polynomial.

Multi-dimensional PR precoding also involves applying $[1/P(D)]_{Mod N}$ to the desired data. To accomplish this, the 2D array to be precoded can be thought of as a 1D array, precoded as such and then returned to its 2D format. In a system with M pixels in a precoded row (one more than in the original array), such as that shown in figure 33, a 2D system polynomial would be made into a 1D polynomial by substituting D^M for D_y .

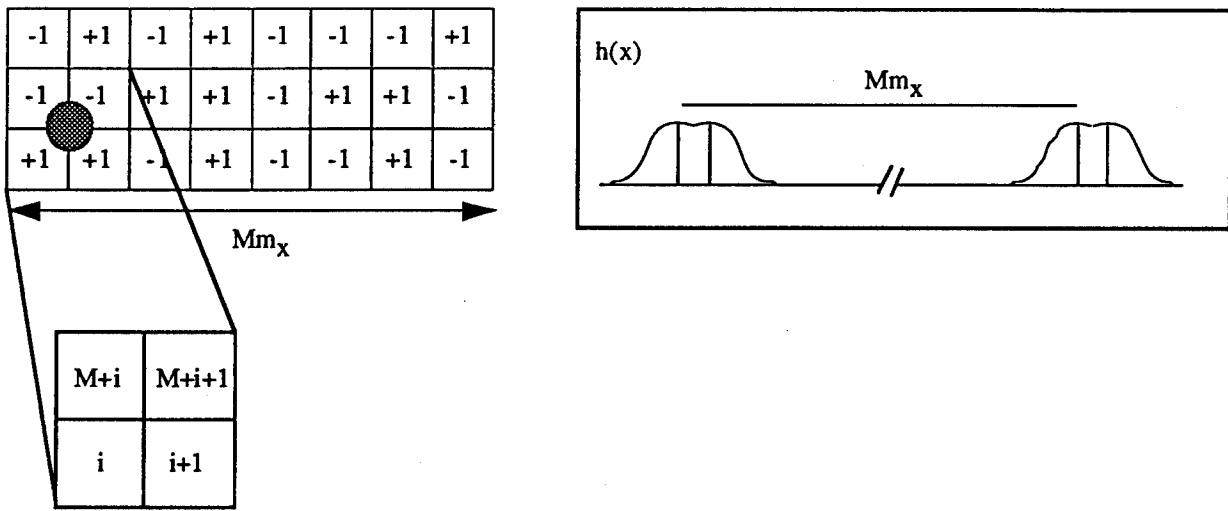


Figure 33: A two-dimensional system polynomial can be converted into a 1D polynomial by substituting D^M for D_y , where M is the number of pixels in the precoded row.

With 3D PR precoding, an additional operator D_z is necessary to express the system polynomial. This system polynomial can also be converted to a 1D representation by substituting D^M for D_y and D^{ML} for D_z , where M is again the number of pixels in a row and L is the number of rows.

4.2.2.2. 2D (1+D) PR precoding: an example of multi-dimensional PR precoding

In this paper, we consider the performance of $(1+D_x)(1+D_y)$ PR precoding applied to a system with binary phase pixels that experiences 2D spatial broadening. We term this 2D (1+D) PR precoding. With this form of precoding, a data value is formed at the center of the overlap of four broadened reconstructed pixels. Figure 9 illustrates the different ways that data values are formed at detection plane, assuming a $[1/((1+D_x)(1+D_y))]_{Mod2}$ precoder. (The four contributing pixels can be in any order.) This particular implementation of 2D PR precoding generates tri-level output values. We refer to these detected values as ZERO, ONE and TWO. Following detection, a TWO would be translated to a ZERO. Mathematically, this translation is the Mod 2 operation mentioned previously and shown in figure 34. Alternatively, multi-valued output could be eliminated by encoding with ± 1 and ± 2 valued storage pixels.

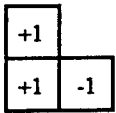
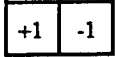
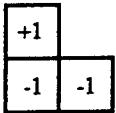
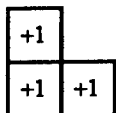
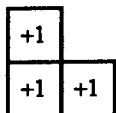
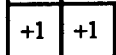
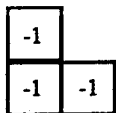
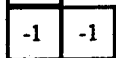
Existing Precoded Bits	Desired Value	Needed Value	Amplitude	Intensity
	ZERO	-1	→ 0	→ 0 "ZERO"
	ONE	+1	→ +2	→ 4 "ONE"
	ZERO	+1	→ 0	→ 0 "ZERO"
	ONE	-1	→ -2	→ 4 "ONE"
	ZERO	+1	→ +4	→ 16 "TWO"
	ONE	-1	→ +2	→ 4 "ONE"
	ZERO	-1	→ -4	→ 16 "TWO"
	ONE	+1	→ -2	→ 4 "ONE"

Figure 34: With 2D (1+D) PR precoding, detected values are created by the coherent summing of four adjacent overlapping reconstructed pixels. Three output levels are generated: ZERO, ONE and TWO. Following detection, a TWO is translated to a ZERO.

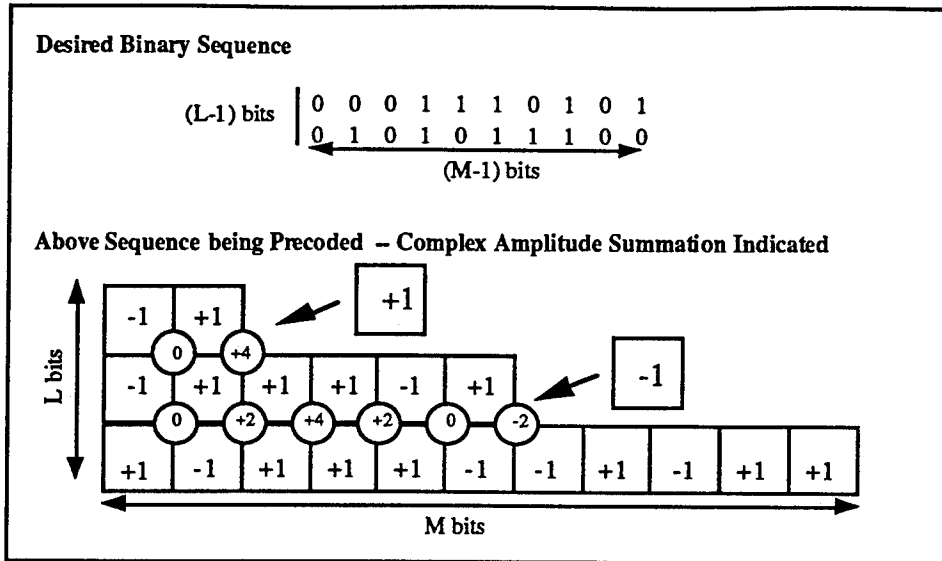


Figure 35: 2D (1+D) PR precoding example. Using the encoding table in figure 34, a 2D array of (L-1) x (M-1) bits is precoded using L x M bits.

To precode, one starts off with two arbitrary orthogonal strips which might be arranged in an "L" or cross shape. These are the "extra" bits. A square array with N^2 bits requires $2N+1$ extra bits. A precoded value is determined by looking at the desired output bit and three already precoded bits, shown in figure 35. This precoding process takes $O(N)$ time if done on a diagonal.

2D PR precoding generates tri-valued output values, while 1D PR precoding produces binary values. Since 2D PR precoding produces an additional output level, one would expect the 2D PR signal separation (in this case, the difference between a ONE and a ZERO or between a TWO and a ONE) to be less than that for 1D PR precoding for the same input energy. This is indeed the case. However, 2D PR precoding obtains more information about the plane being retrieved; for example, certain output patterns, such as 212, are impossible, and this can be used to detect errors.

4.3. Equalizers

The reconstructed spots produced by a coherent imaging system, sampled at the centers of detectors are not exactly described by either a $(I+D_x)$ or a $(I+D_x)(I+D_y)$ system polynomial. While the main lobes of these spots can be described by a $(I+D)$ polynomial, the sampled values of the tails of these spots are non-zero. To improve signal quality, filters, referred to as equalizers, are frequently used in serial communication systems to reshape the overall transfer function of the system and hence the shape of the received data pulse. They can also be used for noise suppression, but this is not considered here. In this study, optimal equalizers for both 1D and 2D

PR precoding are determined by extending the work of Barbosa¹⁵ on the characterization of so-called minimum noise PR channels. The result of this analysis is a frequency description of a 2D zero forcing equalizer (ZFE), which can be realized as an apodizer in the Fourier plane of an optical system.

The overall transfer function of a system can be changed from $H(f_x, f_y)$ to $P(f_x, f_y)$ by using an equalizer described by $G(f_x, f_y)$, where $P(f_x, f_y) = G(f_x, f_y) H(f_x, f_y)$. A reconstructed spot, $p(x, y)$, in such a system would be the Fourier transform of $P(f_x, f_y)$, and the sample values that this spot would have at centers of the detectors in the detection plane would be $p(km_x, \ell m_y)$ for k and $\ell = 0, \pm 1, \pm 2, \dots$. This can be expressed as:

$$p(km_x, \ell m_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(f_x, f_y) e^{-j2\pi(f_x km_x + f_y \ell m_y)} df_x df_y$$

and then rewritten as:

$$p(km_x, \ell m_y) = \int_{-\frac{1}{2m_x}}^{\frac{1}{2m_x}} \int_{-\frac{1}{2m_y}}^{\frac{1}{2m_y}} \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} P\left(f_x - \frac{s}{m_x}, f_y - \frac{t}{m_y}\right) e^{-j2\pi(s km_x + t \ell m_y)} df_x df_y.$$

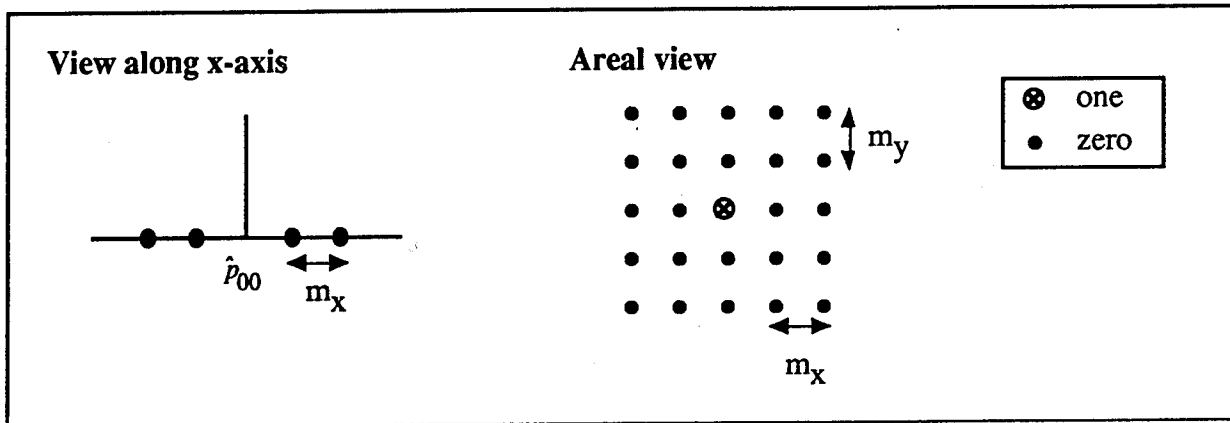


Figure 36a: Portion of a 2D target for a system not using PR precoding.

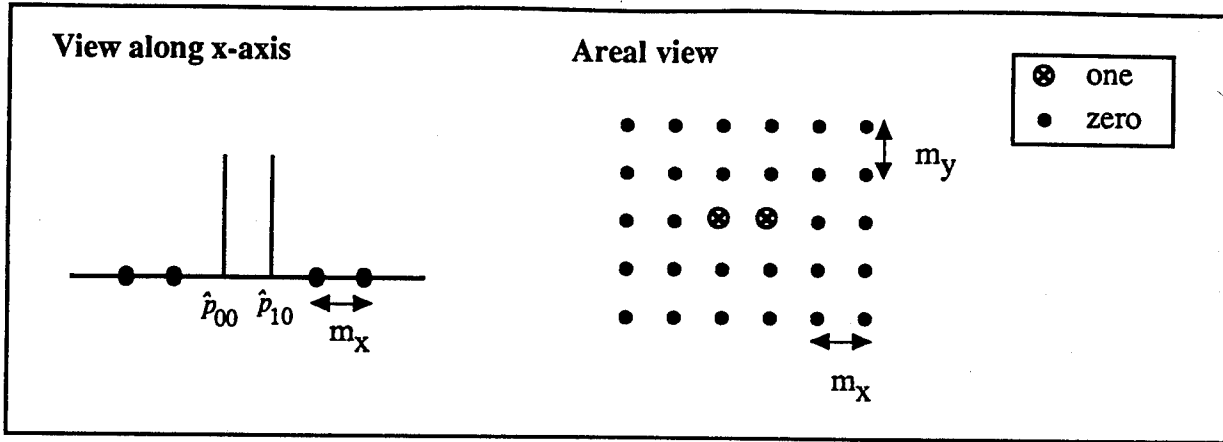


Figure 36b: Portion of a 2D PR target for a system using 1D (1+D) PR precoding. This would be represented with a $(I+D_x)$ system polynomial.

For systems that use PR precoding, as well as systems that do not, there exists a set of ideal sample values $\{\hat{p}_{kl}\}$ that will prevent intersymbol interference (ISI) if one detects by sampling. An array of shifted delta function with these sample values is referred to as a *target*. Portions of targets (which ideally are infinite) are depicted in figure 36 a and b.

In our analysis, we approximate the detection process with sampling and choose $G(f_x, f_y)$ so that the sample values that are formed at the centers of detectors in the reconstruction plane at positions $(km_x, \ell m_y)$ are equal to the set of ideal sample values, $\{\hat{p}_{kl}\}$. For a system to achieve this, the frequency description of the readout system must equal the double summation term in the Fourier integral below.

$$\hat{p}_{kl} = m_x m_y \int_{-\frac{1}{2m_x}}^{\frac{1}{2m_x}} \int_{-\frac{1}{2m_y}}^{\frac{1}{2m_y}} \left(\sum_{q=-\infty}^{\infty} \sum_{r=-\infty}^{\infty} \hat{p}_{qr} e^{+j2\pi(f_x q m_x + f_y r m_y)} \right) e^{-j2\pi(f_x k m_x + f_y \ell m_y)} df_x df_y$$

Equating the last 2 equations yields the condition that $P(f_x, f_y)$, and hence $G(f_x, f_y)$, must satisfy in order to achieve the target described by $\{\hat{p}_{kl}\}$ and preclude ISI.

$$\frac{1}{m_x m_y} \sum_s \sum_t P\left(f_x - \frac{s}{m_x}, f_y - \frac{t}{m_y}\right) = \sum_q \sum_r \hat{p}_{qr} e^{+j2\pi(q m_x f_x + r m_y f_y)} \quad |f_x| \leq \frac{1}{2m_x}; \quad |f_y| \leq \frac{1}{2m_y}$$

In this study, we focus our attention on, *minimum bandwidth channels*, systems where $m_x = m_y = 1$, the spatial Nyquist rate. For minimum bandwidth systems, only one term on the left-hand-

side of this last equation exists, and consequently, there is a unique solution for $P(f_x, f_y)$ and $G(f_x, f_y)$. Table 1 lists the non-zero ideal sample values $\{\hat{p}_{kl}\}$ and corresponding channel functions, $P(f_x, f_y)$, for minimum bandwidth systems using no precoding and PR precoding, as well as the 2D ZFEs, $G(f_x, f_y)$, that would be necessary in coherent imaging readout systems.

Encoding	non-zero \hat{p}_{kl} values	$P(f_x, f_y)$	Equalizer
no precoding	$\hat{p}_{00} = 1$	$rect[f_x]rect[f_y]$	$\frac{rect[f_x]rect[f_y]}{sinc(f_x)sinc(f_y)}$
1D (1+D) Precoding	$\hat{p}_{00} = \hat{p}_{10} = 1$	$cos(\pi f_x)rect[f_y]$	$\frac{cos(\pi f_x)rect[f_y]}{sinc(f_x)sinc(f_y)}$ *
2D (1+D) Precoding	$\hat{p}_{00} = \hat{p}_{10} = \hat{p}_{01} = \hat{p}_{11} = 1$	$cos(\pi f_x)cos(\pi f_y)$	$\frac{cos(\pi f_x)cos(\pi f_y)}{sinc(f_x)sinc(f_y)}$ *

* A phase term has been omitted. This is compensated for by detecting between the centers of reconstructed pixels

Table 1: Targets and their corresponding optimal transfer functions for 2D PR and non-PR minimum bandwidth systems are uniquely determined from equation 9. In coherent imaging systems, the equalizers listed would be used to achieve the transfer functions.

In deriving the ZFEs, it was assumed that detection is accomplished by sampling. In optical systems, however, detection involves integrating intensity over finite areas in the reconstruction plane. We quote a well known theorem frequently used in conjunction with PR signaling (substituting our notation):

If $P(f)$ and its first $K-1$ derivatives are continuous and the K th derivative is discontinuous, $|p(x)|$ decays asymptotically as $1/|x|^{K+1}$.

The optimal transfer function, $P(f_x, f_y)$, which can be generated with a ZFE without precoding has discontinuities at both $|f_x| = 1/2$ and $|f_y| = 1/2$. Thus, while it produces reconstructed spots that are zero valued at the centers of neighboring detectors, the tails of these spots only decay as $1/x$ along the x-axis. In contrast, the optimal transfer functions for PR precoding go to zero in at least one direction, eliminating first order discontinuities at $|f_x| = 1/2$ for 1D PR precoding and additionally at $|f_y| = 1/2$ for 2D PR. Their reconstructed spots, according to the theorem, decay as $1/x^2$ along the x-axis and additionally $1/y^2$ along the y-axis for 2D PR. These decaying terms would suggest better performance in systems with large detection areas, which would be desirable in terms of light efficiency. While the ZFEs for PR based systems reduce tails of reconstructed

spots, they do so at the expense of creating wider main lobes. Since detected values in these systems are formed by the overlap of reconstructed spots, they can tolerate these wider main lobes. Systems not using PR precoding would have to have a reduced data density to use tail reducing filters.

4.4. Simulation results

The performance of 1D and 2D (1+D) PR precoding with and without equalizers was simulated and compared to that obtainable with simple imaging (no precoding). With the PR precoding we use +1 and -1 valued input pixels, whereas with simple imaging we use 1 and 0 valued storage values. We further assume a minimum bandwidth system that retrieves a 2D array of rectangular abutted pixels. We use *worst case signal separation* (SS_{WC}) as a performance metric in our simulations. Memory systems are typically evaluated in terms of their worst case probability of error which for many noise distributions is a function of this. We define SS_{WC} as the difference between the lowest detected ONE and the highest detected ZERO for a given input intensity. The difference between the lowest TWO and the highest ONE for 2D PR precoding was always greater than this and was not used in this metric. The equalizers/apodizers considered in this study were implemented by sampling the equalizer equations (21 times in each direction), quantizing these values (with approximately 200 gray levels) and area encoding this binary amplitude pattern onto a mask. The simulations take into account the quantization of the filters and the loss due to absorption.

Encoding	Equalizer	Simulated SS_{WC} Sampled detection	Simulated SS_{WC} Oversampling for detection
no precoding	no equalizer	.352	2.04
1D (1+D) precoding	no equalizer	< 0	not simulated
1D (1+D) precoding	ZFE equalizer	.444	3.83
2D (1+D) precoding	no equalizer	< 0	not simulated
2D (1+D) precoding	ZFE equalizer	.25	not simulated

Table 2: Simulated SS_{WC} for two-dimensional minimum bandwidth systems with and without equalizers. Without equalizers certain PR precoded data patterns cannot be resolved. Oversampling is used to mimic integration over an area. 1D PR precoding shows improved performance when finite detector size is taken into account.

Table 2 shows the simulated SS_{WC} for simple imaging and the two different forms of precoding with and without ZFEs. In these simulations, the shape of each pixel is described by $rect[x]rect[y]$, in units of the half-width of a diffraction limited spot, and that unit input intensity is used. Detection is accomplished by sampling. All apodizers and apertures are bandlimited with

$|f_x|$ and $|f_y| \leq 1/2$, except for the apertures for the non-equalized PR precoded data. As described in reference 4, a slight reduction in bandwidth in the PR precoding direction increases spot broadening and leads to improved data values.

Without equalizers certain data patterns cannot be resolved with 1D and 2D PR precoding. With an equalizer sampled 1D PR precoding achieves a 26% improvement in SS_{WC} over simple imaging. 2D PR precoding achieves practically an infinite worst case contrast ratio. However, its sampled SS_{WC} is less than that for both simple imaging and equalized 1D PR. This illustrates the limitations of using worst case contrast ratio to evaluate the performance of a recording code. We also ran simulations to determine the performance of equalized 1D PR and simple imaging with larger detectors. In these simulations, we mimicked integration by sampling a reconstructed spot nine times (three times in each direction, with sample points a distance of .2 pixel widths apart) and summing these values. These results are also included in table 2. Close to a factor of two improvement in SS_{WC} is obtained for equalized 1D PR based systems by approximating integration with oversampling.

4.5. Experimental results

The performance of PR precoding, with and without equalizers, was compared to simple imaging (no precoding) experimentally. The test patterns that were used for both the precoded data and the non-precoded data consisted of $42 \times 49 \mu\text{m}^2$ pixels arranged in a 2D array. A binary pattern was surface etched into a glass plate to obtain the +1, -1 values necessary for the PR precoded data, and a binary amplitude pattern was used to obtain the 1 and 0 values for the non-precoded data. The apodizers and apertures, also rectangular, were implemented as a binary pattern on a chrome plate and placed in the Fourier plane of a one-to-one telecentric imaging system. The input data patterns were imaged through the systems onto a CCD camera (with rectangular pixels). In order to study the effects of detector size, the pixels in the input array were chosen to occupy a 5x5 array of pixels on the CCD array. We say a detector is of size .2 when our detection was based on the center value of this 5x5 array, and that it is .6 in size when we used a 3x3 array. Data patterns could not be resolved for any encoding under test when a 5x5 detection area was used.

We express our results in terms of *normalized average worst case signal separation*, $\langle \overline{SS_{WC}} \rangle$. $\langle \overline{SS_{WC}} \rangle$ is the average of the *worst case* experimentally measured ONEs minus the average of the *worst case* measured ZEROs. This value is normalized by the average SS_{WC} that was measured for simple imaging in a minimum bandwidth system with .2 sized detectors. $\langle \overline{SS_{WC}} \rangle$ was

determined by first verifying that the data patterns predicted to yield the worst ZEROs and ONES by simulation also produced the worst data values experimentally. Since there are so many possible data patterns, we were unable to experimentally check all possible data patterns, but did this verification to the best of our ability. These experimental results were then averaged, appropriately subtracted, and the final result was normalized as indicated above.

Encoding	Equalizer	Simulated $\overline{SS_{wc}}$ sampling	Simulated $\overline{SS_{wc}}$ with oversampling	Experimental $\langle \overline{SS_{wc}} \rangle$ detector = .2	Experimental $\langle \overline{SS_{wc}} \rangle$ detector = .6
no precoding	no equalizer	1.00	5.16	$1.00 \pm .06$	$5.39 \pm .51$
1D (1+D) precoding	no equalizer	< 0	not simulated	< 0	< 0
1D (1+D) precoding	ZFE equalizer	1.26	9.69	$1.98 \pm .06$	$12.7 \pm .50$
2D (1+D) precoding	no equalizer	< 0	not simulated	< 0	< 0
2D (1+D) precoding	ZFE equalizer	.710	not simulated	$.423 \pm .05$	$1.74 \pm .44$

Table 3: Normalized experimental $\langle \overline{SS_{wc}} \rangle$ and simulated $\overline{SS_{wc}}$ for two-dimensional minimum bandwidth systems. A factor of two improvement in $\langle \overline{SS_{wc}} \rangle$ is achieved with equalized 1D PR precoding over a system with simple imaging.

Table 3 shows the experimentally measured $\langle \overline{SS_{wc}} \rangle$ values for different detector sizes together with simulated $\overline{SS_{wc}}$ (SS_{wc} normalized by SS_{wc} for sampled minimum bandwidth simple imaging). Experimentally, we measured a factor of two improvement in $\langle \overline{SS_{wc}} \rangle$ for equalized 1D PR. Further gains were made by using larger detectors, since the equalizers used had tail reducing properties. The $\langle \overline{SS_{wc}} \rangle$ values for equalized 1D PR were higher than our simulations predicted. Approximating integration with finer sampling in our simulation might reduce this discrepancy. The results obtained for 2D PR precoding were much lower than we expected. Experimentally, it was quite difficult to align the worst case ONE and ZERO producing patterns to the CCD in this portion of the experiment. We suspect that misalignment could have contributed to these reduced values. When we modeled the behavior of the system, we did not include a model of the CCD readout. This potentially explains the disagreement between our simulated and experimental results. We also did not anti-reflection coat our apodizers, apertures or input planes; Fresnel reflections were noticed which may also have skewed the results.

4.6. Summary

We have presented a precoding methodology for parallel readout optical memory systems that improves signal separation with little coding overhead. We have also developed expressions for 2D zero-forcing-equalizers that can be realized as apodizers in the Fourier plane of PR and non-PR based optical systems. A factor of two improvement in worst case signal separation over a system

with simple imaging was achieved using 1D (1+D) PR precoding with binary phase valued pixels and a 2D optical ZFE . Our current efforts involve applying techniques related to those presented here to parallel readout memory systems with binary amplitude valued input pixels and to incoherent systems as well as considering the effects of noise.

5. Optoelectronic associative memory

There has been considerable interest over the past two decades in the ability to achieve recall of an item stored in a memory device based on a partial query, namely an associative memory ¹⁶. The characteristic that distinguishes an associative memory is its ability to recall the correct output pattern even when the input is distorted or incomplete. Similarly, a Content Addressable Memory (CAM) produces the address of the desired output pattern when an input query is presented. If used in conjunction with a conventional memory device, a CAM also achieves associative recall, since the data can be retrieved in the memory after the CAM computes the address. The key performance measures for a hardware implementation of an associative memory or a CAM are its capacity (number of stored bits) and its search rate (bit-operations/sec). In general, both the capacity and the search rate are limited by the particular technology used for storage and processing and by the choice of the search algorithm.

For purely electronic CAM implementations, there exists a trade-off between the storage capacity and the maximum search rate. While high speed, pipelined VLSI chips (10^9 bit-ops/sec) have been demonstrated ^{17,18,19}, their capacity is severely limited (10-20 Kbit) since all memory resides on-chip. This fact combined with the limited pin-out available on the VLSI chip creates a performance bottleneck for associative memory applications since the serial (or semi-parallel) electronic I/O subsystem forces the processing units to wait for new data from the memory. Here, we describe an optoelectronic associative memory system that removes this limitation and achieves high bandwidth and high capacity by using a parallel accessed optical memory and a custom designed Optoelectronic Integrated Circuit (OEIC) with parallel optical inputs.

The CAM system presently being developed at UCSD ²⁰ (more details on this can be found in the final technical report RL-TR-93-18, March 1993), consists of an optical storage device (e.g. the motionless-head parallel readout optical disk), a photo-detector array, a silicon OEIC with fast local Exclusive-Nor (XNOR) circuitry and summation circuitry (figure 17), and a fast local electronic decision circuit. In this configuration, both the OEIC and the photo-detector array receive a copy of the bit-plane read from the disk since one of the properties of amplitude Fourier Transform

CGHs is that the desired image and its conjugate are reconstructed simultaneously at two different spatial locations (+1 and -1 order). The system operation is based on the page-serial, bit-parallel search method (i.e. it is a template matching network). A 2-D query from the host computer is electronically loaded onto the OEIC. The query is then compared via parallel bitwise XNOR operations to all the binary bit-planes read sequentially from the optical disk. For each bit-plane read from the disk, the outputs of all the XNOR gates are then summed in the OEIC. This sum represents the similarity between the query and the disk bit-plane currently being read; which is a measure of how well they match. It is then fed into the decision circuit which controls the data flow between the photo-detector array and the host computer.

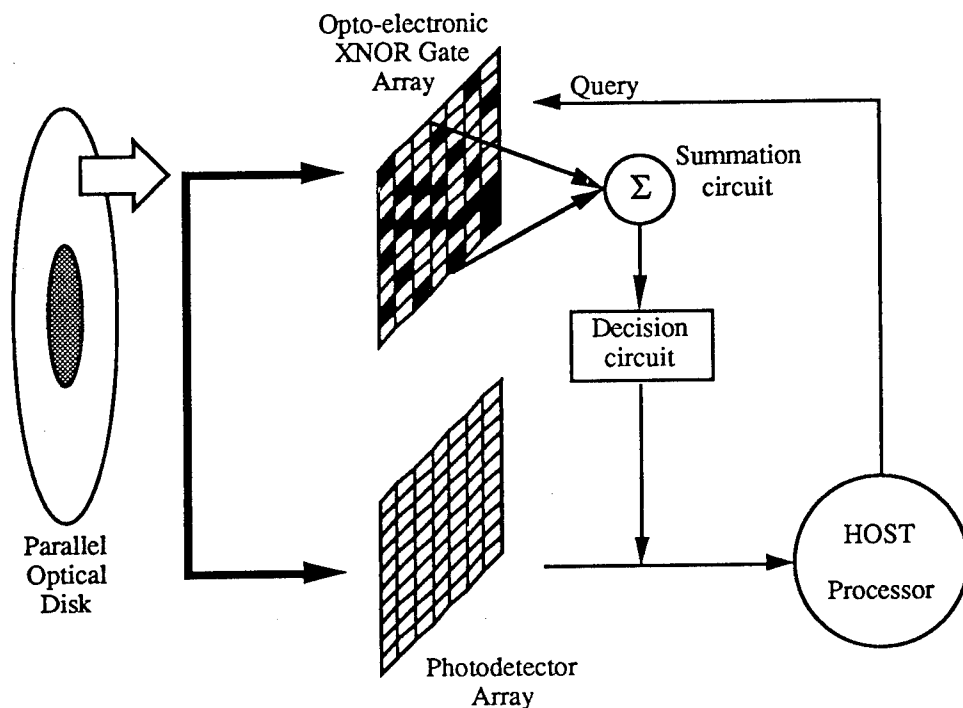


Figure 37: Associative memory system design

When built, the system will support several modes of operation. In the first mode, a threshold value is pre-selected. This is equivalent to asking the system to find all the bit-planes on the disk that satisfy a certain matching criterion. All stored bit-planes for which the Hamming distance with the query is above the threshold; i.e. that are sufficiently close to the query, are then retrieved by the host computer via the photo-detector array. Similarly, the system can also be used to perform classification operations without actually having to retrieve any bit-planes from the disk via the photo-detector array. In this case, the query is compared to all the bit-planes on the disk and the

output is simply a boolean value with an optional convergence accuracy value that states whether or not one or several matches to the query were found on the disk. The last mode of operation finds the best match between the query and the disk bit-planes. On the first rotation, the matching value detected for each bit-plane is input to the decision circuit and the maximum (best) value is stored. The best match is therefore identified and can be retrieved on the subsequent rotation. Since the system cycles through all the bit-planes on the disk sequentially in the order they were stored, it simply keeps track of them by using a counter.

The XNOR logic circuits produce a high output only when a local bit-match occurs, i.e. when one bit from the query and the corresponding bit from the stored bit-plane are identical. The outputs of the XNOR cells need then to be added together in the summation circuit. Both optical and electronic technologies provide solutions to do this. The summation could be done optically by placing a light transmitter at the output of each XNOR cell. A single lens can then be used to collect all the optical signals out of these transmitter and fan them in on a single photodetector. This optical solution potentially provides the fastest systems but has been abandoned because of the requirements on the dynamic range of the fan-in detector and the resulting loss in resolution; where resolution is defined as the smallest number of bits distinguishable between the query and an input bit-plane. Alternatively, an electronic summation circuit can provide very high resolutions (up to one bit) and is the preferred solution since electronic circuits with clock speed of 100 MHz are now available and would not slow down the overall operation of the complete system.

6. Optoelectronic associative memory integrated circuit

The OEIC consists of an array (128 x 128) of unit cells, each having a light detector and local Silicon circuitry that performs the Exclusive-nor function. Each unit cell receives three inputs as well as control information. The query bit is electronically loaded from the host computer. The corresponding bit from the stored binary images arrive from the disk at the detector. The third input is a clock obtained from the disk that signals when a complete image is under observation. For high-speed disk systems, this clock can be distributed to the leaf units with low skew, using an H-tree clock structure²¹. The detector circuits of the optoelectronic XNOR gate array can be designed to provide large noise margins for the detected input bits. Thus the system can operate with images having low contrast ratios. The XNOR logic circuits produce a high output only when a bit-match occurs. The contrast ratios achievable with the disk holograms can therefore be tolerated since each detector circuit restores logic levels. The outputs of the unit cells are then summed electronically using a tree fan-in structure. Each fan-in unit adds the inputs it receives and

sends the result down the tree. The Hamming distance between the query image and a stored image is obtained in $\text{Log}_k N$ levels where k is the fan-in per level. The output of the final stage is fed to a decision/threshold circuit that determines the best matching image. The tree-based operations are performed in an asynchronous manner in order to increase the system throughput. The sum bits of the full-adders at level l are sent down the tree as soon as they are available so that further processing at subsequent levels may begin. Note that more than one image-query result may be in the pipeline at any given moment. The delay of an image-query result is thus $O[\text{Log}N]$ as opposed to $O[\text{Log}2N]$ for bit-serial adders.

Search operations can also be performed on sub-images by incorporating additional thresholding circuitry at the fan-in nodes. The size of the subfield to be searched can be varied by enabling the fan-in units at the appropriate level of the tree to perform thresholding. For example, a threshold can be used in level 3 to recognize a particular 4×4 image, using $k = 4$. In this case the fan-in nodes in earlier stages are performing simple additions. But, if threshold units are used in level 2, it becomes possible to search for a particular 2×2 image. In this case, fan-in units in higher levels (after thresholding) can be used either to count the number of matched sub-images or to find the best matching sub-image by addressing the individual fan-in nodes. Hence, it is possible to dynamically vary the field size of the search operation.

An elementary fan-in unit consists of a simple adder. A fan-in unit at level l consists of a $2(l-1)$ bit adder. If thresholding is required then it becomes necessary to load and store the value of the threshold (θ), and to have a comparator at the fan-in unit. The comparator at the fan-in and root units can also be implemented using a tree-based design. The ability to determine the location of the match is obtained by addressing the individual fan-in units. In addition, the ability to perform a correlation between the disk images and the query is achieved by electronically shifting the disk images between successive XNOR/summation operations. Therefore each fan-in unit is equipped with a counter which increments its value after each successive shift. Note that the number of shifts necessary is proportional to the size of the sub-image and is independent of the overall image size. Matching operations on sub-images are performed concurrently, and the ability to wrap-around is maintained when the image with the multiple copies of the query is kept stationary and the image to be examined is shifted. When the matching image (or sub-image) shifts into place, the threshold circuit produces a high output which then enables the counter to dump its current value into a memory buffer which is then output to a decoder circuit. The address of the active fan-in unit and the contents of the counter uniquely determine the size and position of the match. The number of matches is determined by the number of active fan-in units. It should be noted that the

size (area) of an individual fan-in unit grows (linearly) as one approaches the root of the tree, but their number decreases (exponentially). For instance, for a 128x128 (16K) pixel OEIC, if the minimum field size to be searched is 8x8 pixels, then a total of only 1024 (1K) fan-in units require the added functionality described above. As discussed in the following section, the estimated maximum performance of a full scale 128x128 OEIC is in excess of 10^{11} bit-operations/second, limited by the available optical power at the optical disk.

6.1. Optoelectronic test chip

We have decided to submit first a test chip to verify and validate the design of the prototype. This test chip has one set of all the different functions to be implemented on the chip. This include the leaf unit of the H-tree (with the photodetector, the S-RAM, and the XNOR gate), all the various fan-in units at the different levels of the tree, and the sub-field and full-field comparators. It also includes the circuits for the clock signal detection and broadcast, and the reset lines. The layout of the test chip can be seen in figure 38.

The test chip contains the following circuits:

- Data detector and associated amplifier circuit
- 1 bit SRAM cell (for storing the query bits)
- XOR cell
- One through Eight bit adders
- 6-bit comparator
- 6-bit SRAM cell
- Clock detector and associated clock generation circuit

All the circuits performed logically as expected up to 100 MHz speed. We also measured the photodiode responsivity and the transistor I/V (see figure 39 a and b) curves which are within 10% of the simulations.

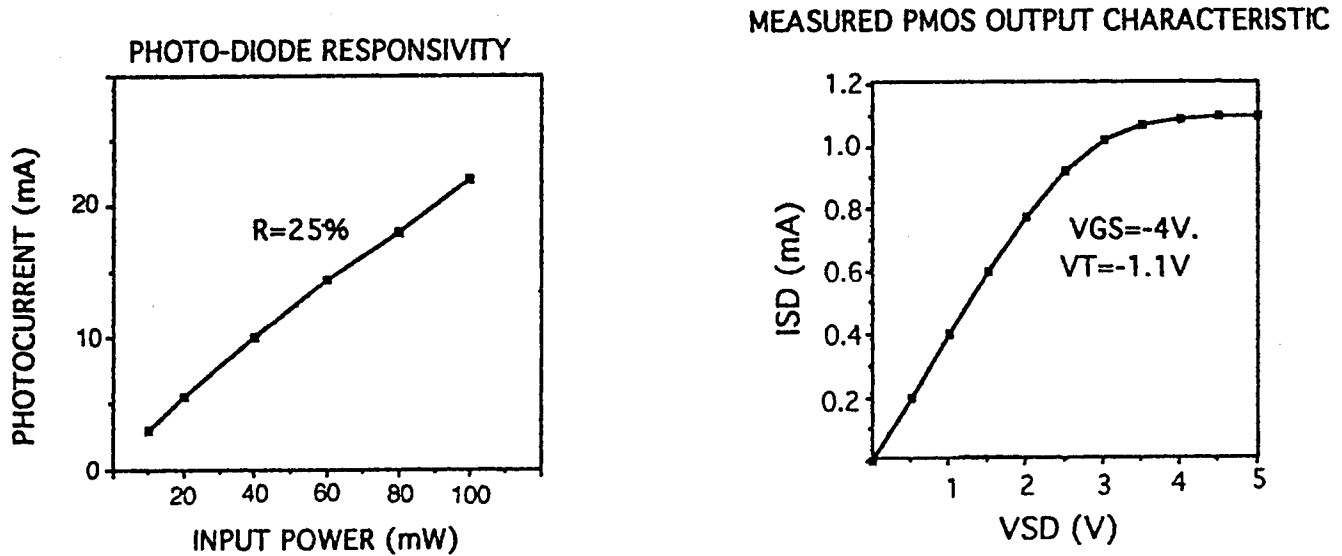


Figure 39: a) Photo-diode responsivity and b) Measured PMOS output characteristic

6.2. 8x8 optoelectronic associative memory chip

6.2.1. Optoelectronic chip functionality

The OE associative memory chip processes 8x8 binary pixel images. The inputs to the chip are :

- Optical memory image coming from the rotating optical disk (data, clock, reset)
- Electronic query image,
- Threshold values for the subfield search regions (explained later in this section).

The outputs from the chip (for each optical memory image) are :

- The degree of matching (full-field) between the electronic query image and the optical memory image (sum of local XNOR's),
- The result of the thresholding operation at each subfield.

Based on output 1, the off-chip processor can :

- 1- Find the addresses of the memory images which satisfy a threshold criteria (by thresholding the sum outputs),

2- Find the address of the memory image that best matches the query (by successively comparing the sum outputs -note that time difference between the construction of images easily allows such an off-chip process-).

Based on output 2, the off-chip processor can :

- 1- Find the addresses of the subfield images which satisfy a subthreshold criteria,
- 2- Find the address of the subfield image that best matches a subfield of the query (by reducing the subfield threshold values starting from the highest).

Initially, the input query image is electronically loaded into the first level units (leaf units) of the H-tree. During the associative search, memory images stored in the optical disk are serially reconstructed onto the OE chip. Each two dimensional optical input received by the chip is converted into a matrix of electronic bits via fast light detectors and XORed bit-parallel with the query image at the first level units of the H-tree (bitwise matching operation). The outputs of these units are then summed by the higher levels (2 through 9) of the binary H-tree and compared to the previously loaded threshold values at the 7th processing level units of the tree (thresholding operation for the 9th level will be implemented off-chip). Thresholding operation included in the 7th level units allows the separation of the associative search into four concurrently working subfields of 8×8 pixels. The results of the four subfield comparisons and the total sum (match value) at the 9th level (root) unit are sent off-chip for further processing.

Proper construction of the memory images on the chip requires fine alignment of the chip with respect to the optical disk. This can be achieved by using 4 light detectors at the 4 corners of the 8x8 chip for alignment. Alignment is ensured when all light detectors produce close enough output intensities.

Note that the signal-to-noise ratio (SNR) of an optical image attains its maximum value at a specific point in time and then decreases as the disk rotates. Therefore, for the full use of the time period between the images with lowest possible optical power, the optical image should be thresholded and latched at its highest SNR point in time. To sense a close enough moment to this point, an additional light detector with a variable threshold is needed: when the intensity of the optical control signal incident on this detector surpasses the threshold, an on-chip synchronization clock signal is generated to latch the optical memory page. If needed, the threshold of this clock detector can be programmed for each memory image .

Furthermore, in order to properly interpret the addresses of the memory images, the host processor must be able to isolate a full rotation of the disk. This necessitates an additional light detector that is optically reset once per rotation.

6.2.2. Logic level designs and simulations

6.2.2.1. XNOR level (Level #1)

In this level, an optical (memory) bit is detected, converted into an electronic bit and XNORed with the locally stored query bit. There is an additional input which is a DC signal called V_{det} to control the detectivity of the light detector. \emptyset is the synchronization clock signal which latches the detector outputs around the highest SNR point (generation of this signal is explained in section 2.4). Since the time difference between the incoming optical images is relatively small ($\geq 1 \mu\text{sec}$), the latched bit can be stored dynamically on the input capacitance of a gate. The query bit, however, needs to be stored during the entire search (period $> 25\text{msec}$) which requires either an SRAM cell or a DRAM cell with periodic refresh. On the other hand, we will see that if the complement of one of the inputs to the XNOR gate is also available, the design of the gate is greatly simplified. This suggests the use of an SRAM cell where both the query bit and its complement are naturally stored.

Note that the detectors of the four XNOR units at the four corners of the chip is also used for alignment. Some extra functionality is needed at these corner units : a control signal S_A connects the detector of the unit to an output pad for off-chip monitoring of the detector output during the alignment. Once the alignment is completed, S_A disconnects the detector from the large capacitance of the off-chip medium for fast operation.

6.2.2.2. Summation levels

In a summation level, the outputs of the previous summation level units are parallel summed. The first summation level consists of a simple half-adder whereas the higher levels necessitate full-adders. We use the conventional half-adder design using an XOR and an AND (NAND+NOT) gate. Based on the simple pass-transistor design of the XOR gate, our full-adder design uses two XOR and three NAND gates. A k^{th} level summation unit consists of one half-adder and $k-2$ full-adders.

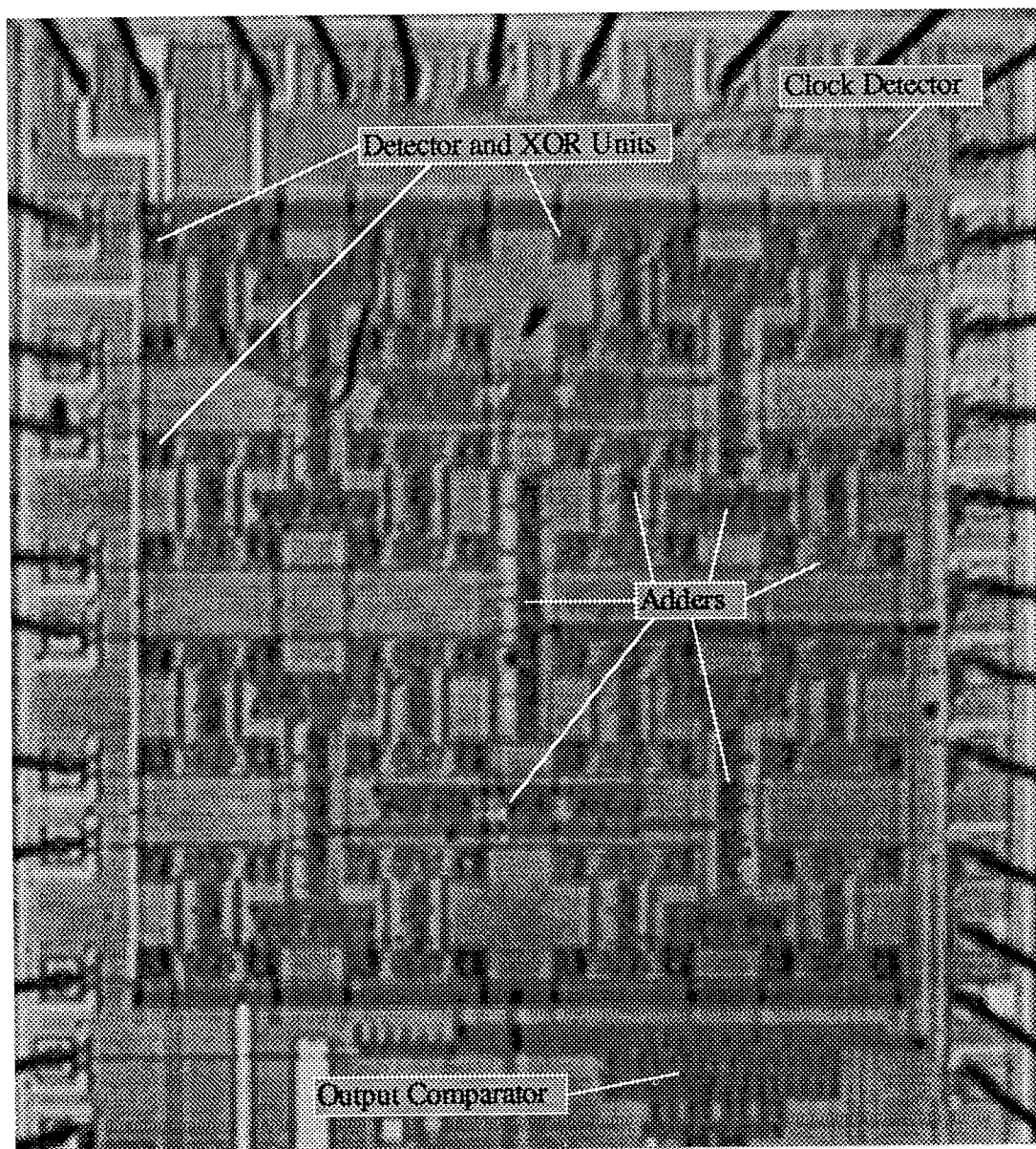


Figure 40: 8x8 associative memory chip

6.2.2.3. *Summation and thresholding level*

In this level units, the outputs of the previous summation level units are parallel summed and compared to a preset threshold value. In addition to the summation logic, this level units contain a 6-bit SRAM memory to store the subfield search threshold value and a comparator to compare the sum of the previous level outputs to the threshold. The logic design of our tree-based comparator unit consists of 20 gates. The design is based solely on inverters and NOR gates.

6.2.2.4. *Generation of the control signals*

The optical clock bit $OP\emptyset$, which is high for each memory image (except the first one on the disk), is detected and restored by the light detector L1. Similarly, the optical reset bit $OP\emptyset_{reset}$ is generated and restored by the light detector L2 at the beginning (and therefore the end) of each full disk rotation. A DC signal V_{clkdet} is used to set the threshold of these detectors independent of the XNOR-level detectors. A separate control signal, V_{extclk} , is used to electronically switch the light detectors for testing purposes. The outputs of these two light detectors are ORed to generate the on-chip latching signal \emptyset , and sent to the off-chip processor for synchronization.

6.2.3. Chip layout and fabrication

Due to the requirement of equal spacing between light detectors and compact H-tree fan-in, we custom laid-out our OE associative memory chip. CADENCE layout tools were used. Figure 40 shows the picture of the layout of the entire chip. Several blocks of the processor including XNOR units, fan-in/summation units and tree-based comparator units are clearly observed. The H-tree fan-in as well as a subfield search region are also illustrated. On the root unit, we have shown a typical half-adder and a full-adder. The 6-bit memory and the tree-based comparator are seen. The chip has been fabricated through ORBIT using their 1.2 micron n-well process.

6.2.4. Chip testing

The chip has been fabricated by ORBIT and received by us. We started the testing by optically testing the various optical inputs available on the chip. The set-up used for the testing is shown in figure 41. A single spot generated from a Laser diode that can be modulated at speeds up to 10 Mhz is created on the chip. The spot can be moved around the chip in order to individually test the detectors and their associated circuits.

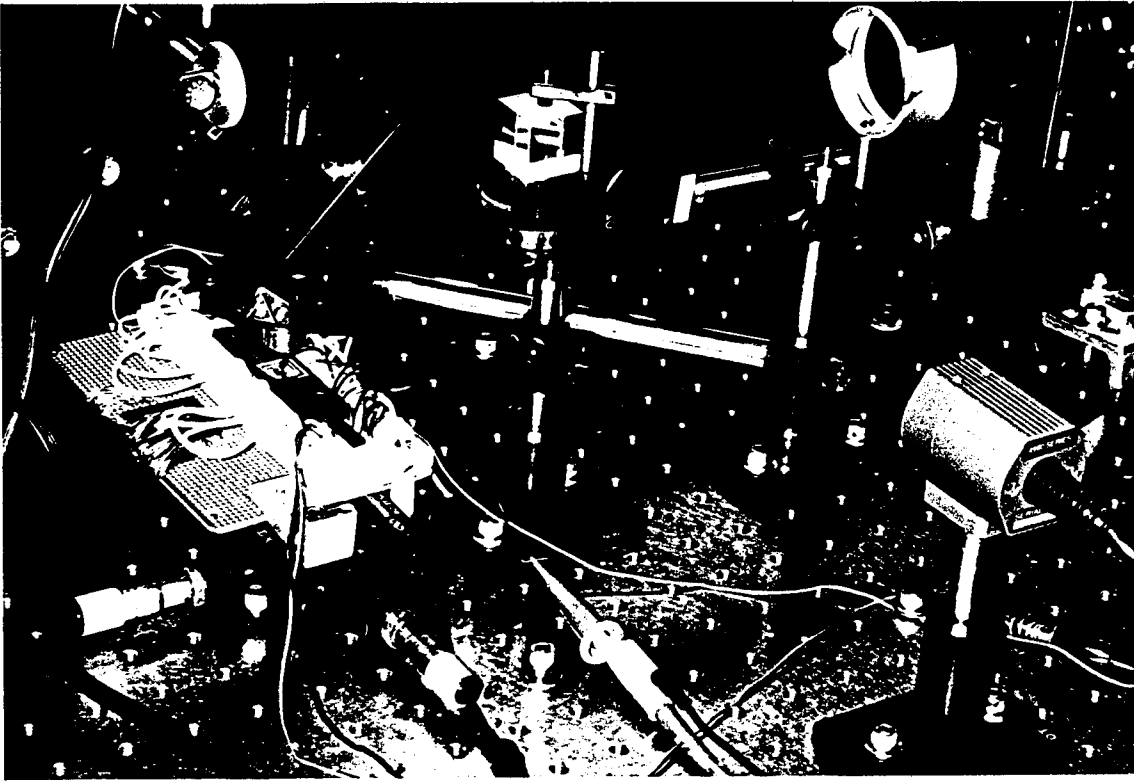


Figure 41: Test set-up for the associative memory chip

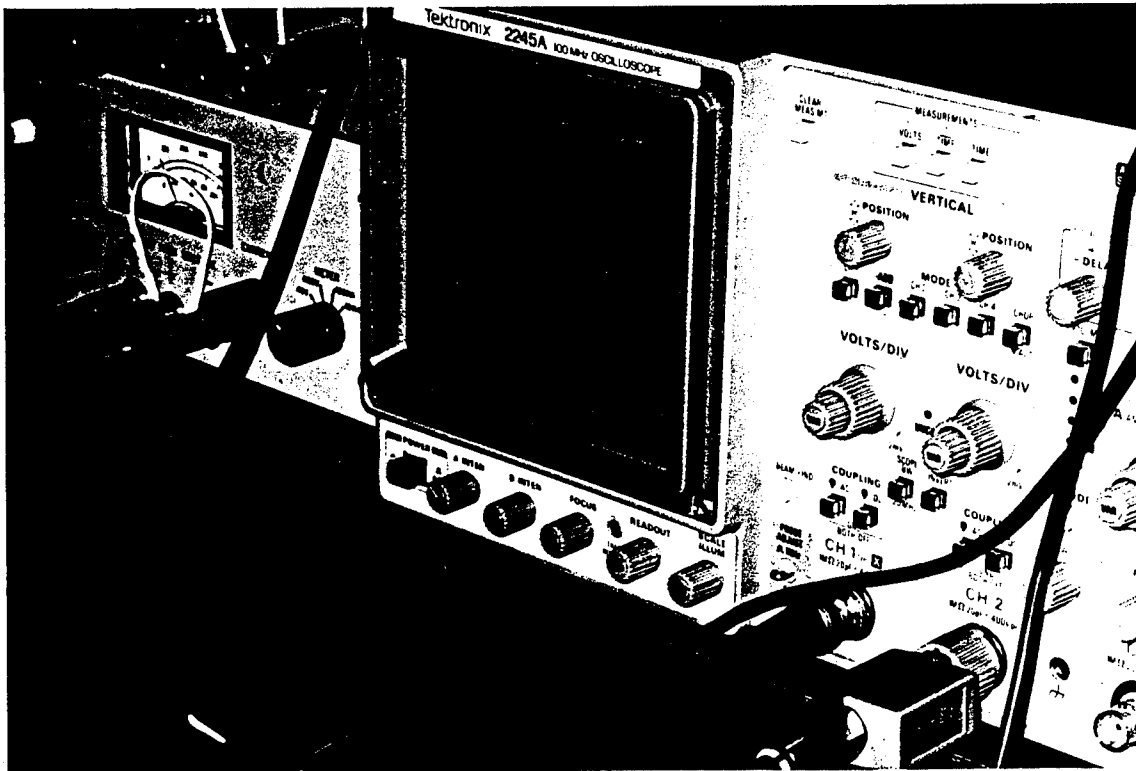


Figure 42: Oscilloscope trace of a 2 MHz signal generated via the clock circuit

The testing is monitored via a CCD camera to align the spot on the desired detector, an optical power meter to measure the incident power on the chip, and an oscilloscope to read the electrical signal generated by the chip. The sensitivity of the detectors was measured to be $4 \mu\text{W}$ at 1 MHz. The clock generation circuit (see figure 42) was operated at 2 MHz and exhibited rise and fall time in the order of 10 nsec, which indicates that it can run at or close to 100 MHz.

The testing also revealed one major mistake in the layout of the circuit. In order to avoid latch-up due to some light reaching the chip outside the detector area, the active region around the transistors has been extended. However, this created some local problems, since this overextended active area happened to short circuit with some of the vias on the metal connections on the chip. For this reason, the 8x8 chip can not be operated in a full system demo.

7. Summary

This Content Addressable Memory system achieves high digital accuracy and high throughput associative search via the electronic tree-based fan-in structure of the OEIC. This enables the system to be applied to a wide variety of database, low-level image processing and pattern recognition tasks. When used in conjunction with the motionless-head parallel readout optical disk system that stores approximately 15,000 bit-planes of 128×128 bits at 2400 rpm, the associative memory system achieves a capacity of about 250 Mbits and a maximum processing speed of 600,000 bit-planes per second or 10^{11} bit-operations per second for 128×128 pixel bit-planes. Note that there is no upper limit on the number of stored bit-planes, contrarily to other systems that are restricted by the algorithm and accuracy (e.g. Hopfield models). If required, the capacity of the system can be increased arbitrarily by using several optical storage devices (e.g. in a jukebox fashion) without reducing the processing speed.

8. References

- 1 B. Robinson, "Grand challenge to Supercomputing," *Electron. Eng. Times*, 18 Sept. 1989.
- 2 1 Gbyte solid state disk, specification sheet, DEC electronics catalog (1993).
- 3 Redundant Arrays of Inexpensive Disks (RAID), specification sheet, Storage Concept Inc. (1993).
- 4 D. B. Carlin et al., "Multichannel optical recording using monolithic arrays of diode lasers," *Appl. Opt.* 23 (22), 3994, Nov. 1984.
- 5 J. Rilum and A. Tanguay, "Utilization of optical memory disk for optical information processing," in *Technical Digest, OSA Annual Meeting*, paper M15 (1988).

- 6 D. Psaltis, M. Neifeld, A. Yamamura, and S. Kobayashi, "Optical memory disk in information processing," *Appl. Opt.* **29**, 2038-2057 (1990).
- 7 P. Marchand, A. Krishnamoorthy, K. Urquhart, P. Ambs, S. Esener, and S. H. Lee, "Motionless-head parallel readout optical disk system," *Appl. Opt.* **32**, 190-203 (1993).
- 8 Code V is a trademark of Optical Research Associates, Pasadena, Ca.
- 9 R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, **35**, 2, 237-246 (1972).
- 10 51/4" Write-Once, Read-Many Optical Disk Specification Sheet, Daicel Chemical Industries, Torrance, CA, 1991.
- 11 OHMT-300 Media Tester Specification Sheet, Apex Systems, Inc., Boulder, Co, 1990.
- 12 P. Kabal and S. Pasupathy, "Partial-response signaling," *IEEE Trans. on Comm.*, vol. 23, no. 9, pp. 921-934, Sept. 1975.
- 13 H. Kobayashi, "Application of partial-response channel coding to digital magnetic recording systems," *IBM J. Res. Dev.*, vol. 14, p. 368, July 1970.
- 14 B. H. Olson and S. C. Esener "Partial response precoding for parallel-readout optical memories," *Opt. Lett.* **19**, p. 661, May 1994.
- 15 L. C. Barbosa, "Characterization of Minimum Noise Partial Response Channels", *IBM Research Report RJ 6475 (62948)*, Oct. 1988.
- 16 see for example, T. Kohonen, *Self Organization and Associative Memory*, Springer Verlag (1984)
- 17 T. Ogura, J. Yamada, S. Yamada, and M. Tan-no, "A 20-Kbit Associative Memory LSI for Artificial Intelligence Machines" *IEEE Journal of Solid State Circuits*, **24**, 1014-1020 (1989).
- 18 H. Takata, S. Komori, T. Tamura, F. Asai, H. Satoh, T. Ohno, T. Tokuda, H. Nishikawa, and H. Terada, "A 100 Mega-access per Second Matching Memory for a data driven microprocessor," *IEEE Journal of Solid State Circuits*, **25**, 95-99 (1990).
- 19 H. Bergh, J. Eneland, and L. Lundstrom, "A Fault Tolerant Associative Memory with High Speed Operation," *IEEE Journal of Solid State Circuits*, **25**, 912-919 (1990).
- 20 A. V. Krishnamoorthy, P. Marchand, G. Yayla, and S. Esener, "Optoelectronic Associative Memory using a Parallel Readout Optical Disk," in *Technical Digest OSA Annual Meeting '90*, Paper MJ5, Boston (1990) and submitted to *IEEE trans. on Neural Networks*.
- 21 H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, Addison-Wesley, 1990.

Rome Laboratory
Customer Satisfaction Survey

RL-TR-_____

Please complete this survey, and mail to RL/IMPS,
26 Electronic Pky, Griffiss AFB NY 13441-4514. Your assessment and
feedback regarding this technical report will allow Rome Laboratory
to have a vehicle to continuously improve our methods of research,
publication, and customer satisfaction. Your assistance is greatly
appreciated.

Thank You

Organization Name: _____ (Optional)

Organization POC: _____ (Optional)

Address: _____

1. On a scale of 1 to 5 how would you rate the technology
developed under this research?

5-Extremely Useful 1-Not Useful/Wasteful

Rating _____

Please use the space below to comment on your rating. Please
suggest improvements. Use the back of this sheet if necessary.

2. Do any specific areas of the report stand out as exceptional?

Yes ___ No _____

If yes, please identify the area(s), and comment on what
aspects make them "stand out."

3. Do any specific areas of the report stand out as inferior?

Yes ___ No ___

If yes, please identify the area(s), and comment on what aspects make them "stand out."

4. Please utilize the space below to comment on any other aspects of the report. Comments on both technical content and reporting format are desired.

***MISSION
OF
ROME LABORATORY***

Mission. The mission of Rome Laboratory is to advance the science and technologies of command, control, communications and intelligence and to transition them into systems to meet customer needs. To achieve this, Rome Lab:

- a. Conducts vigorous research, development and test programs in all applicable technologies;
- b. Transitions technology to current and future systems to improve operational capability, readiness, and supportability;
- c. Provides a full range of technical support to Air Force Materiel Command product centers and other Air Force organizations;
- d. Promotes transfer of technology to the private sector;
- e. Maintains leading edge technological expertise in the areas of surveillance, communications, command and control, intelligence, reliability science, electro-magnetic technology, photonics, signal processing, and computational science.

The thrust areas of technical competence include: Surveillance, Communications, Command and Control, Intelligence, Signal Processing, Computer Science and Technology, Electromagnetic Technology, Photonics and Reliability Sciences.