USC-SIPI REPORT #173

# Photonic Implementations of Neural Networks

by

B. Keith Jenkins and Armand R. Tanguay, Jr.

March 1991

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
Powell Hall of Engineering
University Park/MC-0272
Los Angeles, CA 90089 U.S.A.

# PHOTONIC IMPLEMENTATIONS OF NEURAL NETWORKS

B. Keith Jenkins and Armand R. Tanguay, Jr.

Chapter 15 in *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Bart Kosko, Ed., Prentice-Hall, Englewood Cliffs, New Jersey, 1991.

# CHAPTER 15

# PHOTONIC IMPLEMENTATIONS OF NEURAL NETWORKS

B. Keith Jenkins and Armand R. Tanguay, Jr.

## TOWARDS THE DEVELOPMENT OF A NEURAL NETWORK IMPLEMENTATION TECHNOLOGY

As described in other chapters of this book, neural networks provide a different approach to solving problems as compared with more conventional algorithmic techniques, and can be applied to a wide range of applications. In some of these application domains the simulation performance of neural networks has been comparable to that of more conventional algorithmic approaches. In many application domains, however, a realistic (and therefore large scale) problem may overwhelm the conventional approach, in that it may be too computation intensive to be implemented on a sequential digital computer, and may not parallelize sufficiently well (if at all) for efficient computation on a parallel digital machine. On the other hand, because a neural network algorithm is inherently parallel, it immediately suggests a parallel architecture, which may in turn be implemented using either analog or digital hardware. And for the case of large-scale problems, analog hardware will typically provide a much more efficient neural implementation than digital hardware.

In the previous chapter, fully electronic (primarily VLSI-based) neural networks were described in which the primary functionality of both the neuron units and the weighted

(synaptic) interconnection matrix is incorporated on a planar microelectronic chip. An important advantage of the integrated circuit approach to neural network implementation is the capability for near-term technology insertion, with leverage provided by a well-established technology base characterized by a fully developed computer aided design and computer aided manufacturing (CAD/CAM) device and circuit repertoire. An equally important limitation is the difficulty in scaling up neural chips to incorporate large numbers of neuron units in fully (or near fully) interconnected architectures. This limitation derives from the limited pin-out, off-chip communication bandwidth, and on-chip interconnection density available in both current generation and projected chip designs.

In this chapter, we consider the utilization of optical (free-space) interconnection techniques in conjunction with photonic switching and modulating devices to expand the number of neuron units and complexity of interconnection, by using the off-chip ($3^{rd}$) dimension for synaptic communication. As we shall see, the merging of optical and photonic devices with appropriately matched electronic circuitry can provide novel features such as fully parallel weight updates and modular scalability, as well as both short and long term synaptic plasticity.

Many approaches to the incorporation of photonic and optical technology in the implementation of neural networks are currently being pursued in the research community. The intent of this chapter is not to present a review of these various approaches, the details of which can be obtained from several of the references given in the Suggested Further Reading section at the end of this chapter. Instead, our focus herein is directed toward a description of key photonic devices and techniques based on fundamental optical phenomena, as well as toward a unique and generalizable approach to their potential use in the implementation of large-scale, highly parallel neural network architectures. The unusual nature of some of these techniques has interesting implications for the design of naturally mapped architectures and associated learning/computing algorithms. We will address a number of these unique features in the context of a description of the basic optical phenomena that can be used to advantage, and of the array of photonic devices that comprise the system designer's palette. Because the incorporation of optical and photonic hardware casts the subject of neural network implementations in a somewhat unfamiliar

light, we first discuss a set of desirable and requisite characteristics for neural network implementation technologies.

An important feature of any implementation technology is that of generality: a "building block" approach. The growth and synthesis of material structures, and their incorporation into devices, must be well characterized, understood, and repeatable, for a *small number* of specific material combinations and device structures. These devices are then assembled into circuits or architectures for the implementation of specific computational models. This provides leverage in two ways: (1) the small number of useful and well understood components are used repeatedly in different structures for different applications, and (2) architecture and system level designers need not be experts in the properties of the material and device structures used to configure the components, saving many man-hours in the design of computational systems over a completely custom approach. Such a purely custom approach could preclude the widespread use of these architectures and systems, as has been characteristic, for example, of optical information processing and optical signal processing systems over the past two decades.

Not only is it important for the implementation *technology* to be of a building block nature, but it is also important for the *models* underlying the computational architectures to support such an approach, and in fact to be of a building block nature themselves. Ideally the models would comprise a set of common components and operations at the functional level, such as specific types of neuron units, weighted interconnections, weight updates, and comparisons with desired target values. Then the mapping from model and functional architecture to hardware architecture, layout, and implementation can proceed efficiently as well. This building block approach at both the model and hardware levels has certainly been characteristic of the development of digital electronics, and has been largely responsible for its success.

Assuming that appropriate neural network models and a corresponding technology base can be merged within a compatible building block approach, neural network systems potentially provide a unique capability for large-scale *analog*, *nonlinear* computation. As such, the neural network paradigm potentially alleviates two critical bottlenecks that have impeded the widespread implementation of large-scale analog, nonlinear computing

systems based on non-neural architectures: the lack of appropriate generic hardware components and of the sufficiently leveraged manpower required for their economical design and manufacture, as well as difficulty in establishing efficient techniques for mapping from the application and model domain onto compatible hardware. With regard to the former bottleneck, neural network architectures are generally forgiving with respect to device nonuniformities and imperfections, creating much needed latitude for the device designer. With regard to the latter bottleneck, the neural network paradigm inherently provides a mapping from the problem domain onto a highly parallel architecture, which immediately yields a starting point for its layout in analog hardware.

In the case of photonics for neural network applications, the hardware technology is being developed *simultaneously* with the neural computation model(s). This implies at least two things. First, it is crucial to retain *flexibility* in the functionality of each component, so that as the neural computation models evolve, the hardware can evolve along with it. The development of an entirely new technology base typically takes at least a decade; such a delay between model development and hardware realization is generally unacceptable. Thus, the generic technology base *must* provide sufficient flexibility. Second, not only should the neural computation model steer the technology development, but the reverse can, and indeed *must*, also occur. This assures a mutual compatibility in outcome.

The basic requisite functions for a neural network technology base appear to be: neuron unit response, weighted interconnections (fixed and variable), input/output, learning computation and weight update, and duplication capability (*i.e.*, the capability of making a copy of a network structure). In addition, other features are desirable, such as higher order connection capability. The neuron unit response, at the most common and basic level, is a sum-of-inputs followed by a monotonic nonlinearity. The nonlinearity should have the flexibility of providing different amounts of gain; for example, it is useful to have a high gain to implement a binary threshold, and a low gain to implement a nearly linear response. The neuron unit should be bipolar in that it permits both positive and negative inputs, so that inhibitory and excitatory connections can be realized. This we consider to be the minimal requisite functionality of a neuron unit. In addition, a very desirable feature is the capability for bipolar outputs. Note that biologically this is not necessarily

the case, but useful neural computation models will likely deviate substantially from biological reality and may require this capability. Other capabilities are also desirable, such as leaky integrator effects [Mead, 1989] and more complex behavior such as that required in shunting networks [Carpenter, 1987].

Each weighted interconnection must store a learned or initialized value, and perform a multiplication operation on the signals passing through. An analog multiplication is generally much more efficient than a digital one for reasons of speed and device area or volume. For this reason, it is worth some effort to provide analog storage for the weights. Note that the very large number of weights used in many networks implies that minimization of the incorporated hardware complexity of the requisite storage and multiplication operations is crucial for physical realizability. Input/output is often ignored at the higher levels, but can critically affect the physical architecture and can be a major factor in determining the overall throughput. The input/output function includes the input of signals, the output of results, and the input of weights if necessary.

It is important at the outset to distinguish among systems that have fixed weights; systems that have programmable weights (that are externally computed but loaded into the network); and systems that have full learning capability, in which the learning algorithm is implemented as part of the parallel system. The associated hardware complexity can be quite different in each of these cases. Finally, duplication capability is useful, for example, for replicating a pre-learned network when multiple copies of the network are to be produced and subsequently used with fixed connections. Probing the weight values in a hardware implementation may at first sound straightforward, but implementing very large numbers of weights in a small volume can in some cases preclude such capability.

Applications of hardware implementations of neural networks could include sensor signal processing and fusion, pattern recognition, associative memory, and robotic control. These applications imply a wide range of hardware requirements. For example, most vision processing is characterized by moderately large numbers of neuron units, with small to moderate connectivity and primarily local interconnections. Associative memory, on the other hand, typically requires a very high connectivity.

Semiconductor-based VLSI technology has proven to be very capable for the implemen-

tation of most, if not all, of the above functions. It also has the capability for integration of control circuitry and/or arbitrary digital or logical operations on the same chip as the neural processing circuitry. However, an important issue in neural implementations is that of *scalability*, since many neural network applications are likely to require very large numbers of neuron units and connections. As pointed out above, it is primarily the consideration of scalability that leads us to the conclusion that purely electronic VLSI will work well for certain applications, but can benefit greatly from the incorporation of photonics for other applications.

Two important considerations in VLSI implementations are area complexity and pinout requirements. A fully connected network of $N$ neuron units requires area $O(N^2)$. One can think of this as having only a single linear dimension available for the neuron units themselves, in order to leave room for the connections. This of course limits the size of fully connected networks that can be accommodated on a single chip. For example, chips have been fabricated with 54 neuron units and 2916 ternary (3-level) synapses, and it is estimated that approximately 700 neurons, fully connected with ($10^5$ - $10^6$) similar synapses, could be implemented on a CMOS chip using 0.5 $\mu m$ design rules. Learning capability with analog synapses may require substantially more area per synapse [Jackel, 1988]. On the other hand, networks with low connectivity and only local connections between neuron units permit a much larger number of neuron units to be implemented on a chip. For example, the silicon retina of Mead *et al.* comprises a $48 \times 48$ array of neuron-like units, each connected to its 6 nearest neighbors on a hexagonal grid [Mead, 1988]. A much larger number of neuron units, on the order of $10^5$, could be implemented with such a locally-connected array, since the neuron units and the connections each require area only $O(N)$. So we see that a critical limiting factor in the VLSI implementation of neural networks is the area required for on-chip interconnections, and that the area required for neuron units is relatively inconsequential.

The number of pinouts that can be provided on a chip is proportional to its linear dimension, not to its area. This degree of pinout capacity is well matched to fully connected networks in which the weights do not need to be input or output frequently, and to locally connected networks of low to moderate bandwidth. An example of the latter is a vision

network that operates at video frame rates; in this case the signal can be easily time multiplexed onto a relatively small number of lines for communication onto and off of the chip.

On the other hand, other application areas will require a higher input/output (I/O) bandwidth and/or a large number of neuron units with high connectivity. For example, if the weights are to be fed onto and off of a chip frequently, substantial multiplexing would be required for reasonably large networks (*e.g.*, up to 200 pinouts can generally be accommodated, but as described above, as many as $10^5$ - $10^6$ synaptic weights can be incorporated on the chip). Wafer scale integration with bump bonding techniques can help by providing large, multi-wafer structures, but the number of I/O lines is still likely to be modest due to practical and physical constraints. So we see that a second critical limiting factor in the VLSI (and wafer scale integration) implementation of neural networks is the number of I/O lines that can be practically incorporated. Neural applications utilizing large fully connected networks such as associative memory would benefit greatly from implementations of $10^5$ - $10^6$ fully connected neuron units, implying the need for $10^{10}$ - $10^{12}$ synaptic weights. In addition, the intermediate realm of large networks with partial but moderate-to-large connectivity will likely also prove beneficial to a wide range of applications.

In this chapter, we discuss a variety of issues that impact the development of photonic technology as applied to hardware implementations of neural networks with enhanced capabilities. Photonics has the potential for the implementation of networks with large numbers of neurons ($10^5$ - $10^6$) and high connectivity (approximately $10^{10}$ analog-weighted interconnections) in one "module". The approach taken here is to use electronics to implement the internal function of each neuron unit, and to use optics to implement the connections, weights, and I/O. With this technique most of the area of a two-dimensional (2-D) "chip" can be used to implement the neuron units themselves, and optical free-space propagation and volume holograms can be used to implement the interconnections. Thus the interconnections actually occupy a three-dimensional (3-D) *volume*, which improves scalability dramatically.

The next section of this chapter describes the fundamental optical principles and key

photonic technology concepts that are needed for neural network implementations, and covers photonic analog arithmetic, switching, interconnections, sources and detectors. Architectural considerations are discussed in the subsequent section, including the use of volume interconnections, signal representation, and desired architectural features. The next section then presents a photonic implementation strategy that satisfies most of the desired criteria. In the two concluding sections we investigate the ultimate limitations of photonic implementations of neural networks, and consider the future of such implementations.

# FUNDAMENTAL PRINCIPLES OF PHOTONIC TECHNOLOGY

In order to effectively appreciate the potential advantages as well as the limitations of extending the VLSI (electronic) repertoire to include photonic components and optically inspired functionality, we will first identify and then explain a few truly fundamental principles of the optical and photonic technologies on which such hybrid neural network implementations are based. In this section, therefore, we discuss the basic features of optical analog computation with both coherent and incoherent illumination sources, photonic switching devices and their various neuron-like functions, the characteristics of photonic interconnections that are essential to the implementation of synapse-like interneuron wiring, and the principal features of sources (photonic power supplies) and detectors (photonic-to-electronic power converters).

## Optical Analog Computation

At the present time, most proposed photonic implementations of neural networks are based on analog operations, both in the representation of neuron outputs as well as in the incorporation of interconnection weights. This emphasis of *analog* computation over

perhaps more familiar *digital* computation derives principally from several distinct advantages that accrue to optical systems designed to handle the switching and interconnection of very large numbers of inputs and outputs at each circuit node (neuron unit). These advantages include a significant reduction in the number of switching components required to sum multiple inputs [Abu-Mostafa, 1989], an increase in the degree of fan-in and fan-out allowable from each circuit node, the elimination of analog-to-digital and digital-to-analog converters, a significant decrease in signal routing and interconnection complexity, the potential utilization of natural physical phenomena within certain photonic devices to accomplish difficult computational functions directly, and the possibility of higher computational throughput per unit dissipated energy in operations characterized by high computational complexity. Additionally, two notable disadvantages of analog systems, error accumulation and lack of precision in representation, may prove to be relatively unimportant in the neural network environment, due in part to the self-organizing and error correcting nature of many neural learning algorithms [von der Malsburg, 1987]. We will return to a number of these issues throughout the remainder of this chapter.

The computational operations necessary for the implementation of a wide range of neural and neural-like networks are surprisingly simple, consisting primarily of addition, subtraction, multiplication, and nonlinear thresholding. The operation of addition usually must be performed over a very large number of inputs at a given neuron unit, representing weighted excitatory signals from other interconnected neuron units; subtraction is utilized to differentiate excitatory inputs from inhibitory inputs to a given neuron unit, as is common, for example, in models of the visual process and of associative memories. Multiplication is necessary for the provision of linear interconnections with signal-independent weights, which in turn store learned (or pre-programmed) information, and hence form an important constituent of the neural paradigm. Finally, nonlinear thresholding operations are performed in order to provide the appropriate transfer function between the neuron activation potential (sum of all inputs, both positive and negative, to a given neuron unit), and the output each neuron unit generates in response. It is this nonlinearity in particular that gives multilayer neural networks their computational power, and allows recurrent networks to iteratively approach one of several stable states of the system. The detailed

nature of the nonlinearity itself can affect a number of critical system properties, including the number of iterations (or equivalently the time) required to achieve steady state, and the stability of the network in the presence of noise, inaccuracy, and nonuniformities among both the neuron units and the interconnections. In some envisioned neural network implementations, it is important to also be able to alter the nature of the nonlinear thresholding function dynamically during the computational phase of operation.

In addition to these elementary operations (and combinations thereof), it is necessary to accommodate operationally for both the learning function (which includes input-dependent interconnection weight updates) and for the computational function (which in general requires iterative feedback and the implementation of nonlinear thresholding, usually with fixed interconnection weights). The only additional fundamental operation required by these features, that of input-dependent interconnection weight updates, can usually be reduced to at most a combination or sequence of the previously described fundamental operations of addition, subtraction, multiplication, and nonlinear thresholding. It is important to note, however, that in this case the operations pertain directly to the implementation of interconnection weights, rather than to functions performed by the neuron units; hence, the physical processes involved may in fact be considerably different in nature, and thus subject to a quite different set of constraints. This differentiation between the two different "types" of basic operations (those pertaining to the neuron units and those pertaining to the interconnection pathways and weights) is discussed in more detail below, as well as in the section describing the fundamental physical and technological limitations of neuro-optical computing.

At the outset, we must further differentiate between those fundamental photonic operations that are intended for "optical" implementation, and those that are envisioned for "optoelectronic" implementation. In the first category, we place those types of physical processes in which one or more beams of light interact either *directly* (as in coherent interference) or through an intermediate physical medium (as in the summation of two incoherent beams of light on a single detector). In the second category, we place operations that involve one or more photon-to-electron or electron-to-photon conversion processes prior to the actual implementation of the desired function, which is then assumed to be

accomplished using intermediate (primarily analog, but perhaps digital) circuitry. An example of this latter category might be the subtraction of two optical signals by independent but simultaneous photodetection of each signal, followed by the use of an analog electronic differential amplifier to execute the functional subtraction. Most proposed neuro-optical processors rely to some degree on both types of physical implementation mechanisms (*c.f.* the sections "Architectural Considerations for Photonic Neural Network Implementations" and "An Implementation Strategy", below). In fact, the eventual degree of success achievable by neuro-optical computing techniques rests heavily on an appropriate balance of these mechanisms within a given system, optimized to yield the greatest computational advantage within the allowable physical and system constraints.

Optoelectronic implementations of neural functions for the most part rely on optical signals as inputs and outputs to and from the neuron units in order to allow for both high bandwidth and high interconnection multiplexing capacity, and on electronic signal combination and processing locally within each neuron unit. As such, optoelectronic functions are characterized primarily by the *optical* characteristics of interconnection, and by the *electronic* characteristics of computation. The former will be described below, while the latter has been discussed both by Bang Lee and Bing Sheu elsewhere in this volume [Lee, 1991] as well as by Carver Mead in a recent elegant monograph [Mead, 1989]. On the other hand, optical implementations of these functions rely on communication by, as well as the interaction of, two or more optical signals in order to accomplish the relevant computation, which is often (but not always) followed by a photon-to-electron conversion process in some form of single channel or array detector. It is to the fundamental principles of such optical computational interactions that we next turn our attention.

In order to adequately consider even so basic a process as optical addition, it is essential to differentiate between two basic types of optical interactions (as determined by the nature of the optical signals involved): *incoherent* and *coherent*. Incoherent interactions occur whenever the light wavefronts representing the input signals temporally dephase (do not oscillate in unison) over the relevant time of observation (detector temporal integration window), in that they are either both temporally incoherent at the outset, or are *each* temporally coherent but separated in optical frequency by more than the inverse of the

observation time. Interactions in which the input optical signals spatially dephase over the aperture of the relevant detector wherever the output is utilized (detector spatial integration window) are also incoherent for all practical purposes, and will obey incoherent summation rules as given below. Coherent interactions occur, on the other hand, whenever the light wavefronts representing the input signals simultaneously maintain a constant phase relationship over the detector spatial and temporal integration windows.

From these remarks, it can be seen that it is quite important to understand the distinction between coherent (or incoherent) *light* and coherent (or incoherent) *interactions* as defined by the eventual detector configuration and operational parameters. For example, it is perfectly acceptable to consider a situation in which two mutually coherent optical beams interact to produce an interference pattern with a spatial scale small compared with the relevant detector aperture. In such cases, the interaction will in fact follow *incoherent* summation rules, as the detector effectively *integrates* the space-variant interference pattern over the full detector aperture to produce exactly the same result as the interaction of two mutually incoherent (temporally) optical beams.

Given these preconditions, then, the actual rules for the basic operations are quite straightforward. Consider first the case of addition of two incoherent optical signal beams in a *collinear* geometry, in which the two distinct input beams with intensities $I_1$ and $I_2$ are assumed to emerge from a beam-combining optical system (as yet undetermined) such that the two output beams are collinear. If the beams are combined, for example, by a nondispersive (wavelength-insensitive) 50/50 (50% transmission, 50% reflection) beamsplitter as shown in Figure 15.1(a), two possible output beams $I_{out}$ and $I'_{out}$ are created, each with an intensity given by:

$$I_{out} = I'_{out} = \frac{1}{2}(I_1 + I_2). \tag{1}$$

The output intensity is thus linearly proportional to the sum of the input intensities. Note that this operation cannot be accomplished without an inherent loss, in the case shown above equal to 0.5 or about 3 *dB*. In fact, if we wish to combine $N$ beams collinearly by this technique (using a linear chain of non-dispersive beamsplitters), $N-1$ beamsplitters

are required with transmissivities given by 1/2, 2/3, 3/4, ..., $(N-1)/N$, representing a total loss of $(N-1)/N$ with an overall throughput of $1/N$:

$$I_{out} = \frac{1}{N}(I_1 + I_2 + I_3 + \cdots + I_N). \tag{2}$$

Instead of using a linear chain of beamsplitters with different transmissivities, we could alternatively construct a binary tree structure by pairing the inputs that again requires $N-1$ beamsplitters, but in this case with *equal* transmissivities of $\frac{1}{2}$. This system of beamsplitters also exhibits an overall throughput of $\frac{1}{N}$ (for even values of $N$). Although beam combination of a large number of inputs by the multiple beamsplitter method is impractical, we will utilize this result a little later in order to understand the essential features of holographic beam combiners, which are subject to many of the same constraints.

**FIGURE 15.1** Illustration of optical addition utilizing a 50/50 beamsplitter: (a) collinear *incoherent* beam geometry; (b) collinear *coherent* beam geometry, showing input and output *amplitudes*; (c) collinear *coherent* beam geometry, showing input and output *intensities*.

It should be noted in passing that the overall throughput loss implied by Equation (2) can be circumvented *if* the beams to be summed incoherently are sufficiently distinct in wavelength that a *dichroic mirror* can be used to combine them. A dichroic mirror reflects light within a given wavelength range, and transmits light outside of that range. Multiple dichroic mirrors can be used to collinearly sum multiple beams through appropriate choice of the input wavelengths in each arm, and of the characteristic wavelengths of each succeeding dichroic mirror.

Consider next the case of addition of two *coherent* optical signal beams in a collinear geometry. An example, again using a beamsplitter, is shown schematically in Figure 15.1(b). The presence of the beamsplitter generates a $\pi/2$ phase shift for each transmitted

beam, and a $\pi$ phase shift for each reflected beam [Haus, 1984]. In this case, due to the coherent nature of the two input signal beams, the output intensity is no longer given simply by the sum of the input intensities. In fact, the output *amplitude* is proportional to the sum of the input signal *amplitudes* (provided that the two beams have identical polarizations):

$$a_{out} = \frac{1}{\sqrt{2}}(ia_1 - a_2 e^{i\phi}) \tag{3}$$

in which $\phi$ is the relative phase between the two wavefronts (here assumed constant over the detector aperture). It should be noted parenthetically that optical beams with orthogonal polarizations do not interfere, and hence follow the incoherent addition rule. We assume throughout this chapter that all beams are polarized identically.

From this simple equation, several important principles can be seen to emerge. First, the representation we must choose for simple addition to occur with coherent light is different than in the case of incoherent light: in the coherent case, we must use the *amplitudes* (containing phase information), whereas in the previous (incoherent) case addition is linear in the *intensities*. Second, the input-output transformation represented by Equation (3) reveals an easy method for implementing both addition and subtraction: we merely set the phase difference to $-\pi/2$ for addition, and to $\pi/2$ for subtraction. This can be accomplished either by adjusting the relative path lengths of the two input beams, or by inserting an appropriately oriented wave plate in one of the two beams. In the incoherent case treated above, no such algorithm exists since we are adding intensities (which are positive definite quantities), and direct subtraction is not possible without intermediate intervention by an active optical or optoelectronic device. Third, note that the second output beam is now not symmetric with the first:

$$a'_{out} = \frac{1}{\sqrt{2}}(-a_1 + ia_2 e^{i\phi}). \tag{4}$$

This asymmetry in output amplitudes directly results from the asymmetry between the phases of the reflected and transmitted components in a partially transmitting mirror [Haus, 1984].

Since the intensity of an optical beam is related to its amplitude by the relation:

$$I_m = (\mathbf{a}_m^*)^T \cdot \mathbf{a}_m \tag{5}$$

in which $\mathbf{a}_m$ is the vector amplitude of the wave representing its polarization, $*$ represents the complex conjugation operation, and $T$ represents the transpose of the vector, the output intensities in the two coherently summed channels are given by:

$$I_{out} = \frac{1}{2}[a_1^2 + a_2^2 + 2a_1a_2 \sin\phi], \tag{6}$$

and

$$I'_{out} = \frac{1}{2}[a_1^2 + a_2^2 - 2a_1a_2 \sin\phi], \tag{7}$$

as shown in Figure 15.1(c). From these two equations, it can be seen that for arbitrary values of the phase shift $\phi$, the output intensity is not simply related to the sum of the input intensities, but instead has a seemingly undesirable cross term. We can use this cross term to advantage by noting that for phase shifts of 0 or any integer multiple of $\pi$, both output intensities are equal, and reduce to the expression previously noted for the incoherent case (Equation (1)). Thus for coherent illumination, we can either perform addition directly with the amplitudes, or with the intensities if we are careful about proper phasing of the input signals. One difficulty with the former approach is that most detectors are linear in intensity but not in amplitude, as will be discussed in further detail below.

In direct analogy with the analysis presented for the case of incoherent multiple signal beam summation, we can extend the above equations to include the case of multiple collinear coherent inputs using appropriate combinations of beamsplitters. For $N$ coherent input beams summed optimally, the output amplitude is given by:

$$a_{out} = \frac{i^{N-1}}{\sqrt{N}}[a_1 + a_2 + a_3 + \cdots + a_N]. \tag{8}$$

In order to achieve this result, one must again use $N-1$ beamsplitters with (intensity) transmissivities identical to those employed in the incoherent case, and in addition the

relative phases must be arranged such that the phase difference between $a_2$ and $a_1$ is $-\pi/2$, and each successive beam is *increased* in phase by $\pi/2$.

In all of the cases described above, we have constrained the problem by requiring that the output beams all be *collinear*, and in fact many proposed neuro-optical architectures implicitly demand such a constraint. We will show in later sections, however, that this is perhaps an unnecessary and in many cases undesirable constraint. Hence we consider here also the case of optical addition with *noncollinear* output beams, requiring instead only that the summed beams fill the same detector aperture. There are in fact a number of interesting variants of these two constraints, but we will limit the discussion to the two principal cases only. One possible such configuration is shown in Figure 15.2(a), in which two incoherent signal beams are summed within the detector aperture by using two 100% reflecting mirrors, producing an output intensity given by:

$$I_{out} = I_1 + I_2. \tag{9}$$

This output intensity is uniform across the detector aperture, as shown schematically in the figure. In addition, relaxation of the requirement for collinearity can be seen to now allow for the use of mirrors instead of beamsplitters, eliminating the loss we found in the previous case at the expense of increased *angular* multiplexing.

**FIGURE 15.2** Illustration of optical addition utilizing mirrors: (a) angularly multiplexed *incoherent* beam geometry; (b) angularly multiplexed *coherent* beam geometry.

Up to this point in the discussion, we have had to consider the *phase* of the optical wavefronts only for the case of collinear, coherent addition. In that case, we only needed to use the *relative* phase shift between the two beams to derive Equations (6) and (7), since the phase shift is constant in both space and time over the detector spatial and

temporal integration windows. In order to consider the case of *noncollinear, coherent* addition, however, we must allow for the space-variant phase shifts that naturally result when two coherent wavefronts cross at a non-zero angle. These effects are automatically taken into account if we express each wave (beam) amplitude in a form that incorporates both its *magnitude* and its *phase* everywhere in space at a given instant of time. For a plane wave (in which the planes of constant phase are oriented normal to the direction of propagation), it proves convenient to use the form $\mathbf{a}\exp(i\mathbf{k}\cdot\mathbf{r})$, in which $\mathbf{a}$ is a vector representing the polarization of the optical wave (its amplitude in each of the principal coordinate directions), $\mathbf{k}$ is the *wave vector* of the optical wave (defined as a vector with direction normal to the planes of constant phase, and with magnitude $|k| = 2\pi n/\lambda$, in which $n$ is the refractive index of the propagation medium and $\lambda$ is the wavelength of the light wave in vacuum), and $\mathbf{r}$ is a position vector defined from an arbitrary origin in space ($\mathbf{r} \equiv x\hat{x} + y\hat{y} + z\hat{z}$, in which $\hat{x}, \hat{y}$, and $\hat{z}$ are unit vectors along the Cartesian coordinate axes.

For the case of coherent illumination, then, the result of noncollinear summation is as shown schematically in Figure 15.2(b). In this case, the phase of each input wave varies across the detector aperture (assumed to lie in the plane $z = 0$) at a rate that is a function of the angular separation of the incident beams, as well as of the angular deviation of the bisector from normal incidence. This condition can be represented by writing the wave amplitudes with a space-variant phase, which is in turn dependent on the $x$-component of the wave vector $k_x$ in the form $a_1\exp(ik_x x)$. The two waves will thus interfere in the plane of the detector, forming essentially a new wave with a local amplitude given by the sum of the incident amplitudes. The resulting intensity pattern has both a space-invariant (uniform) and a space-variant (sinusoidal) component:

$$I_{out} = a_1^2 + a_2^2 + 2a_1a_2 \cos 2k_x x. \tag{10}$$

If we assume that the detector is linear in intensity over the dynamic range represented by this equation, and furthermore that the detector is uniform in responsivity over its aperture, then the output from the detector will be the *spatial average* of this interference

pattern, resulting in an output intensity that is in fact a sum of the input intensities, as represented by Equation (9). In this case, the result is independent of the relative phase shift (difference) between the two beams at their points of entry into the beam combiner system, since such a phase shift will merely result in a translation of the interference pattern without altering its integrated value. This result also can be extended to the case of multiple input signal beams, with the stipulation that mirrors cannot be allowed to occlude each other; hence, for a given beam width, only a certain number of beams can be combined without loss by means of this method without overcrowding the available angular spectrum.

The operation of optical multiplication is fundamentally different in a number of ways from those of addition and subtraction. Perhaps the most important difference is that the multiplication of either beam amplitudes or intensities cannot be accomplished directly, but must instead utilize a nonlinear medium of some form within which the beams can interact. There are two principal types of interactions to consider: those in which the two beams must be present *simultaneously* in order to form the desired product, and those in which the two beams are utilized *sequentially* in time. In general, the former interactions tend to operate on the amplitudes and hence require mutual coherency, whereas the latter interactions typically form products of (incoherent or coherent) intensities, which are therefore more straightforward to detect with currently available intensity-sensitive detectors.

. Simultaneous multiplication of two optical beams is suggested by Figure 15.2(b), in which two coherent signal beams are angularly multiplexed to form the interference pattern given by Equation (10). Note that the space-variant part of the output intensity in the plane of the "detector" is proportional to the product of the amplitudes, *i.e.*, is of the form $2a_1a_2 \cos 2k_x x$. If instead of employing a uniform (spatially averaging) detector as before, we were now to employ a space-variant detector sensitive to the local intensity, it is possible to record this modulation term along with the unmodulated (uniform) bias represented by the squares of the two amplitudes. If in addition the "detector", for example, is assumed to generate a change in either its absorption coefficient or its refractive index as a function of the recorded intensity pattern (for a given exposure), a diffraction grating

will be formed. The resulting diffraction grating can be characterized by an amplitude that is proportional to the product of the input signal beam amplitudes, and that can be probed by a third so-called "readout" beam. This is at once the basic principle of holographic recording (as explained in more detail below in the subsection on "Photonic Interconnections"), and at the same time allows the implementation of the multiplicative operation for coherent inputs. It should be noted that although this process produces a useful result for the case of two inputs, extension to larger numbers of inputs is not trivial, and requires the utilization of higher order terms in the susceptibility tensor (representing the complex dielectric constant) for implementation. The one exception to this rule is the use of the probe (readout) beam intensity $I_p$ as a third effective input, in which case output intensities proportional to either $I_p a_1 a_2$ or $I_p I_1 I_2$ can be detected depending on the operational parameters of the recording medium and readout configuration.

If the simultaneity requirement is relaxed to allow for sequential interactions in an intervening photosensitive medium, then it is possible to multiply two incoherent input signals by means of the simple generic scheme shown in Figure 15.3. The medium again acts as an effective detector for beam 1, generating a transmittance (in its range of linearity) proportional to the intensity of beam 1. This transmittance can be generated either directly, or through the exposure given by the product of the intensity and the exposure time as in the familiar case of photographic film. Beam 2 is effectively employed as a probe beam, such that the output intensity is given by:

$$I_{out} = cI_1 I_2, \tag{11}$$

as desired, in which $c$ is a proportionality constant subject to the constraint that:

$$cI_1 \leq 1. \tag{12}$$

This process, as in the previous case, is extendable to accommodate an arbitrary number of inputs by iteration, unfortunately resulting both in a lengthy generation sequence for a large number of inputs, and in the potential for significant nonlinear effects with a heavily constrained overall dynamic range. For the case of $N$ input beams, we can utilize $N-1$

exposure steps in combination with $N-1$ intermediate readout steps and a final readout step with beam $N$ to generate an output intensity of the form:

$$I_{out} = c^{N-1} I_1 I_2 I_3 \cdots I_{N-1} I_N, \text{ with } c^{N-1} I_1 I_2 I_3 \cdots I_{N-1} \leq 1. \tag{13}$$

In direct analogy to the case of summation, we could instead utilize a binary tree structure, which requires only $log_2 N$ time steps but uses the same *number* of devices.

Finally, it should be noted that this latter process of incoherent beam multiplication through an intervening medium by sequential illumination is suggestive of the process of *spatial light modulation*, in which the same basic concept is extended to cover a two-dimensional array of multiplication elements. In fact, this process is an essential component of the general area of photonic switching, to which we will turn our attention below.

**FIGURE 15.3** Illustration of optical multiplication utilizing a medium with variable transparency.

Before turning to the topics of photonic switching and photonic interconnections, we conclude this section with a discussion of the fourth principal computational process that can be performed optically (as opposed to optoelectronically, as discussed below): that of the incorporation of functional nonlinearity. Although many types of functional nonlinearities are of interest in a generalized analog computational system, those of primary utility in the neural network environment are for the most part threshold-like in nature. A threshold function $f_T(x)$ of some input variable $x$ (such as the input intensity, for example) can be described in general by:

$$\begin{aligned} f_T(x) &\cong T_{min}; \quad -\infty \leq x < x_1 \\ f_T(x) &= m(x); \quad x_1 \leq x < x_2 \\ f_T(x) &\cong T_{max}; \quad x_2 \leq x \leq \infty \end{aligned} \tag{14}$$

in which the function $m(x)$ is a monotonic function with a minimum value of $T_{min}$ and a maximum value of $T_{max}$. For a step function response, the function $m(x)$ can be eliminated by setting $x_1 = x_2$. In many cases, a smoother transition between the two extreme states has been found to generate enhanced network stability and faster settling times. In such cases, the function $m(x)$ may be taken, for example, as a sigmoid with an exponential onset and an asymptotic approach to the saturation level.

The incorporation of such nonlinear functionality by direct optical means can be achieved through the use of a number of different types of nonlinear materials; such materials typically exhibit a change in their refractive index or absorption coefficient proportional to the first and higher order powers of the local optical intensity. One example of such a material is photographic film, which after development exhibits a (negative) sigmoid-like exposure characteristic, with a saturation value determined by the maximum optical density achievable within a given film thickness. (The optical density (OD) of a medium is given by the negative of the decadic logarithm of its transmittance; for example, a film that transmits 1% of the incident illumination has an optical density of two (OD2)). Another common example of an optical nonlinearity is the photoconductive saturation behavior of certain semiconductor materials such as cadmium sulfide, zinc selenide, and silicon. In this latter case, the distinction between an "all-optical" nonlinearity and an optoelectronic nonlinearity becomes somewhat blurred, as the photoconductor can be thought of as a light-sensitive electronic device.

Such optical techniques for the generation of functional nonlinearities at present suffer several inherent disadvantages, in that they often require either on off-line post-exposure development step (which is unsuitable for real time operation at high frame rates), long response times, or very high optical intensities to achieve saturation. In addition, such materials as of yet have not proven to be readily programmable, which is often a desirable feature from the systems perspective in order to accommodate variable threshold functions, gain, saturation values, and offsets. As we will see in the next section, the incorporation of electronic circuitry with optical detectors and modulators to achieve *optoelectronic* nonlinearities can in fact greatly increase the threshold sensitivity and operational bandwidth of nonlinear switching elements, while simultaneously providing flexible programming ca-

pabilities.

## Photonic Switching

The switching function, that of providing an output that is (perhaps nonlinearly) dependent on one or more inputs, is a principal distinguishing characteristic of neuron units. Electronic circuit elements (particularly as configured by very large scale integration techniques) are quite well suited to the switching task, as long as the number of inputs (representing the fan-in) and the number of outputs (representing the fan-out) are both kept relatively small (less than a few hundred or so for the case of analog fan-in and fan-out). However, for neural network implementations that demand a high degree of connectivity (with a concomitantly large number of neuron units), the required gate count as well as the *area* required for interconnection routing in purely electronic implementations rapidly gets out of hand.

The fundamental aspects of the fan-in and fan-out components of the switching function are quite distinct, and lead to different types of demands on the chosen implementation technology. The *fan-in* of a number of inputs requires that a particular functional relationship be established between the generated output, on the one hand, and the set of inputs, on the other. In the case of a neural network, the output typically depends on both sums and differences of various combinations of the inputs. Therefore, a given implementation technology must properly generate the requisite logical or functional relationship, as well as provide for an appropriate physical input mechanism (*e.g.* the input leads in the case of an electronic implementation). For electronic circuits, the network area required for the provision of input leads and functional circuitry typically scales directly with the number of inputs, which is an unfortunate dependence when the number of inputs is large. *Fan-out*, on the other hand, usually implies the broadcast of a single output value to a number of (input) locations or nodes. In electronics, the output power required to drive the inputs to a large number of nodes scales directly with the number of these nodes, which again does not scale favorably (but turns out to be an unavoidable penalty in any case). Significant

($N$-fold) fan-out often involves the incorporation of high power driver circuitry, which may have to be duplicated $M$ times ($M < N$) in order to avoid unacceptable loading of the output stages.

The combination of both fan-in and fan-out components of the switching function reveals a further demand on the real estate required for the establishment of weighted *interconnections*. In a fully connected neural network with $N$ neurons, for example, area must be provided for the incorporation (storage and programming) of $N^2$ independent weights as well as $N^2$ independent signal pathways. Hence, the chip area required in a VLSI circuit implementation of such a fully connected neural network will scale at least as the square of the number of neuron units ($O(N^2)$). Network segmentation into a number of interconnected chips can help somewhat to expand the network size beyond the limitation imposed by applying this constraint to a single chip. However, the limiting factor in the multiple-chip case rapidly becomes interchip communication (I/O), as pinouts from VLSI chips of greater than two hundred or so are not technologically feasible at present.

Photonic implementations of neural networks take advantage of the simple beam-combining mechanisms outlined above to multiplex inputs and outputs, and as such exhibit much higher capacity for fan-in and fan-out than do typical electronic implementations. The utilization of *optical* rather than electronic interconnections for the fan-in and fan-out functions provides for completely different scaling laws at large numbers of inputs and outputs to a given neuron unit, as described in more detail in the following subsection ("Photonic Interconnections").

Even given photonic interconnections with a high degree of fan-in and fan-out capability, the nonlinear functional (switching) relationship between the output and combinations of the inputs must still be provided for. For purposes of neural network implementation, the primary photonic switching component is the *spatial light modulator*, a device that alters either the amplitude or phase across an expanded probe beam in response to the local intensity (or exposure) across an input (writing or recording) beam.

The simplest example of a spatial light modulator, albeit one that cannot operate at real time frame rates, is photographic film. Following exposure to an information-bearing optical field, in which an image of a given scene is brought into focus on the two-dimensional

plane of the film, a "latent" (undeveloped) image is formed within the photographic emulsion on the surface of the film. Chemical development is used to transform the latent image into a measurable change in the optical transparency (transmissivity) of the film, which can then be "read out" or probed by secondary illumination to reveal features of the recorded scene. In this context, slide projection is in fact the equivalent of *amplified* readout with a probe beam, in the sense that the reconstruction of the image is accomplished at a much higher level of intensity (for a longer period of time) than the original exposure.

As can be seen from this example, the basic functions performed by a spatial light modulator are those of *detection, functional transformation*, and *optical modulation*, as shown schematically in Figure 15.4(a). In the case of photographic film, the detection process occurs at photosensitive centers during exposure, the functional transformation (the transfer function that relates the output transparency to the input exposure) is incorporated during development and fixing, and the optical modulation process is accessed during readout. This division of the spatial light modulation function into three key elements is particularly useful in the discussion of optoelectronic spatial light modulators, which typically consist of separately identifiable detectors, control circuitry, and modulation elements, as shown schematically in Figure 15.4(b-d). This functional division also allows extensive use to be made of sophisticated electronic circuitry deployed locally within each pixel, both to generate programmable nonlinear control functions and to compensate to a certain degree for the nonidealities inherent in the optical detection and optical modulation elements.

**FIGURE 15.4** Fundamental principles of spatial light modulator function: (a) block diagram of the principal functions of an optically-addressed spatial light modulator, including the detection, functional implementation, and modulation functions; (b) schematic diagram of an $N \times N$ array of spatial light modulator pixels, in which three pixels are shown in different transmission states; (c) expanded view of the pixel array, showing an incomplete fill factor within each pixel; (d) expanded view of a single pixel within the array, illustrat-

ing one possible pixel configuration that incorporates two detector elements $D_1$ and $D_2$, control electronics for impedance matching and functional implementation, and two modulator elements, shown here in different transmittance states.

Up to this point in the discussion, we have focused on *optically-addressed* spatial light modulators (OASLMs) that respond locally to the incident light intensity, as this light detection function is common to most envisioned photonic neural network architectures. Another way of controlling the modulation within an array on a pixel-by-pixel basis is to configure the spatial light modulator such that it can accept a serial or parallel electronic input signal, which can be decoded (or demultiplexed) to drive each individual modulator element. Such *electrically-addressed* spatial light modulators (EASLMs) can be driven, for example, by the output of a television camera to again combine the functions of detection, functional transformation (which may be accomplished in an external circuit), and modulation. One advantage of such a combination is the current advanced state of the art in closed circuit television cameras (CCTVs), which exhibit exceedingly high performance at relatively low cost. One notable disadvantage, however, is the implied limitation on the frame rate of the combined device, since most high resolution TV cameras are designed to operate at less than one hundred frames per second.

Over the past two decades, a wide range of physical modulation mechanisms have been investigated for use in various types of spatial light modulators. Such mechanisms include the modulation of the index of refraction or birefringence in single crystal materials by means of an applied electric field (the *electrooptic* effect), the reorientation of liquid crystal molecules (producing in turn a change in the index of refraction or birefringence) by either an applied electric field or by local optically-induced heating, changes in coloration produced by optical absorption (the *photochromic* effect), modulation of the polarization of reflected light by application of local magnetic fields (the *magnetooptic* effect), surface deformations in a thin film or membrane induced by either applied electric fields or local optically-induced heating, changes in the local refractive index induced by the application of pressure or by the transmission of an acoustic wave (the *acoustooptic* effect), and

electric field modulation of the absorption or dispersion properties of semiconductor device structures. The utilization of these physical modulation mechanisms in various spatial light modulator configurations has been addressed in a number of review articles [Tanguay, 1985; Warde, 1987], journal special issues [*e.g.*, *Spatial Light Modulators for Optical Information Processing*, 1989], and topical conference proceedings [*e.g.*, *Spatial Light Modulators and Applications*, 1990].

The principal configurational and operational characteristics of spatial light modulators that are of interest for application to neural networks include optical sensitivity, write (input) wavelength, read (output) wavelength, input-output transfer function, functional programmability, operational bandwidth, degree of integration, pixel size, total number of pixels per chip, output modulation contrast ratio, dynamic range, and dissipated power density. In many cases, these characteristics are interdependent, and thus impose at times contradictory design constraints that must be optimized in the overall systems context. The fundamental and technological limitations that affect device design and performance are discussed further below and in a succeeding section.

As is the case for electronic circuitry, both monolithic and hybrid approaches to the development of optoelectronic spatial light modulators with suitable functionality have been employed. In the *monolithic* approach, the detectors, control circuitry, and modulation elements within each individual picture element (pixel) are integrated within a single class of materials on a supporting substrate, as shown schematically in Figure 15.5. An example of such an approach is the integration of $p$-$n$ or $p$-$i$-$n$ junction photodiodes with metal-semiconductor field effect transistors (MESFETs) [Sze, 1981a] to drive multiple quantum well (MQW) optical modulators based on the quantum confined Stark effect (QCSE) [Miller, 1990], all fabricated by means of photolithographic processing with multiple mask levels on gallium arsenide ($GaAs$) substrates. In Figure 15.5, two distinct approaches to the monolithic integration of spatial light modulators are illustrated, differentiated primarily by the method employed to physically or electrically isolate (pixelate) the modulator elements.

Two particularly critical parameters of spatial light modulators used in neural network implementations are the contrast ratio and dynamic range of the modulator. Their values

can in certain cases be increased by incorporating the active modulation layer (for example, a multiple quantum well device) within a symmetric or asymmetric optical (Fabry-Perot) cavity [Whitehead, 1989a; Whitehead, 1989b; Whitehead, 1989c; Yan, 1989]. The asymmetric case is shown schematically in Figure 15.5(a), in which two multilayer Bragg mirrors are used to form a reflective cavity with a high reflectivity ($R$) on the substrate side, and a lower reflectivity on the air-incident side. One of several advantages of monolithic integration is the potential for utilizing common components for multiple purposes. For example, the basic MQW modulator structure can also be used as a $p$-$i$-$n$ photodetector by application of appropriate bias voltages, as shown in Figure 15.5(b). To date, significant progress in such monolithically integrated optical modulators has been achieved, although spatial light modulators with large numbers ($> 10^4$) of pixels have not yet been fabricated that exhibit the relatively high degree of integration described above.

**FIGURE 15.5** Examples of monolithically-integrated spatial light modulators. The chosen examples incorporate photodetectors, control circuitry, and multiple quantum well modulators within each pixel on a single gallium arsenide ($GaAs$) substrate. In (a), the control electronics and photodetector elements are fabricated following the photolithographic definition and physical isolation of the modulator elements, while in (b) a buffer (isolation) layer is used to allow fabrication and interconnection of all of the elements without chemical or ion beam etching.

In the *hybrid* approach, on the other hand, certain of the device functions may be integrated on a substrate within one materials system (with its associated process technology) in order to optimize either their performance characteristics or manufacturability, while others are integrated on a separate substrate within a different materials system (with a necessarily distinct process technology). Following separate processing sequences for each individual component, the two substrates are then interconnected (bonded together) such

that the mating pixels on each substrate are in pairwise electrical contact. For example, several currently investigated types of spatial light modulators (SLMs) incorporate the detection elements and control circuitry on a silicon ($Si$) substrate utilizing standard VLSI design rules, while the modulation elements are based in a separate technology (such as multiple quantum well structures integrated on a $GaAs$ substrate). Alternatively, hybrid spatial light modulators can be fabricated on a single common substrate, with additional functionality provided by the growth, deposition, or coating of a second active material onto the substrate. Examples of this type of hybrid SLM include silicon VLSI/ferroelectric liquid crystal devices [Drabik, 1990] and silicon/PLZT devices [Lin, 1990]. Such hybrid SLMs are also in the early stages of advanced development, and are the subject of current intensive research and development efforts [*Spatial Light Modulators for Optical Information Processing*, 1989; *Spatial Light Modulators and Applications*, 1990].

Using either of these two approaches to spatial light modulator fabrication, devices based on both transmissive and reflective readout can be constructed, with different implications on the overall systems design in each case. In particular, the reflective mode can be used to advantage in configuring a hybrid-integrated SLM to mate the detection and control circuitry functions of the device with the optical modulation function. Use of the reflective readout configuration allows the detection and control circuitry to be integrated on a substrate that is opaque to the readout illumination wavelength [Kyriakakis, 1990], as shown schematically in Figure 15.6.

**FIGURE 15.6** Example of a hybrid spatial light modulator, in which the photodetectors and control electronics are fabricated on a silicon substrate, and the multiple quantum well modulator elements are fabricated on a gallium arsenide ($GaAs$) substrate. The two sets of devices are bump contacted on a pixel-by-pixel basis to provide parallel electrical continuity.

As an example of the degree of functional integration currently envisioned for spatial

light modulators that are specifically designed for photonic implementations of neural networks, a silicon-based CMOS chip has recently been designed and fabricated [Asthana, 1990a; Asthana, 1990b] that incorporates two input detectors, control circuitry, and two (optical modulator) output drivers within each $100 \times 100$ $\mu m$ pixel as shown schematically in Figure 15.7(a). It should be noted that these current dimensions do not in any sense represent a lower limit, but rather a practical size for laboratory demonstrations and experiments, as well as a useful size from the perspective of neural network applications. The pixel layout allows for two $30 \times 50$ $\mu m$ detectors, followed by a 15 transistor dual input, dual output differential amplifier that implements a sigmoid-like transfer function, with externally programmable saturation characteristics. Output pads are also provided for hybrid bonding (by bump contact techniques [Shirouzu, 1986]) to an $InGaAs/GaAs$ multiple quantum well modulator structure fabricated on a $GaAs$ substrate [Kyriakakis, 1990]. Utilizing 2 $\mu m$ CMOS design rules, the control circuitry easily fits within $25 \times 100$ $\mu m$, leaving adequate space for the modulator output pads as shown in Figure 15.7(b, c). This currently allows for the integration of $10^4$ pixels per $cm^2$, or $6 \times 10^4$ pixels per $in^2$. The functional operation of this circuit will be discussed in the section, "An Implementation Strategy", below.

**FIGURE 15.7** VLSI layout of a generalizable silicon-based spatial light modulator structure: (a) neuron pixel layout; (b) photograph of a single neuron unit in VLSI implementation, with probe pads substituted for the two detectors (bottom) and for contact to the two modulation elements (top); (c) photograph of a $6 \times 6$ array of neuron units on a VLSI chip that incorporates additional test circuitry.

Before leaving the subject of photonic switching, it should be noted that the general principles outlined above can be used to design a wide variety of mutually compatible devices with different functionalities as well as different tradeoffs among the set of con-

figurational and operational parameters. For example, it is relatively straightforward to design time-integrating and time-differentiating circuits; sharp (step-like) thresholds; level slices; sigmoid-like functions, their complements, and their derivatives; inverters; and logarithmic amplifiers. Many such functions can be implemented with only a few integrated components, such as capacitors, diodes, transistors employed as current amplifiers, and biased transistors employed as resistors [Mead, 1989]. Therefore, these functions can easily be incorporated within each pixel (neuron unit) of a two-dimensional spatial light modulator, as well as in some cases between pixels for the implementation of non-local (other than pointwise) operations such as automatic gain control and nearest-neighbor inhibition.

## Photonic Interconnections

Given that the neuron units are to be represented by individual pixels within a two-dimensional spatial light modulator, interconnections must now be established between each individual neuron unit (pixel) and many (if not all) other neuron units. As such, the chosen interconnection scheme must be capable of the appropriate degree of fan-in and fan-out, be characterized by sufficient transmission bandwidth in each channel, and be scalable to relatively large numbers of neuron units. In addition, the neural network paradigm presents the additional requirement that the interconnections be *weighted*, such that the output from a given point-to-point interconnection is proportional to the product of the input and a stored constant or weight. It is in fact this requirement that eliminates from consideration a large number of possible switching networks that provide full reconfigurability in a non-blocking manner (such as a crossbar or shuffle-exchange network), but without the capacity to incorporate weights within each interconnection pathway. In adaptive networks (those that incorporate learning algorithms), these interconnection weights must have the capability of being updated in a manner determined by the particular learning algorithm employed. A nontrivial consequence of these last two requirements is that the interconnection weights must be *stored* for at least as long as the average iterative computation, if not *much* longer; yet, they must simultaneously exhibit dynamic

programmability if the network is to exhibit either short-term or long-term plasticity.

For very small numbers of neurons with a low degree of connectivity, one possible way of forming the interconnection network would be to use fiber optic transmission lines with modulated semiconductor laser diodes as sources and optical receivers as detectors, much like a fiber optic local area network. The weights could be incorporated by means of a variable gain amplifier at either end of each fiber optic link, with weight storage in local dynamic random access memory (RAM) or static read only memory (ROM) circuits. Unfortunately, the sheer bulk of each transmitter, receiver, and fiber optic link precludes scalability to large neural network systems. For example, a fully connected twenty neuron network would involve four hundred sets of sources, transmission lines, and detectors, which would currently represent a prohibitive requirement. The same would be true of a fifty neuron network with a fan-out and fan-in of eight, representing a relatively low degree of connectivity.

In order to be able to satisfy the interconnection requirements for a large number of neuron units that are fully or nearly fully interconnected, the appropriate photonic technology to employ is that of *holographic* interconnections, in which the weights as well as the interconnection patterns themselves are stored as holograms in either a fixed (static) or real time (dynamic) holographic recording material. In this section, we first discuss the basic principles that apply to the utilization of holographic recording for point-to-point interconnections. Next, we describe the physical origins of a number of complexities with holographic interconnection schemes that lead to both interchannel crosstalk and throughput losses. An architecture that lends itself to the minimization of such complications will be described in detail in the section, "An Implementation Strategy", below. Finally, the potential for incorporation of real time volume holographic recording media such as photorefractive materials in holographic interconnection networks is addressed.

The essential principle of holographic recording, that of the space-variant interference of two mutually coherent wavefronts, was discussed briefly in reference to Equation (10) and is illustrated in Figure 15.2(b). In this figure, two angularly separated (noncollinear) collimated beams are incident on a photosensitive material, such that their mutual interference locally exposes the material to the intensity distribution given by Equation (10).

In Figure 15.2(b), the photosensitive material was assumed to spatially integrate across the interference pattern, producing an output that depends on only the spatial *average* of the intensity distribution. Suppose now that we use instead a photosensitive material with the property that its *local* index of refraction or absorption coefficient depends on the *local* incident intensity (exposure), which allows the complete interference pattern to be *recorded.* The resulting change in the local optical properties of the medium may either be immediate (as in the case of a photochromic transformation, for example), or may require development following exposure (as in the case of bleached photographic negatives or dichromated gelatin thin films). Figure 15.8(a) shows such a detection or recording geometry in which a thin semi-transparent layer of photosensitive material acts as a quasi-planar holographic recording medium. The interference pattern produced by the mutually coherent signal and reference beams within the holographic recording medium is recorded to form a diffraction grating within the volume accessed by both beams simultaneously, as shown in the figure. For simplicity in Figure 15.8 (as well as in subsequent figures), we have not shown the refraction of the incident and transmitted beams at the input and output faces of the holographic recording medium that occurs due to a difference between the refractive indices of the medium and its surround. The amplitude (and intensity) of the reflected beam shown in Figure 15.8(b) depends directly on the index difference, and represents a throughput loss on readout.

**FIGURE 15.8** A simplified holographic recording configuration: case of plane wave signal and reference beams, and a *thin* holographic recording medium; (a) recording, and (b) reconstruction with a plane wave readout beam.

Consider first the case of an exposure-dependent refractive index variation. Illumination of such a space-variant modulation of the refractive index by a coherent collimated beam of the same wavelength $\lambda$ as the exposure (writing) beams will result in a diffraction pattern consisting of several collimated beams, each emanating in a characteristic direction as

shown schematically in Figure 15.8(b), and as given by the following equation:

$$\mathbf{k}_{mx} = \mathbf{k}_{rx} + m\mathbf{K}_G; \quad |\mathbf{K}_G| = \frac{2\pi}{\Lambda_G} = |\mathbf{k}_2 - \mathbf{k}_1|; \quad m = 0, \pm 1, \pm 2, \cdots \tag{15}$$

In this equation, $\mathbf{k}_{mx}$ is the $x$-component of the wave vector of the $m^{th}$ diffracted beam (diffraction *order*), $\mathbf{K}_G$ is the wave vector (assumed oriented along the $x$-axis) of the interference pattern (diffraction grating) formed by the two writing beams (with wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$), $\mathbf{k}_{rx}$ is the $x$-component of the wave vector of the incident readout beam propagating in the $x$-$z$ plane, and $\Lambda_G$ is the spatial wavelength of the diffraction grating. The multiple diffracted orders result from the phase modulation of the readout (probe) beam by the refractive index modulation $n(x)$ of the thin holographic grating; the magnitude and phase of the readout beam amplitude immediately after passing through the hologram (located at the position $z_0$) can be written in the form:

$$A_{diff} = a_r e^{i(k_{rx}x + k_{rz}z_0)} e^{i\phi_G(x)}, \tag{16}$$

in which $A_{diff}$ is the amplitude of the diffracted wavefront, $\phi_G(x) = 2\pi n(x)d/\lambda$ is the local phase shift induced by the diffraction grating (assumed to be of thickness $d$), $a_r e^{i(k_{rx}x + k_{rz}z_0)}$ is the incident readout beam amplitude at the exit plane of the hologram ($z_0$), and $k_{rx}$ and $k_{rz}$ are the $x$ and $z$ components of the wave vector $\mathbf{k}_r$, respectively. Each of the diffracted orders can then be directly associated with a corresponding Fourier component of the modulated amplitude [Goodman, 1968], which can be expanded in terms of the form:

$$A_m e^{imK_G x}. \tag{17}$$

In order to assess the effectiveness with which the holographic grating diffracts the incident beam into a particular diffraction order, it is convenient to define the *diffraction efficiency* $\eta$ of each order as:

$$\eta \equiv \frac{|A_m|^2}{|a_r|^2}. \tag{18}$$

The essential diffraction properties of thin absorption gratings (in which the modulation

occurs in the local absorption coefficient) are the same as for the case of thin pure phase gratings, with two principal exceptions: (1) for sinusoidal absorption gratings, the diffracted orders are limited to $m = -1, 0$, and 1; and (2) the presence of absorption significantly decreases the maximum first order diffraction efficiency that can be achieved.

In order to illustrate how such holographic gratings can be employed to generate weighted point-to-point interconnections, we need to introduce two additional concepts: the lens as an angle-to-position encoder, and the superposition of holographic gratings recorded with different diffraction efficiencies. The first concept can be understood with reference to Figure 15.9, in which a simple lens is placed one focal length away from a point source in the input plane of a photonic interconnection, and a second simple lens is placed one focal length away from the output plane. What is normally thought of as the focal property of a lens results in the generation of a collimated beam (a beam comprising both parallel rays and planar wavefronts) following the first lens, with an *angle* (both in and out of the plane of the page) that depends on the *position* of the point source in the focal (input) plane. In this sense, the first lens acts as a position-to-angle encoder, providing a one-to-one correspondence between the input location and the output collimated beam angle. Depending on the nature of the grating stored within the holographic optical element, the collimated beam will be diffracted into a new direction characterized by a *different* angle. The second lens will then focus the diffracted beam to a point in the output plane that depends on this angle, thus acting as an angle-to-position encoder. The utilization of different orientations of gratings within the holographic optical element allows for the interconnection of any arbitrary point in the input plane to any other point in the output plane.

**FIGURE 15.9** A point-to-point interconnection system, using a holographic optical element (HOE) for interconnection routing, and lenses as position-to-angle and angle-to-position encoders. In this example, the holographic optical element effectively performs an input angle to output angle transformation, such that light emitted (or transmitted) at point $p_1$ in the input plane ($P_1$) is detected at point $p_2$ in the output plane ($P_2$).

Suppose now that we do in fact choose to superimpose a number of planar gratings within the holographic medium, each with a different wave vector (orientation and grating period) and grating modulation (variation of the refractive index or the absorption coefficient). Assuming for the moment that the diffraction process is linear, each input point will be interconnected with a number of output points as determined by the set of recorded gratings. Likewise, each output point will be interconnected with a specified number of input points. Each interconnection will be weighted by its diffraction efficiency as determined by Equation (18), which is in turn dependent on the index of refraction (or absorption coefficient) variation recorded for each grating. As such, the holographic optical element acts as a multi-port variable beamsplitter, redirecting (diffracting) a given fraction of each input beam to a specified set of output beams. By employing lenses as described above, this feature allows the construction of a point-to-point interconnection with weights and arbitrary fan-out/fan-out (delimited only by the number of gratings recorded).

There is at least one obvious problem with the interconnection scheme outlined above, however, in that any *given* grating will connect *any* of the input points to specific output points pairwise, as shown by Equation (15). This particular feature occurs because each input point generates a collimated beam with a distinct wave vector $\mathbf{k}_r$ corresponding to a particular direction (angle) of propagation, each of which satisfies Equation (15) with a different diffracted wave vector (for each diffracted order) $\mathbf{k}_m$. The result of this degeneracy is that any recorded hologram that is designed to connect a single input point to one or more output points will in fact also connect *every* other input point to corresponding sets of output points, using the same relative interconnection pattern for each input point. This effect can be utilized to advantage, for example, in parallel digital optical computing systems with interconnection symmetry or regularity, since one simple hologram can in effect implement a very large number of point-to-point interconnections (the equivalent of wires in the case of an electronic implementations) [Jenkins, 1984]. For neural networks, however, the common requirement of nearly arbitrary (highly irregular) interconnections makes this feature undesirable.

A second problem with the proposed interconnection scheme is the presence of a multiplicity of diffracted *orders* for each diffraction grating, as shown in Figure 15.8, which occasions the connection of each input point to a number of geometrically related output points even for the case of a single stored grating.

The solution to this seeming dilemma is to extend the holographic medium into the third dimension (the direction of light propagation), creating a *volume* holographic optical element (VHOE) to take the place of the thin planar element discussed above. There are two essential properties of VHOEs that bear directly on the utilization of such elements in photonic interconnections. The first is that diffraction is limited to the first order only and all higher diffracted orders are suppressed if the holographic medium is thick enough, as defined below and as shown schematically in Figure 15.10. This occurs because each additional "layer" in the thickness direction of the holographic medium provides an additional constraint on the diffraction phenomenon; these constraints act collectively to enhance the amplitude diffracted into the first order by means of constructive interference, at the expense of the other diffracted orders.

**FIGURE 15.10** Volume holographic recording with plane wave signal and reference beams; (a) recording, and (b) reconstruction, showing the elimination of the higher diffracted orders.

The second important property of a volume holographic optical element is that of *angular selectivity*; specifically, the range of input angles that can diffract from a given grating decreases as the thickness of the grating is increased. The central angles that are allowed in the case of a thick grating are the same angles that define the two beams that initially *created* the holographic grating. This property therefore eliminates the inadvertent connection of all input points pairwise to a matching set of output points, and allows for the generation of *independent, weighted* interconnections as are desired for neural network applications.

In order to differentiate "thin" grating diffraction behavior (the so-called Raman-Nath diffraction regime) from "thick" grating behavior (the so-called Bragg diffraction regime), it is convenient to define a dimensionless "thickness" parameter $Q$ such that:

$$Q = \frac{2\pi\lambda d}{n\Lambda_G^2} \tag{19}$$

in which $n$ is the average refractive index of the holographic recording medium, and the remaining parameters are as specified previously. In general, gratings for which $Q \geq 10$ operate well within the Bragg regime, while gratings with $Q$ parameters less than unity exhibit unacceptable degrees of Raman-Nath character for truly independent multiplexed interconnection applications. The angular response characteristics of both planar and volume diffraction gratings are shown as a function of the $Q$ parameter in Figure 15.11, in which the transition from pure Raman-Nath to pure Bragg behavior for increasing values of $Q$ can be seen. Note that the number of input points that can be independently connected to an equally-sized array of output points is a decreasing function of the width of the angular response.

**FIGURE 15.11** The angular alignment sensitivity of a volume holographic optical element, as a function of the dimensionless $Q$-parameter defined in the text. The grating strength for all of the curves (3.14 radians) is optimized to produce 100% diffraction efficiency in the limit of large $Q$ (Bragg diffraction regime), and is not optimized for low $Q$ gratings. Note that the diffraction efficiency is essentially independent of angle for low $Q$ gratings, and is very strongly peaked at the Bragg angle (7.5 degrees in this case) for high $Q$ gratings.

The throughput efficiency of a volume holographic optical element as used in an interconnection application is determined to first order by the diffraction efficiency of each

individual interconnection grating, in direct analogy to the definition of the diffraction efficiency for the planar hologram case in Equation (18). For example, for the case of an unslanted pure phase grating with equiphase fronts (*i.e.* planes of constant phase) parallel to the bisector of the recording beams with wave vector $k$ and perpendicular to the entrance face of the volume holographic recording medium, the diffraction efficiency at the Bragg (optimum readout) angle is given by [Kogelnik, 1969]:

$$\eta = e^{-\alpha d/\cos\theta_B} \sin^2\left(\frac{\pi \Delta n d}{\lambda \cos\theta_B}\right), \tag{20}$$

in which $\alpha$ is the absorption coefficient of the holographic recording medium of thickness $d$ at the optical readout wavelength $\lambda$, $\Delta n$ is the amplitude of the refractive index modulation, and $\theta_B$ is the Bragg angle defined by $2k\sin\theta_B = K_G$. As can be seen from Equation (20), the diffraction efficiency of the first order for a single grating can approach 100% if the absorption coefficient satisfies the requirement $\alpha d \ll 1$, provided sufficient index modulation $\Delta n$ can be produced by the exposure process. The dependence of the diffraction efficiency on the grating strength is shown in Figure 15.12 for both thin (Raman-Nath) and thick (Bragg) pure phase diffraction gratings. The grating strength $v$ is defined as the integrated peak phase modulation of the grating in each case, and is given by:

$$v = \frac{2\pi \Delta n d}{\lambda \cos\theta_B}. \tag{21}$$

The maximum diffraction efficiency of the thin diffraction grating is about 34%, which occurs at a grating strength of 1.8 radians. Thick diffraction gratings achieve 100% diffraction efficiency at a grating strength of $\pi$ radians, at which point the diffraction efficiency of the thin grating has peaked and is nearly at its first node, as shown in Figure 15.12.

**FIGURE 15.12** The diffraction efficiency of thin (Raman-Nath diffraction regime) and thick (Bragg diffraction regime) holographic gratings as a function of the grating strength.

The extremely narrow angular alignment characteristics of volume diffraction gratings in principle allow the simultaneous multiplexing of large numbers of independent, weighted interconnections to be recorded between the input plane and the output plane (*c.f.* Figure 15.9). In addition, the use of angular multiplexing allows for both fan-out from a given input point to a number of output points, as well as fan-in from a number of input points to a single output point.

The holographic implementation of the fan-out from a single input point to a number of output points uses several multiplexed (superimposed) holographic gratings to achieve the desired weighted fan-out, one for each output point. Consider a 4 input, 4 output interconnection as shown in Figure 15.13. For each input point $x_j$ that we wish to interconnect to an output point $y_i'$, the recording process requires the pairwise coherent interference within the holographic recording medium of $x_j$ with a second beam $y_i$ corresponding to $y_i'$. The interconnection of $x_1$ to $y_1', y_2', y_3'$, and $y_4'$ therefore requires the pairwise coherent interference of $x_1$ with $y_1, x_1$ with $y_2$, and so on. This process results in the fourfold fan-out of $x_1$ to all of the outputs.

The fan-out from a single reference beam to a number of output beams is directly analogous to the readout of a traditional hologram (of, for example, a two-dimensional or three-dimensional image), provided that the full set of beams $\{y_i\}$ is coherently recorded with the given reference beam $x_j$. Although up to this point we have formulated the point-to-point holographic interconnection problem in terms of collimated (plane wave) input and output beams that record individual diffraction gratings (characterized by a single grating wave vector) within the holographic recording medium, many alternative recording and reconstruction geometries can be envisioned that produce equivalent results. In the case of traditional holography, for example, the input transparency bearing the image to be recorded is illuminated with a collimated beam, resulting in a complex diffraction pattern at the front entrance plane of the holographic recording medium. Collimated, converging, or diverging reference beams can be utilized to produce reconstructed images with a wide variety of optical imaging characteristics. Likewise, various input and output beam geome-

tries can be used in a point-to-point interconnection system to optimize the overall system characteristics, such as freedom from interchannel crosstalk, optimum use of the spatial frequency recording characteristics of the holographic recording medium, optical system complexity, and convenience of the optical layout (particularly when viewed in conjunction with associated optical subsystems).

**FIGURE 15.13** Schematic representation of a 4 input, 4 output holographic interconnection, showing 4 coherent input beams $x_1$-$x_4$ and 4 coherent recording beams $y_1$-$y_4$, each of which corresponds to a desired output $y_1'$-$y_4'$. In (a), the sets $\{x_j\}$ and $\{y_i\}$ interfere within the volume holographic medium, recording the desired interconnection diffraction gratings. In (b), a new set of input beams $\{x_j\}$ illuminates the volume holographic medium, reading out the weighted interconnection pattern and forming appropriately weighted sums at each of the outputs $\{y_i'\}$.

The fourfold fan-in of inputs $x_1, x_2, x_3$, and $x_4$ to $y_1$ can likewise be accomplished by recording each of the necessary interconnections pairwise, as before for the fan-out case. The recording process for the fully implemented 4 to 4 interconnection therefore involves the generation of 16 individually weighted diffraction gratings that connect the full set of inputs $\{x_j\}$ to the full set of outputs $\{y_i'\}$. The multiplexed hologram that accomplishes this function can be recorded in a number of ways, each characterized by certain advantages and disadvantages [Psaltis, 1988].

In the fully coherent approach, the requisite gratings can be recorded by illuminating the holographic recording medium with $\{x_j\}$ and with $\{y_i\}$ simultaneously. This can be accomplished, for example, by using a spatial light modulator to store each of the sets of values, and a pair of mutually coherent readout beams to encode these values and interfere them within the holographic element. In this manner, all of the required gratings are recorded in a single exposure; however, there are two difficulties inherent in this single

exposure, fully coherent approach. The first problem is that a fully independent $N$ to $M$ interconnection requires $NM$ stored interconnection weights, whereas the single exposure described above supplies only $N+M$ input values that can be used to generate the weights. The resulting interconnection matrix can in fact connect all of the input points to all of the output points, but the relative fan-out weights from each input point will be degenerate. One way to avoid this degeneracy is to illuminate the holographic recording medium with each input $x_j$ and a full set of corresponding outputs $\{y_i\}$, sequencing through all $N$ of the inputs (and changing the set of corresponding outputs) one at a time. This procedure generates an independent fan-out from each input point. The second problem with the single exposure, fully coherent approach is that undesirable gratings will be recorded among the $\{x_j\}$ and among the $\{y_i\}$ that can lead to considerable coherent crosstalk among the *desired* interconnection pairs. This coherent interference process diminishes the degree of independence of the interconnections.

This coherent-recording-induced crosstalk can also be avoided by sequencing the recording, but in this case each desired grating pair is recorded separately such that only one input beam $x_j$ interferes with one output beam generator $y_i$ (recording beam for the desired output beam $y_i'$) at a time. This scheme effectively eliminates the coherent crosstalk, but does not eliminate another form of crosstalk (called *beam degeneracy* crosstalk [Jenkins, 1990a; Jenkins, 1990b; Asthana, 1990c], the origin of which is described below) that can be equally severe; in addition, the complication imposed by the incorporation of such a sequential recording schedule can be a serious constraint for large $N \times M$ ($N$ input points to $M$ output points) interconnections, as $NM$ independent recording steps are required for full programming of the interconnection. This proves to be particularly problematic for the rapid generation of weight updates in a large scale neural interconnection network that incorporates synaptic plasticity. Furthermore, sequential recording of holographic exposures can cause partial erasure of previously recording interconnection weights in certain types of holographic recording materials, necessitating the use of recording schedules that attempt to balance the weights recorded at the beginning of the sequence (and hence partially erased by all subsequent exposures) with the weights recorded at the end of the sequence [Psaltis, 1988]. The use of such recording schedules usually implies an overall

decrease in both the exposure efficiency and throughput efficiency of the resulting holographically recorded interconnection matrix, as well as the buildup of noise resulting from the series of space-variant erasures.

One potential scheme for reducing coherent-recording-induced crosstalk, beam degeneracy crosstalk, and sequential recording schedules involves the use of an array of coherent but mutually incoherent sources to simultaneously expose the holographic recording medium to only the desired sets of gratings [Jenkins, 1990a; Jenkins, 1990b; Asthana, 1990c]. This scheme will be discussed in detail in a later section.

The fan-out process is illustrated in Figure 15.14, in which implementations using both beamsplitters and volume holographic optical elements are shown. The case of fan-out utilizing beamsplitters is shown schematically in Figure 15.14(a). As can be seen in the figure, the input beam can be divided among the output beams with arbitrary weights set by the transmissivities of the beamsplitting elements $BS_i$. If the final beamsplitter is a mirror, the fan-out process can be accomplished with essentially zero throughput loss. By analogy to the beamsplitter case, as well as by direct analysis, it can be proven that the holographic fan-out process shown in Figure 15.14(b) can also be accomplished with essentially arbitrary weights, with no optical throughput loss inherent in the fan-out process itself. It is interesting to note that these two implementations differ in at least one essential feature, in that the beams fanned out from the holographic implementation originate within the same volume, while the beams fanned out from the beamsplitter implementation originate from vertically displaced beamsplitters. If we were to extend the fanned out beams in the latter case backwards toward the left hand side of Fig. 15.14(a), we could imagine replacing the three discrete, vertically displaced beamsplitters with a single, multiplexed "virtual" beamsplitter that generates the same set of output beams. One physical realization of such a "virtual" beamsplitter component is in fact the multiplexed volume hologram shown in Figure 15.14(b).

**FIGURE 15.14** Schematic representation of the fan-out process for optical beams, for the case of one input and three outputs: (a) with beamsplitters

($BS_1$ - $BS_3$); (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings.

The collinear fan-in process is illustrated in Figure 15.15 for both types of implementations. As was discussed above, for the beamsplitter implementation an intrinsic fan-in loss is encountered for the case of collinear fan-in, while the intrinsic loss can be circumvented by resorting to mirrors and employing angular multiplexing. For the case of volume holographic optical elements, the situation is identical, such that collinear fan-in is grossly inefficient for large numbers of fan-in interconnections to the same node. On the other hand, appropriate use of angular multiplexing can eliminate this seemingly inherent fan-in loss, giving rise to a highly multiplexed, efficient interconnection element [Jenkins, 1990a; Jenkins, 1990b; Asthana, 1990c] as described in a later section.

**FIGURE 15.15** Schematic representation of the fan-in process for optical beams, for the case of three angularly distinct inputs and one combined collinear output beam: (a) with beamsplitters, showing the unavoidability of a throughput loss associated with the set of transmitted (and multiply reflected) beams; (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings, showing an analogous throughput loss.

The physical origin of this intrinsic optical throughput loss in the case of collinear fan-in is directly related to the mechanism that gives rise to beam degeneracy crosstalk. In Figure 15.16 we show a 4 to 4 holographic interconnection that is assumed to have been recorded by the sequential exposure technique described above in reference to Figure 15.13, in order to include all 16 individually weighted interconnection gratings but none of the undesirable gratings that can give rise to coherent-recording-induced crosstalk. In

this case, readout by the input beam $x_1$ generates the four output beams $y_1'$ through $y_4'$, with values given by the stored interconnection weights $w_{ij}$:

$$y_i' = w_{i1}x_1. \tag{22}$$

Within the holographic medium, however, each of the four output beams can in turn act as an *input* beam, generating undesired output beams in the directions $x_2', x_3'$, and $x_4'$. These undesired output beams are a result of diffraction from the gratings recorded between each output generating beam $y_i$ and the full set of input beams $\{x_j\}$. Each output beam is automatically Bragg matched (at the correct Bragg angle) to the full set of input beams due to the collinear recording geometry employed. We refer to the fan-in as *collinear* in this case because each input beam $x_j$ that is fanned in to a given output $y_i'$ produces an output beam in the *same* direction. The generation of diffracted intensity in the directions $x_2'$-$x_4'$ from readout with $x_1$ results in a throughput loss for the interconnections between $x_1$ and the set of output beams $\{y_i'\}$. In addition, the throughput losses of the individual output beams $\{y_i'\}$ will not be equal in general. Furthermore, the undesired diffracted beams $x_2'$-$x_4'$ can *also* act as input beams, generating additional output beams in the directions $\{y_i'\}$ that coherently interfere with the beams directly diffracted in those directions by the input beam $x_1$. The combination of interconnection-dependent losses from the output beams $\{y_i'\}$ into the "cross-coupled" beams $\{x_i'\}$, and of interconnection-dependent coupling from $\{x_i'\}$ into $\{y_i'\}$ gives rise to an undesired redistribution of the intensities of the output beams. This phenomenon is referred to as *beam degeneracy* crosstalk, as it arises from the beam direction degeneracy (collinearity) of the output beams fanned into a single output point.

**FIGURE 15.16** Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: recording/readout configuration. The input beams $\{x_j\}$ are assumed to have interfered within the volume holographic medium with the set of recording beams $\{y_i\}$, producing

the desired set of interconnection gratings with weights $w_{ij}$. Illumination of the volume holographic medium with beam $x_1$ produces a 1 to 4 fanout into the output beams $\{y_i'\}$, as well as the zeroth order beam $x_1'$. Due to the effects of beam degeneracy, power is also coupled into the zeroth order beams $x_2'$-$x_4'$, and crosstalk terms $\{c_i\}$ are introduced into the outputs.

Both the throughput loss and the beam degeneracy crosstalk that characterize holographic interconnection geometries with collinear fan-in can be estimated by numerical simulation of the diffraction process from a multiplexed grating [Asthana, 1990c]. By using the optical beam propagation method [Johnson, 1986] to simulate the diffraction process, we can analyze the 4 to 4 interconnection described above for the case of a single beam readout, as shown in Figure 15.16. The results of such an analysis are presented in Figure 15.17, which shows the diffraction efficiency of each of the four beams fanned out from the single input point, as well as the three cross-coupled beams in the directions $\{x_j'\}$ and the zero order (undiffracted) beam. For this illustration, all 16 interconnection weights are equal in magnitude. As the grating strength is increased, a significant amount of intensity is coupled into the cross-coupled components, robbing the desired fan-out beams of the desired diffraction efficiency. In addition, the *relative* diffraction efficiencies observed in the designated fan-out beams are no longer independent of the grating strength, as desired in a fully independent weighted interconnection. Extensive modeling of $N$-to-$N$ holographic interconnections with collinear fan-in suggests that the throughput loss increases approximately as $1/N$, which is potentially catastrophic for large interconnection networks. In a later section, we will describe an alternative holographic recording approach that obviates this $1/N$ loss.

**FIGURE 15.17** Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: diffraction efficiency as a function of grating strength for the readout configuration of Figure 15.16. Shown

are the depletion of the zero order beam $x_1'$ and the rise of the desired output beams $y_i'$, accompanied by a strong buildup of the cross-coupled beams $x_2'$-$x_4'$.

The development of a viable photonic interconnection technology is based in no small part on the availability of appropriate photosensitive recording materials [Psaltis, 1988; Smith, 1977; Gunter, 1988; Gunter, 1989]. Many interconnection demonstration experiments have been performed in the laboratory on bleached photographic emulsions and dichromated gelatin films, both of which are thick enough (10 - 30 $\mu m$) to exhibit sufficient Bragg-like diffraction behavior to allow a limited degree of multiplexing to be incorporated. Neither material, however, exhibits capacity for real time operation, which is essential for the implementation of photonic neural networks with at least some degree of synaptic plasticity. On the other hand, one principal advantage of photographic film and dichromated gelatin is their essentially infinite read-write asymmetry, which is highly desirable in many applications as described below.

By *real time operation*, we mean that the holographic interconnections can be programmed (exposed) and used (read out) on roughly the same time scale (perhaps at $kHz$ frame rates), without the necessity of chemical development processes or the like. By *read-write asymmetry*, we mean that the readout of a programmed interconnection should not erase the stored weights at an accumulated readout exposure equal to that of the recording exposure. Ideally, we would like to have the capability of exposing the holographic interconnection to the recording beams with essentially instantaneous "development" of the stored gratings, with the recording process characterized by very high sensitivity during the "learning" process. At the same time, we would like to be able to initiate readout of the stored interconnection pattern without altering the stored weights for a length of time equal to the desired "computation" time. Although in many applications the learning and computation times may differ by only an order of magnitude, in other cases it is desirable to compute for very long times compared with the learning phase, and yet still maintain the capacity for (slowly varying) weight updates.

The class of photosensitive recording materials that has been most extensively investi-

gated for photonic interconnection applications does in fact have the capacity for sensitive holographic recording, is available in "thick" samples that allow for the formation of Bragg-regime diffraction gratings, exhibits a high multiplexing capacity, and allows for the inclusion of modest read-write asymmetries. This class is that of the so-called "photorefractive" materials [Gunter, 1988; Gunter, 1989], which includes single crystals of semi-insulating optical materials such as bismuth silicon oxide ($Bi_{12}SiO_{20}$), bismuth germanium oxide ($Bi_{12}GeO_{20}$), lithium niobate ($LiNbO_3$), strontium barium niobate ($Sr_{1-x}Ba_xNb_2O_6$), potassium niobate ($KNbO_3$) and barium titanate ($BaTiO_3$), as well as semi-insulating semiconductors such as gallium arsenide ($GaAs$), indium phosphide ($InP$) and cadmium telluride ($CdTe$). The use of the term "photorefractive" to describe these materials exclusively is somewhat misleading, in that many other classes of materials are known to undergo a refractive index change following illumination as well as those traditionally included in the class described above. But at least the term is descriptive of the basic phenomenon involved, as outlined below.

In photorefractive materials such as bismuth silicon oxide, exposure to an interference pattern at an appropriate wavelength (characterized by significant photosensitivity) generates free charge carriers (electrons or holes) liberated from deep traps. The number of photogenerated carriers is in general proportional to the local intensity absorbed by the crystal; as such, the photogenerated carrier population mimics the exposure pattern in both amplitude and phase. The photogenerated carriers are free to diffuse to regions of lower intensity, or they can be assisted out of the brightest regions by application of a bias electric field to produce carrier drift. In either case, they tend to be retrapped, in turn creating a space charge distribution that has the same spatial frequency as the interference pattern. This space-variant space charge distribution produces a locally modulated electric field with the same spatial frequency (as determined by the grating spacing or grating wavelength), which in turn induces a local change in the refractive index of the photorefractive material through the linear (Pockels) or quadratic (Kerr) electrooptic effect [Kaminow, 1974]. The refractive index grating can then be probed by a readout beam to generate a diffracted beam, just as in the case of the pure phase gratings described previously.

An excellent set of review articles on the physical properties and applications of photorefractive materials has been assembled by Gunter and Huignard [Gunter, 1988; Gunter, 1989]. The state of the art is such that 1 $cm^3$ crystals of many of these materials have been grown, and shown to exhibit a very high degree of optical quality. Exposure sensitivities vary widely, but several crystals require of order 500 $\mu J/cm^3$ for full exposure to saturation (the highest grating strength that can be achieved in that particular crystal). This corresponds to the absorption of about 50 $mW/cm^2$ of optical intensity throughout 1 $cm^3$ of material for an exposure period of 10 *msec.* The range of spatial frequencies that can be supported in these materials ranges from a few lines/$mm$ to over 2000 lines/$mm$. Diffraction efficiencies close to 100% have been observed in several types of crystals, while others saturate nearer to 10% for thicknesses of order 1 $cm$.

Optimization of photorefractive materials for interconnection device applications is under way, including the development of growth processes for large photorefractive crystal boules with a high degree of optical uniformity; the characterization of both unintentionally incorporated impurities and intentionally incorporated dopants that alter the holographic recording, readout, and storage characteristics; the use of applied d.c. and a.c. bias electric fields to enhance the holographic recording sensitivity; the use of polarization effects to enhance the reconstructed image signal-to-noise ratio; and the antireflection coating of the (typically high index) front and rear surfaces to increase the diffraction efficiency and avoid the presence of unwanted gratings due to multiple reflections [Karim, 1988; Karim, 1989a; Karim, 1989b]. In addition, the origin of electric field nonuniformities that occur within photorefractive crystals during grating recording is under active investigation, and several methods of eliminating the field collapse have been discovered [Herbulock, 1988]. Use of these methods increases both the saturation diffraction efficiency and the grating response time, resulting in more efficient interconnection devices that operate at higher recording sensitivities.

## Sources and Source Arrays

In reviewing a large fraction of the journal articles published over the past decade on optical information processing and computing, including the most recent coverage of photonic implementations of neural networks, you will be inspired perhaps by the cleverness of a particular proposed architecture, or intrigued by the novel features of a particular device structure. But you will also be amazed at the apparent lack of emphasis on certainly one of the most fundamental components in any proposed photonic computational system: the source of the light! This oversight may be caused in part by direct analogy to the situation in VLSI electronics, in which it is a bit unglamorous (and probably also to a certain extent unnecessary) to concentrate on the battery or the power supply. After all, electrical power is relatively inexpensive, widely available, well characterized, and reasonably abundant. At peak usage, your home probably uses about 10 $kW$, most of which is dissipated in the air conditioner.

However, the situation in photonic technology is quite different. Sources of coherent optical radiation that can produce average output powers in the 10 $kW$ range exist in only a few laboratories, are very large (about 15 $m^3$), usually emit in the far infrared (10.6 $\mu m$), and are far from inexpensive. Incoherent sources in the range of 100 - 1000 $W$ are available (xenon-mercury ($Xe$-$Hg$) gas discharge lamps, for example), but this type of source is typically noisy (exhibits large intensity fluctuations), difficult to collimate, and characterized by a very short lifetime (from the systems perspective). In addition, gas discharge lamps are broadband sources, and as such usually require wavelength filtering in order to provide compatibility with wavelength sensitive devices such as volume holographic optical elements and spatial light modulators. A broadband source that has been suitably filtered to allow readout of a typical volume holographic optical element (within the allowable spectral bandwidth of the stored diffraction gratings) might generate only about $10^{-5}$ - $10^{-6}$ of its total rated power in the wavelength region of interest. For the 1000 $W$ $Xe$-$Hg$ lamp, this results in only about 1 - 10 $mW$ of quasi-monochromatic optical power.

Coherent, monochromatic optical power can be provided by an array of different types of laser sources [Milonni, 1988], including the argon-ion ($Ar^+$) laser, the neodymium-YAG (*Nd-YAG*) laser, the helium neon (*He-Ne*) laser, the helium cadmium (*He-Cd*) laser, dye lasers, excimer lasers, and semiconductor laser diodes. Typical monochromatic (single laser line) power outputs from the first two types range from about 500 $mW$ to 25 $W$. Helium neon and helium cadmium lasers are readily available as well as relatively inexpensive, but have output powers that are typically in the range 1 - 5 $mW$, peaking out at about 50 $mW$. Dye lasers are often optically pumped by argon-ion lasers, and hence exhibit power outputs slightly lower than that of the pump laser. Excimer lasers are typically operated in the pulsed mode of operation at repetition rates of 10 - 1000 pulses per second, and emit average powers in the 10 - 100 $W$ range. Finally, semiconductor laser diodes are available with very long lifetimes at output powers of 1 - 20 $mW$, and much shorter lifetimes in the 100 $mW$ - 1 $W$ range.

Of these six different types of coherent sources, the first five are still relatively bulky (about 0.1 $m^3$), consume considerable electrical power, generate significant amounts of heat (many must be water cooled to ensure stable operation and practical lifetimes), and are very expensive (especially when compared with a comparable electronic power supply!). Although these sources can be (and indeed are) employed in current systems-level demonstrations, their collective liabilities do not augur well for their eventual incorporation in commercially viable computational systems in general, and perhaps neural network applications in particular. This leaves the last category, that of semiconductor diode lasers (including, possibly, miniaturized diode-pumped *Nd-YAG* lasers), for further consideration.

Before discussing the properties of semiconductor diode lasers as optical power sources any further, we should at least note that the range of output powers available from these sources (1 - 100 $mW$ for single element devices) is rather limited. Taking an upper bound (with continued research and development) of about a watt per device gives us a realistic estimate of the amount of average coherent source power available for at least circuit level implementation of photonic neural networks, though certainly at the systems level phased arrays of stripe laser diodes and/or multiple sources could conceivably be employed.

Semiconductor diode lasers [Kressel, 1977; Casey, 1978a; Casey, 1978b] have been ex-

tensively investigated and developed over the past two decades for a broad range of commercial applications, including compact disk player recording and readout, fiber optical communications systems [Jones, 1988], merchandise optical scanners, and laser printers. The physical size of these lasers is small enough (about $0.3 \times 1 \times 5$ $mm$) to fit in a standard transistor (or IC) package, as long as external cooling is not required. Lasers with power outputs of 1 - 10 $mW$ are relatively inexpensive, costing a few tens of dollars in quantity on the average. Higher output power lasers are considerably more expensive, however, as are lasers with very narrow spectral linewidths (so-called *single longitudinal mode* lasers). For the higher power lasers (as well as for the intermediate power lasers that are required to maintain a high degree of center wavelength accuracy), external cooling (*e.g.* by means of a thermoelectric cooler) must be provided in order to maintain thermal stability in both wavelength and output power.

The wavelength ranges spanned by semiconductor diode lasers are dictated by the direct bandgap materials used to fabricate the coherent light-emitting diode (semiconductor $p$-$n$ junction). Aluminum gallium arsenide/gallium arsenide *(AlGaAs/GaAs)* lasers grown on single crystal gallium arsenide substrates emit at wavelengths in the range 780 to 900 $nm$, while lasers based in the quaternary indium gallium arsenide phosphide (*InGaAsP*) compound semiconductor system (and grown on indium phosphide substrates) emit at wavelengths further into the infrared (1.2 to 1.6 $\mu m$). The aluminum gallium arsenide/gallium arsenide lasers in particular are nearly wavelength matched to the peak sensitivity of both silicon and gallium arsenide photodetectors, as might be employed for photonic switching in spatial light modulator arrays, or for detection of computed results in a system diagnostic or output plane.

Within these ranges, a typical multimode semiconductor diode laser has a spectral bandwidth of 0.5 - 2 $nm$; a single longitudinal mode laser has a much narrower spectral bandwidth of order $10^{-4}$ $nm$ (about 50 $MHz$ centered at an optical frequency of $3.5 \times 10^{14}$ $Hz$). Both multimode and single longitudinal mode diode lasers can be used to write and read holographic optical interconnection elements, as long as the coherence length of the laser is larger than the thickness of the holographic recording medium. The coherence length of a laser is essentially the maximum path difference over which two

beams derived from the same laser can maintain the stable phase relationship necessary to exhibit an interference pattern. In applications requiring high multiplexing capacity within the holographic interconnection medium (or significant path differences among beams that must coherently interfere), the narrower linewidths of the single longitudinal lasers are often preferable since their coherence lengths are several orders of magnitude longer. For example, typical multimode semiconductor diode lasers operated above threshold exhibit coherence lengths in the range 0.1 - 10 $mm$, while stabilized single longitudinal mode diode lasers can have coherence lengths exceeding 1 $m$.

Employing a single, high intensity optical power source in a typical neural network application carries with it a potential penalty: an inherent tradeoff between energy efficiency on the one hand, and the need for array generation optics on the other. This tradeoff arises from the fact that most optoelectronic implementations of neuron unit arrays have either photodetectors or modulation windows (in some cases both) that are smaller in size than each individual pixel, as was shown schematically in Figures 15.4 and 15.7. The ratio of the area of a given photosensitive element to the entire pixel area is referred to as the *fill factor* of the pixel (with respect to that particular element). Typical fill factors for the photodetectors and modulation windows may range from less than 0.1 in the case of monolithic integration to about 0.5 for hybrid integrated devices. Light that falls outside the appropriate areas within a given pixel will at best contribute to the overall system throughput loss, and at worst may adversely affect the function of adjacent devices that exhibit photosensitivity.

In order to efficiently channel the optical illumination to the correct photosensitive regions, we need to (a) *expand* the source illumination uniformly to fill the entire aperture of the device in question (a spatial light modulator or volume holographic optical element, for example), (b) in many cases *collimate* (or re-collimate) the light source to produce a planar wavefront with a beam of constant width, (c) *spatially filter* the beam to enhance its uniformity by eliminating significant fixed-pattern noise, (d) *focus* the light within each individual pixel to a size compatible with the relevant photosensitive area (in effect thereby generating a two-dimensional array of focused beamlets), and (e) *align* the resulting array of focused beamlets with each succeeding device in the optical path.

The procedures and optical elements required for beam expansion, collimation, and spatial filtering are well understood among the optical community for the case in which the source beam is initially *axially symmetric*, as is typical of gas and excimer laser systems. In typical semiconductor laser diodes, however, the planar nature of the light-emitting heterojunction region often gives rise to a diffraction-induced beam divergence *parallel* to the junction of 3 - 10 degrees, and a corresponding beam divergence *perpendicular* to the junction of 20 - 60 degrees. Comparable procedures and optical elements for such *anamorphic* (non-axially symmetric) beams are more complex, and are currently under development. Also under development are a number of types of semiconductor diode lasers that emit approximately axially symmetric beams suitable for standard collimation and filtering systems.

The optical source array generation problem has received considerable attention recently, due primarily to significant interest in optical interconnection systems. In one promising approach, a two-dimensional array of computer generated and photographically reduced amplitude-encoded Fresnel zone plates has been used to form an $8 \times 8$ grid of microlenses that function by means of *diffraction* (from what is, practically speaking, a computer generated hologram (CGH)) rather than *refraction* [Marrakchi, 1990]. In another well-developed approach, computer generated binary phase holograms (so-called *Dammann* gratings [Dammann, 1971) have been configured to form large grid patterns of regularly spaced illuminated spots with predetermined locations and fill factors [Morrison, 1989]. Using this latter technique, $32 \times 32$ arrays have been generated with both high throughput efficiencies and low scattered light by crossing two fabricated $1 \times 32$ grating arrays. In addition, an $81 \times 81$ array has been experimentally demonstrated by using two pairs of crossed $1 \times 9$ grating arrays in an optical arrangement that generates multiple images by means of a convolution operation [McCormick, 1989]. In both of these techniques, all of the resulting light beamlets are mutually coherent, as they derive from the same source. This mutual coherence has an impact on the utilization of such source arrays for the generation of independent holographic interconnection networks, as described in the subsection on "Photonic Interconnections" above.

An interesting alternative to the generation of pixelated optical sources by modification

of the properties of a *single* source is that of direct fabrication of *multiple source arrays.* One striking example is the recent successful fabrication of over one million independent surface-emitting semiconductor diode lasers on a single gallium arsenide chip [Jewell, 1990]. Both cylindrical and square cross-section microlasers have been fabricated with diameters and edge dimensions in the range 1 - 5 $\mu m$, with heights above the surface of the wafer of about 5.5 $\mu m$ as shown schematically in Figure 15.18. In the fabrication process employed, the laser mirrors are arranged to generate laser emission *through* the 500 $\mu m$ thick gallium arsenide substrate, as shown in the Figure. In order to accomplish this without significant absorption in the substrate, the active (lasing) medium is composed of $InGaAs$ quantum wells with $GaAs$ barriers, giving rise to an emitted infrared wavelength ($\approx$ 950 $nm$) that lies in a region of substrate transparency.

**FIGURE 15.18** Illustration of a surface-emitting laser diode source array [after Jewell, 1990]. In this example, the individual semiconductor laser diodes are isolated by chemically assisted ion beam etching techniques, must be individually contacted, and emit *through* the $GaAs$ substrate.

In the present configuration, the lasers are essentially optically isolated, and hence are not designed to be mutually coherent (phase locked). In fact, over time constants typical of holographic recording in currently available photorefractive crystals (milliseconds), it is likely that such arrays are for all practical purposes *mutually incoherent*, due both to the optical isolation as well as to process-induced variations in device parameters that alter the wavelength emitted from each individual laser. Arrays of surface-emitting semiconductor lasers that have been specifically *designed* to have uniformly separated wavelengths have also been demonstrated [Chang-Hasnain, 1990]. We shall return to this characteristic in a succeeding section that addresses a particular strategy for photonic neural network implementation.

At present, each laser within the array operates at a threshold voltage of about 10

volts at a threshold current of a few milliamperes, resulting in a power dissipation of 10 - 50 milliwatts per device at threshold, and higher for power outputs significantly above threshold. In order to keep the overall power density within established limits (1 - 10 $W/cm^2$) and thus to keep from overheating the substrate (resulting in potentially deleterious effects on wavelength stability and/or catastrophic failure), the lasers must either be spaced appropriately, operated in a pulsed (on/off) mode at less than unity duty cycle, or temporally multiplexed (turning on only a few lasers at a time) by resorting to individual rather than parallel addressing. Given the current rate of progress in the development of these and other types of surface-emitting laser arrays, it is reasonable to expect demonstration of continuous operation of up to $10^4$ microlasers per square centimeter within the near future.

It should be noted that the array shown in Figure 15.18 is not currently configured for parallel operation of all of the sources simultaneously, which would require electrical contact to the tops of each selectively etched microcavity. This feature could likely be provided by an additional surface passivation and metallization step. Matrix-addressable surface-emitting laser arrays have recently been fabricated by forming columns of lasers separated by etched isolation grooves, and interconnected across the grooves by striped row contacts [Orenstein, 1990a]. Application of an appropriate bias voltage across a given pair of electrodes (column and row) activates the laser diode at the intersection, allowing for raster-scanned operational modes as well as fully parallel operation [Von Lehmen, 1990].

Other currently investigated approaches to surface-emitting laser array fabrication use various techniques to form the microlaser cavities *within* the planar substrate without the need for deep etched isolation grooves, such as by the use of ion implantation to form electrically insulating isolation layers between the laser cavities [Tai, 1989a; Orenstein, 1990b] or by the current confinement that results from photolithographic definition of one of the two laser mirrors and its associated electrical contact [Tai, 1989b]. Fabrication processes that yield planar or quasi-planar device structures allow for direct parallel contact if desired without the complications of depositing contacts on vertical sidewalls.

Before leaving the subject of semiconductor laser diodes and surface-emitting laser arrays, it is worthwhile to note a very useful feature of such devices: their capacity for *high*

*bandwidth direct modulation.* By this we mean that the output intensity of the semiconductor laser source can be modulated (at full modulation depth, *i.e.* from well below the threshold for lasing to peak output power) at frequencies up to a few gigahertz by direct variation of the voltage applied across the device. This attribute can be used to advantage in many neuro-optical implementation architectures by eliminating the need for mechanical or electrooptical shutters, as well as by offering temporal multiplexing as an additional degree of freedom for the systems designer.

One additional type of solid state device that is capable of both single source and source array fabrication is the light emitting diode (LED). Closely related to the semiconductor laser diode, the LED is also a *p-n* junction device that can be fabricated with considerably less processing complexity by elimination of the high reflectivity mirrors that form the semiconductor laser cavity. An additional advantage is the lack of a threshold for operation, allowing the LED to emit over a much wider dynamic range of applied voltages. One drawback of light emitting diodes is that they are relatively broadband (incoherent) sources, and as such are not usable as sources for holographic recording applications (and in many cases for readout of multiplexed holographic optical elements as well). In addition, they are relatively inefficient emitters with typical electrical-to-optical conversion efficiencies of a few percent. This feature tends to make LEDs rather power consumptive for a given amount of usable output intensity.

## Detectors and Detector Arrays

Detectors are optoelectronic components that act as photon-to-electron converters, in that they transform incident optical intensity into electronic form, usually a voltage or a current. Detectors therefore allow the optical representation of neuron unit outputs, for example, to be converted into an electronic representation for further processing. As such, they are important components for the photonic implementation of neural networks in at least two functional areas: (a) as input transducers for the necessary optical detection function of optically addressed spatial light modulators, and (b) as output transducers

for the translation of optically generated intermediate and final results to an appropriate electronic format. After all, once you've gone to all of the trouble of learning and computing with a neural network, it might prove worthwhile occasionally to actually get the answer out and use it to initiate some other useful process!

In both of these functional areas, we can further categorize detectors as (a) single pixel detectors, and (b) interconnected detector arrays. In the first category, we include both single element detectors that have one optical input aperture and one output channel, as well as the single pixel detectors employed as part of an array in two-dimensional spatial light modulators. This latter assignment is made because even though detectors used in spatial light modulators are perhaps *configured* in an array, their outputs are used only within one or at most a few local pixels. In the second category, we include arrays of detector elements that are interconnected in such a way that the *entire* parallel (one- or two- dimensional) array can be read out electronically through one or more output channels. An example of a detector array in this category might be the light sensitive element in the CCD (charge-coupled-device) camera, now commonplace in many solid state cameras and video cassette recorders.

This distinction between single pixel detectors and detector arrays is important because the technologies that are commonly employed in these two cases differ in a number of respects, and as a result can exhibit wide differences in performance characteristics such as bandwidth, sensitivity, linearity, and dynamic range. In the case of single pixel detectors, for example, it proves easier to jointly optimize performance parameters because of the larger number of degrees of freedom available to the device designer in a single input, single output system. The detector array designer, on the other hand, often must make additional tradeoffs dictated by the nature of the charge storage and readout process employed over the full set of integrated pixels.

In the context of photonic neural network implementations, single pixel detectors have two primary functions. The first is to act as optical signal to electronic signal converters within optically addressed spatial light modulators, to translate a pixel's worth of incident light intensity (representing, for example, the weighted sum of signals from the output of a plane of neuron units) into a voltage or current. The resultant electronic signal can then

be processed by local intrapixel circuitry to produce the desired neural threshold function for subsequent optical encoding (modulation). This process could be accomplished either onboard a monolithic or hybrid integrated optically addressed spatial light modulator (OASLM), or on a separate detector chip that interfaces with an electrically addressed spatial light modulator (EASLM). In this latter case, the detector will most likely fall under the *detector array* category discussed further below, since a parallel-to-serial conversion is typically required to extract the array of data (*e.g.*, an image) from the detector chip (followed by a serial-to-parallel conversion to load the signal into the EASLM). It should be noted that even in the case of monolithic spatial light modulators that do not feature discrete detectors, electronic control circuitry, and modulators, converting a two-dimensional optical input distribution into a modified two-dimensional output distribution *necessarily* involves a local detection function, even if it is not particularly easy to separate the detection process from the modulation process.

The second important single pixel detector function is to provide for single point monitoring functions within the system, such as the output power from a given laser source, the average power emitted from a laser source array, or a particular system output that activates a desired process or function (for example, the identification of a specific defect pattern on a manufactured part within the input image field of a neural image processor, that in turn results in rejection of the part).

Perhaps the simplest type of detection element that can be incorporated in a single pixel is the *photoconductor*, which typically consists of a thin film of material that alters its resistance to electrical current in response to the intensity of incident illumination. The most commonly used single pixel photodetectors, however, are based in some way or other on the *semiconductor p-n junction diode*. Under reverse bias in a *p-n* junction diode, photocarriers created by light absorbed within the region of the junction between *n*-type and *p*-type semiconductor layers are swept *out* of the junction region by the internal electric field across the junction, and collected in the external circuit. If the internal electric field is high enough, each photocarrier can acquire enough energy during sweepout to generate an avalanche of additional carriers, leading to significant gain in the class of so-called *avalanche photodiodes*.

The inclusion of an intrinsic (undoped or compensated) layer of semiconductor material between the $n$-type and $p$-type layers allows for a significant reduction in the junction capacitance of the device, with a corresponding improvement in signal bandwidth. Such devices are commonly referred to as $p$-$i$-$n$ photodiodes, packaged versions of which are commercially available for a wide variety of photosensor functions. Typical $p$-$i$-$n$ photodiodes exhibit risetimes of a few nanoseconds, are linear in output over seven orders of magnitude of input intensity, and are sensitive to very low light intensity levels. For silicon $p$-$i$-$n$ photodiodes, sensitivities of about 0.4 milliamperes of output current per milliwatt of optical input power at a wavelength of 830 $nm$ are common, which represents a conversion efficiency from photons to electrons of approximately 60%.

*Phototransistors* are light sensitive devices that exhibit current gain in exactly the same manner as a transistor, with the exception that the controlling base current is injected *optically* rather than through the base lead. In fact, most VLSI transistors (both bipolar and MOS) are photosensitive (though perhaps not optimized for the photodetector role), and must be protected from stray light in order not to compromise their performance characteristics. The principal advantage of a phototransistor is its inherent current gain of order 100 to 1000, which often makes the interface of the photodetector to following circuitry more straightforward. In cases requiring exceptionally high gain in the front (photodetection) end, two transistors can be paired as shown in Figure 15.19 so that one acts as a phototransistor, and the other as a current amplifier. Such a two transistor combination has achieved widespread use, and is referred to as a *photo-Darlington pair* [Sze, 1981b]. The tradeoffs for increased gain in both of these cases are risetime (which translates directly into signal bandwidth) and area required for integration. Typical risetimes for phototransistors are almost three orders of magnitude higher (a few microseconds) than those characteristic of $p$-$i$-$n$ photodiodes. Photo-Darlingtons are yet another factor of ten or so slower in response time. Optimized phototransistors and photo- Darlingtons require relatively large collector-base junctions in order to provide an appropriately sized photosensitive region that can be accessed by optical imaging techniques.

**FIGURE 15.19** Schematic diagram of a photodarlington pair utilized as a high gain detector/amplifier combination.

In many if not most cases, the type of photodetector chosen for use as a single pixel detector in a spatial light modulator application depends on its integrability with associated control electronics and molulation elements. This, in turn, depends on whether the particular spatial light modulator in question is monolithically or hybrid integrated, as discussed in the section on photonic switching above, and on which semiconductor substrate the photodetection element itself is to be fabricated. In some cases, the desire for integration of a high density of neuron units may place strict bounds on the area allocated to each separate function in general, and on the photodetection and requisite amplification function in particular.

In traditional applications of photodetector technology, for example in spectroscopy and optical metrology, linearity of response (output voltage or current as a function of the input intensity) is prized, as is a wide dynamic range over which linearity is assured. In neural network applications, however, linearity is typically less of an issue. In fact, it is often convenient to *use* the inherent nonlinearity of the input-output characteristic of a particular photodetector device to generate part or all of the nonlinearity required of the overall neural unit function. This can result in a lower overall expenditure of real estate for each neuron unit, increasing the neuron array density, as well as in a reduction of circuit complexity within each pixel. One such example is the output current saturation characteristic of phototransistors at high input intensities, which can be used to emulate the upper saturation regime of the sigmoidal neuron response function.

Detector arrays are employed whenever the intensity distribution of a one- or two-dimensional image field requires conversion to electronic form for interface with succeeding computational or output stages of the system. In a very real sense, a two- dimensional detector array is nothing more than the business end of an optoelectronic *camera* that can be positioned anywhere within the optical system that the local intensity distribution represents a desired result. In fact, low reflectivity beamsplitters can be used to merely

"sample" the local intensity distribution of a given beam of light, allowing most of the incident light to propagate in a further computational arm of the optical train for use elsewhere.

Detector arrays are inherently different in at least one key respect from the single pixel photodetectors (as well as arrays of photodetectors used in optically addressed spatial light modulators) discussed previously: the need to provide for some form of output channel multiplexing, in order to avoid the requirement for a one-to-one correspondence between pixels in the array and output pins. For example, in a $1000 \times 1000$ element detector array, fully parallel readout requires one *million* output channels or pinouts. As a result, detector arrays are usually configured to perform some form of parallel-to-serial conversion of the data into a single high bandwidth serial channel prior to the readout of each frame (though multiple output channels can also be used). This can either add significant circuit complexity to the area surrounding each pixel in order to accommodate for the parallel-to-serial conversion and interpixel communication function, or be directly incorporated into the design of the photodetection elements themselves, as in the case of the CCD arrays discussed below.

The state of the art of detector arrays has advanced tremendously even over the past decade, to the point where solid state detector arrays with quite spectacular performance are used everywhere from astronomical applications (as detectors for even the largest telescopes), to earth observation satellites (infrared focal plane arrays), to photomicroscopy (in place of the traditional film-based photographic camera), to consumer products (electronic still photography and video cameras).

One of the most successful and generally available types of solid state detector array is the *charge-coupled-device* (or CCD) array [Sze, 1981c; *Optical Engineering*, 1987a; *Optical Engineering*, 1987b]. In this technology, usually based on MOS fabrication techniques in silicon (but adaptable to compound semiconductor substrates as well), incident illumination within a given pixel causes the accumulation of photogenerated charge in an electrostatic potential well formed by the application of bias voltages on a set of electrodes, as shown schematically in Figure 15.20. In operation, the CCD array is illuminated for a given exposure time (slightly less than one full frame interval), during which time the

charge generated by the incident illumination is integrated within each primary well. Subsequently, appropriate voltages are applied by means of multi-phased electrode structures to *spill* the accumulated charge packet into the neighboring well, while simultaneously moving the charges in the neighboring well to *its* neighboring well, and so on throughout the array.

**FIGURE 15.20** Schematic diagram of a charge coupled device (CCD) photodetector array fabricated on a silicon substrate. Electrostatic potential wells are created by application of appropriate voltages to the three phase bias electrode structure, with electrical isolation provided by the gate oxide layer. Light incident through the transparent electrodes creates stored charge that can be transferred to an output signal terminal by proper sequential phasing of the bias voltages ($P_1$ - $P_3$).

The overall operation resembles the function of an array of one-dimensional shift registers. At one edge of the structure, the charge packets from each row are collected into a single column that is read out by a very high speed shift register (a linear, usually buried channel CCD array) to form the single output channel. Full readout of the array must occur before the next frame is exposed (except in specifically designed cases such as the time-delay-and-integrate or TDI mode of operation, in which only *one* shift is interposed between successive exposures).

One-dimensional arrays of CCD elements have been successfully fabricated in sizes of 1 × 2048, while special purpose two-dimensional CCD imaging arrays 2048 × 2048 in size are commercially available [Blouke, 1987]. This represents a parallel detector with 4,194,304 individual pixels! In one particular 2048 × 2048 CCD array, the imaging area is 5.5 × 5.5 centimeters, with a pixel size of 27 × 27 microns. This array exhibited a dark (unilluminated) noise buildup in each pixel of only 6 to 12 electrons when read out at a rate of 50,000 pixels per second, which allows for detection of extremely low level signals

with excellent signal-to-noise ratio. Given a well capacity of about 700,000 electrons, this very low noise figure suggests a dynamic range in excess of 70,000, nearly five orders of magnitude! For well charge densities less than 200,000 per pixel, the linearity is better than 0.5% over this portion of the full dynamic range. Finally, this array exhibited an extraordinarily high charge transfer efficiency of 0.999992, representing the fraction of charge within a given pixel that is routinely transferred to an adjacent pixel without loss.

The integration of large scale detector arrays by means of VLSI techniques provides the prospect of special purpose arrays that perform part of the computational function *within* the confines of the array. One example of such special purpose chips is the incorporation in a CCD array of charge-coupled analog circuitry to perform arithmetic operations such as addition, subtraction, and magnitude comparison [Fossum, 1987]. Such an array could allow for detection of parallel differential outputs, with both positive (excitatory) and negative (inhibitory) weighted sums as dual optical inputs in a (positive definite) intensity representation.

## ARCHITECTURAL CONSIDERATIONS FOR PHOTONIC NEURAL NETWORK IMPLEMENTATIONS

We now turn our attention to the *use* of the photonic components and fundamental principles described above in the implementation of highly parallel neural network *architectures.* The focus in this section is on a general framework that emphasizes characteristics common to different approaches to photonic and optical neural network implementations, as well as on illuminating some of the key fundamental differences among the various implementation approaches. A review of recent and ongoing research in photonic and optical neural network implementations is beyond the scope of this chapter; sources of such information can be found in the Suggested Further Reading section at the end of this chapter.

Photonic neural network implementations can be adaptive or non-adaptive, can represent the signal using different physical quantities, and can be built using one-dimensional

(1-D) or two-dimensional (2-D) arrays of neuron units with two-dimensional or three-dimensional (3-D) interconnection elements. These issues, in addition to other features that are desirable in any photonic implementation of a neural network, are discussed in this section. Throughout, one should keep in mind the distinctions that exist among systems with fixed interconnections, programmable systems, and truly adaptive systems. We will initially concentrate on the implementation of a single layer of a network, and subsequently show how this generalizes to multiple layers.

The computation process of any one layer of a neural network can be represented by:

$$y_i = f\left[\sum_j w_{ij}x_j\right] \tag{23}$$

in which neuron unit $j$ is situated at the input to the layer of interconnections, neuron unit $i$ is situated at the output of the layer of interconnections, $y_i$ is the output of neuron unit $i, x_j$ is the output of neuron unit $j$, $w_{ij}$ is the weight associated with the interconnection between them, and the function $f$ represents the neuron unit nonlinearity. The term inside the brackets, the activation potential, will be denoted by $\rho_i$. Note that the term in brackets is a matrix-vector product between an interconnection weight matrix and an input vector. The function $f$ then operates independently on each element of the resulting vector; this is called a *point nonlinearity*, and as such lends itself to implementation with a spatial light modulator (SLM).

Most current learning algorithms fall into one of a small number of classes. For example, one such class can be specified by:

$$\Delta w_{ij} = \alpha\delta_i x_j - \beta w_{ij} \tag{24}$$

in which $\Delta w_{ij} = w_{ij}(k+1) - w_{ij}(k)$ is the weight update, $k$ represents the iteration index, $\alpha$ is the learning gain constant, and $\beta$ is a decay constant that is included primarily for hardware convenience; $\beta$ can be set to 0 when so desired. Suitable choices of $\delta_i$ give different learning algorithms, such as Hebbian, Widrow-Hoff, single-layer least minimum squares (LMS), and for the case of multilayer networks, backward error propagation. (For

example, an optical architecture that potentially implements backward error propagation in a multilayer neural network is described by Wagner and Psaltis [Wagner, 1987]). In this chapter we will restrict our attention to this particular class of algorithms for illustrative purposes. Although other classes of learning algorithms can likely also be implemented using photonic hardware, research to date has focused primarily on the class represented by Equation (24). An important aspect of Equation (24) for implementation is the outer product between the training vector $\delta$ and the input vector $\mathbf{x}$ for the weight matrix update.

An example of a photonic neural system is shown in block diagram form in Figure 15.21. This system utilizes a 1-D array of neuron units at the input and output, and a 2-D interconnection mask. Each pixel in the input is expanded optically (using cylindrical lenses) and illuminates the corresponding row of the interconnection mask. The mask stores the analog weights, and provides a pointwise multiplication before the beam is contracted so that one column from the mask is incident onto one corresponding output pixel. The optical system in effect provides a fully parallel analog optical matrix-vector multiplication as represented by the bracketed term in Equation (23), performed over all $i$. Threshold functions and feedback connections are provided by means of either photonics or electronics. The first experimental demonstration of such a system applied to neural network implementations used an array of light emitting diodes (LEDs) as inputs to, and a linear detector array as the output from, the optical interconnection [Psaltis, 1985; Farhat, 1985]. This particular system utilized electronics to provide the threshold functions and feedback connections.

**FIGURE 15.21** Block diagram of a 1-D to 1-D photonic neural network, in which a one-dimensional neuron array is fully interconnected to a one-dimensional detector array by means of a two-dimensional interconnection mask.

It should be noted that many variants of Figure 15.21 are possible; some of them are

more compact than others, though all of them share essentially the same basic characteristics. The interconnection mask can be fixed (*e.g.*, photographic film) or variable (*e.g.*, an SLM). In the latter case the SLM can be electronically or optically addressed. Electronic addressing is appropriate for straightforward interfacing to an electronic machine that supplies the (updated) interconnection weights, whereas for a maximum adaptation rate an optical addressing technique would ultimately be optimal. Currently available SLMs with large numbers of pixels tend to be slow (500 × 500 analog pixels with 1 - 100 $ms$ frame times) [Tanguay, 1985]; much faster technologies are being developed for future use [see, for example, Lentine, 1988; Lentine, 1991; McCormick, 1989b]. Such a system, with 1-D inputs, 1-D outputs, and 2-D interconnections, will likely scale up to 100 - 1000 fully connected neuron units.

A photonic system that can implement larger numbers of neuron units and interconnections is shown in Figure 15.22. All neuron unit planes are now 2-D arrays, and the interconnection medium is a 3-D structure, implemented in a volume holographic material. In effect, there is a separate volume grating connecting each input neuron unit $j$ to each output neuron unit $i$. The diffraction efficiency of each grating is proportional to the weight, $w_{ij}$, of the corresponding interconnection. Note that each such grating is analogous to a beamsplitter, as discussed in the previous section, with the primary difference that the volume gratings are direction (and wavelength) selective. Thus, beams incident on such a "beamsplitter" at other than the correct angle are not affected by the presence of the holographic beamsplitter. Properly recorded, then, the grating $w_{ij}$ is situated in angular orientation and grating period so that it affects only the inputs at the angle corresponding to $x_j$ , and will direct the corresponding output $w_{ij}x_j$ to the correct summation node $\rho_i$. The achievable numbers of neuron units and interconnections are currently subjects of considerable debate, but would likely be $10^4$ - $10^6$ neuron units per plane and on the order of $10^{10}$ independent interconnections with weights, assuming continued research unveils no impassable boundaries.

**FIGURE 15.22** Block diagram of a 2-D to 2-D photonic neural network,

in which a two-dimensional neuron array is fully interconnected to a two-dimensional output array by means of a three-dimensional volume holographic optical interconnection mask. The input plane, output plane, and optional training plane are shown. Many variants of this geometry with similar properties are possible.

For the case of an adaptive network, we use a variable (typically photorefractive) holographic material for recording and implementing the interconnections. To incorporate learning, a training plane comprising a 2-D array of nodes generates the $\delta_i$ terms (Figure 15.22). During a weight update, an exposure is made of the interference pattern between beams emanating from the two left hand planes in the figure. Each of the two left hand planes could be implemented using, for example, a 2-D spatial light modulator illuminated by an expanded beam. This results in a change in the refractive index modulation representing the current weight that is dependent on the product $\delta_i x_j$, so that with appropriate choices of parameters, the increment in diffraction efficiency can be made proportional to $\delta_i x_j$. Ideally, this records changes (updates) in the interconnection weights within the hologram given by Equation (24), above, in the form of gratings situated with appropriate angular orientation and grating period. It should be noted that generating and recording these weight updates is not a simple matter, and care must be taken to insure that the appropriate interference terms are recorded and that not too much crosstalk is inadvertently created. Recording and recall of the correct values is primarily a number representation issue and is discussed below; undesirable crosstalk depends on the recording and reconstruction technique as previously discussed in the subsection on "Photonic Interconnections".

An example of one source of holographically-induced interconnection crosstalk is an inadvertent degeneracy of gratings. Even though each volume grating affects only the beams incident at a particular angle with respect to the grating, it affects *all* of the beams at that particular angle. Because of this, an entire cone of beams (with its axis of symmetry aligned with the grating wave vector) can be affected by a single diffraction grating. This

degeneracy creates an undesired coupling between different interconnections in a fully connected network. For neuron unit sources on an ideal, rectangular grid, this coupling can be eliminated by removing neuron units from certain locations in the array, leaving sparsely distributed neuron units arranged in a degeneracy breaking pattern. This eliminates the undesired coupling, at the expense of a reduction in the number of neuron units from $N^2$ (for an $N \times N$ array) to $N^{1.5}$ [Psaltis, 1989].

The case of a non-adaptive network is likely to be an important one as well. In this case the interconnection hologram does not have to be recorded in accordance with a specific learning algorithm. If the weights are known *a priori*, then any applicable recording technique will suffice. In many cases, however, the weights may not be known. A common scenario may involve the training of a "master" network; once it has been trained, copies of the network could be produced in a production environment. If the network is large, and particularly if it utilizes volume holographic optical interconnections, then probing the values of all of the weights could be impractical. The most efficient production means in this case would be to make direct copies of the volume hologram. Thus, the capability of rapidly copying a multiplexed volume interconnection hologram is important.

The physical representation of the signal directly impacts the operation of a photonic neural network. The physical quantities available for optical representation of a signal level are field amplitude, phase, intensity, polarization, spatial position or frequency, and wavelength. We will consider only the most likely candidates: field *amplitude* (with phase) and *intensity*. For the case of an amplitude (with phase) representation, the signals may in general be complex valued; bipolar signals, of course, represent a subset of these numbers, and thus can be represented. Given that $x$ and $y$ are represented as (electric or magnetic) field amplitudes, the resulting detected activation potential of neuron unit $i$, $\rho_i$, is given by

$$\rho_i^{(\text{coh})} = |\sum_j w_{ij} x_j|^2 \tag{25}$$

for the case of a coherent sum, and by

$$\rho_i^{(\text{incoh})} = \sum_j |w_{ij}x_i|^2 \tag{26}$$

for the case of an incoherent sum (*c.f.* the preceeding section on "Fundamental Principles of Photonic Technology"). In both Equations (25) and (26), the weight $w_{ij}$ is represented physically by the *amplitude* diffraction efficiency. The coherent sum given by Equation (25) has the advantage of allowing for the addition of both positive and negative numbers in computation of the neuron unit potential, as desired for the incorporation of both excitatory and inhibitory neuron unit inputs. Clearly, Equations (25) and (26) deviate from conventional neural network models. The effects on different neural network models of such deviations in the summation before thresholding are not currently well understood.

If we instead encode the signal levels as intensities, the activation potential becomes

$$\rho_i^{(\text{int})} = \sum_j w_{ij}x_j, \tag{27}$$

which is the desired activation potential, but at the expense of all terms in the summation being nonnegative. In this case the weight $w_{ij}$ is represented physically by the *intensity* diffraction efficiency. A technique for effectively achieving bipolar signals in this case will be discussed in the section describing "An Implementation Strategy".

The signal representation used also impacts the nature of the weight updates. The physical weight updates can be derived using common models of photorefractive (or other) recording materials. Such a derivation requires a number of approximations and assumptions to be made regarding the chosen operational mode. By appropriate choice of the operational mode, the ideal weight update rule given by Equation (24) can be approximately obtained for both intensity representation and amplitude representation cases. The operational mode may not prove to be the same in each case, and may differ in such parameters as the size of the weight updates, the size of the existing weights before the update, and the exact characteristics of the holographic material used. The "second order" terms that deviate from the precise form of Equation (24) are also different in the two cases; the effect of such terms on learning algorithm performance is not well characterized or

understood, and is currently an active area of research.

So far we have discussed only a single interconnection layer with neuron units for inputs and outputs. If such a physical network includes feedback, it can be generalized to functionally implement an arbitrary multilayer feedforward or recurrent network. Figure 15.23 illustrates this principle, showing one *physical* layer of neuron units, one layer of interconnections from the neuron units to a set of fan-in nodes, and feedback from each fan-in node to the corresponding neuron unit. These neuron units can be conceptually divided into groups corresponding to different *functional* layers. Some of the physical interconnections then represent functionally feedforward connections (represented by solid lines and boxes in Figure 15.23), and some represent functionally lateral connections within a layer (represented by broken lines and boxes in Figure 15.23). Feedback connections to previous layers, and feedforward connections that bypass the next subsequent layer, can also be incorporated in a similar manner, but are not shown in the figure. This technique for implementing multilayer networks using a single physical layer has been discussed by Farhat for the case of 1-D neuron unit arrays interconnected by a 2-D mask, and used in the implementation of parallel optoelectronic simulated annealing [Farhat, 1987]. Thus any photonic (single physical layer) architectures discussed herein generalize to multilayer networks, provided that they have capability for arbitrary connections and feedback.

**FIGURE 15.23** A single layer physical neural network with feedback, used to implement a multilayer recurrent functional network. The solid boxes indicate feedforward connections, and the broken boxes indicate lateral connections.

In summary, the desirable characteristics of a photonic implementation of neural networks include: (1) modularity, so that multiple "modules" can be cascaded; (2) capability for lateral, feedforward, and feedback interconnections, which can be achieved physically by use of a single layer network with feedback and arbitrary interconnection capability;

(3) analog, weighted connections with analog signals; (4) bipolar signals and weights; (5) scalability to large numbers of neuron units with high connectivity; (6) generality, so that different neuron models, network models, and learning algorithms can be implemented within the same basic technology; (7) compatibility of different components within a given architecture; and (8) overall feasibility of the proposed combination of algorithm, architecture, devices, and materials. In addition, the optical/photonic hardware would ideally incorporate the following features: (1) simultaneous, parallel updates of all interconnection weights at each iteration; (2) high optical throughput; (3) low interconnection crosstalk; and (4) flexible functionality for neuron unit response, so that different neuron models and learning algorithms can be accommodated.

## AN IMPLEMENTATION STRATEGY

In this section a photonic technique for the implementation of neural networks is described that potentially satisfies the aforementioned desirable characteristics and features [Jenkins, 1990a; Asthana, 1990a; Jenkins, 1990b; Asthana, 1990b; Jenkins, 1990c]. This photonic neural network implementation technique utilizes optoelectronic spatial light modulators (SLMs) for the 2-D neuron unit and training term planes. Each neuron unit incorporates dual channel encoding to allow for the representation of bipolar input and output signals, and comprises two integrated detectors, two modulators and integrated electronics. The neuron unit input and output signals are represented in the optical system by intensity. The interconnections are based on a 3-D holographic material with a novel incoherent/coherent recording and reconstruction technique that permits simultaneous updates of all weights during each iteration. In addition, the interconnections utilize a unique double angular multiplexing arrangement to minimize interchannel crosstalk and throughput losses, in which each pixel of the object beam SLM is illuminated by a set of mutually incoherent beams, each at a different angle. This implementation technique is explained in the remainder of this section.

A key feature of this implementation strategy is the use of an array of individually

coherent sources that are mutually incoherent to generate an array of coherent beam pairs used for holographic recording and reconstruction in the interconnection network. Consider the problem of recording two holograms, object $A$ recorded with reference beam $x_j$ and object $B$ recorded with reference beam $x_{j'}$, as shown in Figure 15.24(a). The objects $A$ and $B$ could each be a 2-D array of data. In order to write both holograms simultaneously, $A$ and $x_j$ originate from the same coherent source and are mutually coherent; similarly for $B$ and $x_{j'}$. However, $B$ and $x_{j'}$ originate from a different source than $A$ and $x_j$, so that each pair is incoherent with respect to the other pair. In this way, there are no extra (crosstalk) holograms written, such as that between $A$ and $x_{j'}$, or between $x_j$ and $x_{j'}$. This technique can be used for more than two multiplexed holograms, in which case a separate source is assumed for each hologram written.

**FIGURE 15.24** Incoherent/coherent technique for recording and reconstructing multiple holograms simultaneously, in which all solid lines represent mutually coherent beams, and all broken lines represent a separate set of mutually coherent beams: (a) recording; (b) reconstruction; and (c) holographic representation, in which each hologram represents the fanout from a given neuron unit.

During reconstruction, the holograms are illuminated by the same set of reference beams $x_j$ and $x_{j'}$. This simultaneously reconstructs the arrays $A$ and $B$ (Figure 15.24(b)). If the arrays are in registry upon reconstruction, a pixel-by-pixel incoherent sum will be achieved in the output array. If we now consider each reference beam $x_j$ to be the output of a neuron unit at the input to an interconnection layer, then each reconstructed hologram corresponds to the fan-out from one neuron unit, with a contribution to each pixel in the output array proportional to the weight of the corresponding interconnection. This is depicted in Figure 15.24(b) and (c), in which the two signals fanning in to a given neuron unit are derived from separate, mutually incoherent optical sources. Note that this

technique provides an incoherent sum for the potential of each neuron unit (Equation (26) or Equation (27), depending on the chosen representation), as desired.

Another critical as well as unique feature of the photonic architecture described herein is a "double angular multiplexing" technique in which one input node or pixel in the object beam path has multiple beams passing through it at different angles. Thus, a set of angularly multiplexed beams is introduced for each object beam node $\delta_i$, as shown in Figure 15.25. A three-fold angularly multiplexed fan-in from $x_1, x_2$, and $x_3$ to yield neuron unit potential $\rho_1$ is depicted in this figure; solid lines represent mutually coherent beams (all dashed lines represent a mutually coherent set as well; similarly for mixed dashed lines). Note that this multiplexing technique eliminates the fan-in beam degeneracy characteristic of collinear geometries referred to above in the subsection "Photonic Interconnections". Thus, the ensuing cross-coupling terms are absent, and a much more accurate set of weights can be recorded and reconstructed at each iteration.

**FIGURE 15.25** Doubly angularly multiplexed volume holographic optical interconnection, designed to circumvent the effects of beam degeneracy. The mutually incoherent input beams ($\{x_j\}$) are angularly multiplexed over $j$, as are the corresponding sets of output beams from the training plane ($\{\delta_i^{(j)}\}$) generated by the coherent sources $S_j$, to produce an angularly multiplexed fan-in at each summed output, thus yielding the neuron activation potentials $\{\rho_i\}$.

A photonic architecture for neural network implementation that utilizes these principles is shown in Figure 15.26, for the case of Hebbian learning ($\delta_i = y_i$). The components shown in the figure comprise one module; inputs and outputs refer to this particular module. Only feedforward connections are shown. The upper spatial light modulator, $SLM_1$, generates the training terms $\delta_i$ that also represent neuron unit outputs in this case. The lower spatial light modulator, $SLM_2$, is the array of input neuron units. An array of coherent but mutually incoherent sources is used to illuminate the system; they are provided

by a mutually incoherent laser diode array or by a coherent beam passing through an SLM that temporally modulates the phase of each pixel independently. (It can be shown that the latter method is equivalent to the former for the particular type of holographic recording and reconstruction used herein.) A volume holographic material stores the requisite weighted interconnections, and can implement either fixed or adaptive interconnections depending on the material used.

**FIGURE 15.26** Photonic architecture for neural network implementation that incorporates a parallel source array, double angular multiplexing, and incoherent/coherent recording and reconstruction; the Hebbian case is depicted.

Both spatial light modulators in Figure 15.26 consist of an array of pixels, each of which comprises three elements: (1) two integrated detectors for input of positive and negative parts of the neuron unit activation potential, (2) integrated electronic circuitry to provide the neuron unit (sigmoid or hard threshold) nonlinearity, and (3) two hybrid or monolithically integrated modulators for separate optical readout of the positive and negative neuron unit outputs. The SLMs, as shown, are read out in transmission, and have detectors situated so as to receive optical inputs on the right face of the SLM.

In the learning phase, the shutter is open as shown schematically in Figure 15.27. Light from each source $S_j$ is approximately collimated so that it illuminates the entire array on $SLM_1$, at an angle dependent on the position of the $j^{th}$ source. Thus, for an $N$ by $N$ array of sources, there are $N^2$ beams reading out the contents of $SLM_1$ simultaneously, each at a different angle; the entire array of terms $\{y_i\}$ is encoded onto each of these beams. Each such beam then interferes only with its corresponding reference beam $x_j$, derived from the *same* source and encoded by $SLM_2$, in the holographic medium. This writes the set of desired weight update terms $\alpha x_j y_i$.

**FIGURE 15.27** Photonic architecture for neural network implementation: recording configuration. This configuration implements the learning function in the photonic architecture of Figure 15.26. The sets of beams emitted from the source array (two are shown) interfere in the volume holographic medium to update the weights stored in the interconnection holograms.

During the computation phase, the shutter is closed to prevent learning as shown schematically in Figure 15.28. The array of sources is imaged onto $SLM_2$ as a set of readout beams, so that each individual source corresponds to one pixel (neuron unit) on the SLM. The SLM modulates each beam so that the transmitted beam has an intensity proportional to the output value of the corresponding neuron unit. Thus, the $j^{th}$ source illuminates the $j^{th}$ pixel of this SLM, providing the signal $x_j$ that becomes a reference beam to read out the $j^{th}$ hologram. This hologram reconstructs an array of spots, similar to that depicted in Figure 15.24, that contribute to the input of each neuron unit in the output plane. The optics is set up so that this array is imaged onto the detector array. In the complete neural network architecture of Figure 15.26, additional optical elements (mirror $M_2$, lens $L_4$ and beamsplitter $BS_2$) are used to displace the detector array plane to the detector side of $SLM_1$, providing the neuron unit activation potentials. In addition, this beam is sent through beamsplitter $BS_2$ to a subsequent layer in the next module or to the output layer.

**FIGURE 15.28** Photonic architecture for neural network implementation: reconstruction configuration. This configuration implements a single forward pass of the computing function in the photonic architecture of Figure 15.26. The lower set of beams acts as a set of reference beams, and generates a set of weighted output arrays that are imaged onto the detector array. Each stored hologram is reconstructed by a single neuron unit $x_j$, and fans out with appropriate weights to illuminate the detector array. The full set of reconstructed

holograms sums within each pixel to yield the neuron activation potentials $\{\rho_i\}$.

A generalized architecture that incorporates learning algorithms of the form of Equation (24) is shown in Figure 15.29. Instead of $SLM_1$, as in Figure 15.26, this architecture utilizes a training term ($\delta_i$) generator that is implemented via one or more optoelectronic SLMs. In general, target values $t_i$, actual neuron unit outputs $y_i$, or possibly activation potentials $\rho_i$ may be provided as inputs to the training term generator. The physical arrangement of optical beams passing through the training term generator (from left to right) is the same as that shown passing through $SLM_1$ in Figure 15.26. Lateral and feedback connections can be incorporated by including an optical feedback path from the output of the hologram to the input side of $SLM_2$.

**FIGURE 15.29** Generalized photonic architecture for neural network implementation, including provision for the generation of arbitrary training terms ($\delta_i$).

For many applications, both $SLM_1$ and $SLM_2$ can be fabricated using the same technology. Let's consider the case of a sigmoidal response with bipolar inputs and bipolar outputs. The electronics within each neuron unit can take the difference between the two detector inputs to yield the (bipolar) neuron potential. It can then perform the sigmoidal nonlinearity, and send the result to appropriate (positive channel or negative channel) modulator(s). For example, we have fabricated a number of silicon chips that integrate the necessary control electronics with appropriate detectors. One possible circuit that has been designed to incorporate the necessary functionality is shown schematically in Figure 15.30. Outputs from the two photodetection stages ($V_{in1}$ and $V_{in2}$) are differentially amplified using two pairs of CMOS transistors ($M_1 - M_3, M_2 - M_4$), generating two separate and complementary outputs. The differential amplifier has been designed to saturate for

large values of the input signal difference, producing the upper asymptotic limit behavior characteristic of the sigmoid function. Each output signal is then inverted and clipped by another CMOS transistor pair ($M_{21} - M_{22}, M_{11} - M_{12}$), which asymmetrizes the transfer curve and adds the lower asymptotic limit of the sigmoid function. Finally, each output signal is inverted yet again and shifted in level by a dual transistor sub-unity gain amplifier stage ($M_{23} - M_{24}, M_{13} - M_{14}$), producing complementary output signals ($V_{out1}$ and $V_{out2}$) that control the dual channel modulation elements. External provision is made in each pixel (neuron unit) for the adjustment of the voltage offset (zero crossing point) of each characteristic curve. This external bias adjustment allows for post-fabrication fine tuning of the overall response of the circuit, given process-induced variations in device characteristics.

**FIGURE 15.30** Schematic diagram of a dual-input, dual-output differential amplifier that effects a sigmoid-like transfer characteristic.

A sample set of characteristic curves measured from one of these chips is shown in Figure 15.31, with the voltage on the second detector input channel as the parameter. These curves show the differential function of the dual channel circuit, as well as the desired sigmoidal response characteristic. A $6 \times 6$ array of these $100 \times 100$ $\mu m$ neuron units has also been fabricated with excellent uniformity.

**FIGURE 15.31** Experimentally obtained transfer characteristics from the circuit shown in Figure 15.30, showing the output voltage in both channels ($V_{out1}$ and $V_{out2}$) as a function of one input voltage ($V_{in1}$), with the other input voltage ($V_{in2}$) as a parameter.

The modular nature of this photonic neural net architecture can be inferred from Figure 15.29; the upper right SLM is $SLM_2$ of the subsequent module. Feedback paths from one module back to previous modules can be added, if desired, in a relatively straightforward manner. Bipolar signals are incorporated by the dual channel nature of the SLMs, with positive and negative channels for each neuron unit. Since each neuron unit has two physical outputs and two physical inputs, each interconnection between two neuron units physically consists of four separate weighted connections (positive modulator to positive detector, positive modulator to negative detector, *etc.*). Thus, even though each physical weight is nonnegative in value, their combination permits effective implementation of bipolar functional weights. In fact, the extra degrees of freedom provided by four independent weights can require revised weight update rules to ensure convergence of the learning process [Petrisor, 1990].

With the current spatial light modulator design at 100 $\mu m \times 100$ $\mu m$ per neuron unit, $10^4$ neuron units per $cm^2$ can be implemented on each SLM. By constructing an SLM as a mosaic of such arrays, a 3 $in \times 3$ $in$ SLM could implement approximately $5 \times 10^5$ neuron units. Each "tile" or small array within such a mosaic need not be carefully aligned with respect to the other tiles, as the optical system just images the array back onto itself; in the case of multiple modules or lateral/feedback connections, there is, however, a requirement that all SLMs are similarly tiled, within an appropriate tolerance figure. Note that the current design utilizes only 2 $\mu m$ feature sizes in CMOS; this could eventually be scaled down by a factor of 4 in each dimension, yielding more than an order of magnitude increase in the number of neuron units implementable per unit chip area (or an equivalent reduction in the overall size with the same number of neuron units).

It should now be clear that this architecture can be generalized to implement certain other neural models. The use of electronic circuitry for the neuron unit function and training term generation provides significant inherent flexibility. For example, we have completed preliminary designs of units for forward and backward propagating signals in a backpropagation-style multilayer neural network. Although the optical weight updates in the holographic medium are restricted to outer-product terms (Equation (24) in the architecture as shown, variants of the architecture may permit other learning scenarios.

Finally, we consider the important question of making duplicates of a network that has already been trained. Since a volume hologram may store on the order of $10^{10}$ independent weighted interconnections, the preferred technique is to make direct copies of the multiplexed volume hologram. Here we describe a technique for copying such a multiplexed volume hologram in one step [Jenkins, 1990c]. To our knowledge this has never previously been achieved, but the use of incoherent/coherent holographic recording and reconstruction makes this in principle quite straightforward. Figure 15.32 shows an optical setup for duplicating the hologram. The master hologram is illuminated with the same set of reference beams as those employed during exposure; all of the mutually incoherent reference beams illuminate the master volume hologram simultaneously, recalling all of the stored holograms in parallel. The source array is imaged so that it generates an identical set of reference beams on the secondary (copy) holographic medium. Similarly, the reconstructed object beams are also imaged, so that they are incident on the secondary holographic medium, with amplitude and phase identical to that during recording of the master hologram. The appropriate pairs of beams interfere in the secondary holographic medium, making a complete copy of the original hologram. (As shown in Figure 15.32, the copy will actually be a spatially inverted version of the original. A slight variant of the optical system depicted in the figure can produce a copy that is identical to the original.) Thus it is conceivable to mass produce copies of a previously trained interconnection pattern, without ever knowing exactly what the interconnection weights are.

**FIGURE 15.32** Optical layout for copying the entire contents of a three-dimensional volume holographic optical element (VHOE) into a second VHOE, utilizing a two-dimensional array of individually coherent, but mutually incoherent sources.

We conclude this section with a brief summary of the current implementation status of this particular photonic approach to neural network fabrication. For the neuron unit

arrays, $6 \times 6$ arrays of dual-channel detectors integrated with neuron function electronics have been fabricated in silicon and operate correctly. Individual multiple quantum well ($InGaAs/GaAs$) modulators have been successfully fabricated and tested, and exhibit drive voltages compatible with the electronics. The novel doubly angularly multiplexed incoherent/coherent interconnection technique has been tested experimentally at the level of two inputs/two outputs, and simulated at the level of four inputs/four outputs all with very favorable results [Jenkins, 1990c; Asthana, 1990b; Asthana, 1990c]. In addition, several learning algorithms that incorporate some of the unique features of the optical hardware have been successfully designed and simulated. Large 2-D arrays of laser diodes that are not mutually coherent have been fabricated recently [Jewell, 1990; Orenstein, 1990a; Von Lehmen, 1990]. Photorefractive crystals are routinely grown commercially, and can be purchased from vendors for use at visible as well as infrared wavelengths. In addition, the basic requisite features of the doubly angularly multiplexed incoherent/coherent holographic recording techniques have been demonstrated in single crystals of bismuth silicon oxide ($Bi_{12}SiO_{20}$), though not as yet at infrared wavelengths compatible with both the laser diode source array and the multiple quantum well spatial light modulators. All of the other components in the architecture (lenses, beamsplitters, *etc.*) are essentially available off the shelf.

As with any research project in progress, several questions pertaining to the photonic approach outlined herein remain partially unanswered. Consider, for example, the incoherent/coherent source array. Given the current state of the art of laser diode arrays, the total power dissipation will limit the number, maximum optical power, and spacing of the individual sources. Cross-coherence among the sources can cause undesirable crosstalk among corresponding interconnections, although in some neural network models a small to moderate degree of interconnection crosstalk is not likely to cause intolerable degradation in performance. Fortunately, a larger spacing of sources implies that each laser can output a higher power, and also assures a higher degree of mutual incoherence. Other remaining questions include the achievable contrast ratio and uniformity of the spatial light modulators; suitable monolithic or hybrid techniques for integrating detectors, electronics and modulators; optimization of the learning algorithm relative to the chosen holographic ma-

terial's storage and erasure time constants; and linearity and limitations of the hologram copying process. The next section discusses fundamental and technological limitations of the photonic hardware and their impact on the performance of photonic neural network architectures.

# FUNDAMENTAL PHYSICAL AND TECHNOLOGICAL LIMITATIONS OF NEURO-OPTICAL COMPUTATION

Even though we are relatively early on in the development of viable neuro-optical computing systems, it is not too early to begin asking questions about the ultimate boundaries that may impact our future achievements. This line of inquiry can have a two- fold impact. First, discovery of inherently *fundamental physical limitations* that affect all forms of computation can, if correctly applied to the neural computational paradigm, both provide us with an ultimate goal worthy of achievement, and perhaps warn us in advance of architectural choices that will prove unworthy of technological implementation. Second, careful analysis of the *technological limitations* (device performance boundaries within a given technological implementation) that affect system performance can provide us with necessary guidance in choosing among many possible implementation strategies. The goal, of course, is to come up with the right combination of implementation strategy and technological choices to achieve the highest computational throughput (or perhaps learning rate) based on any one of a number of metrics. In this section, then, we discuss both the fundamental physical and technological limitations that impact the future performance of neuro-optical computational systems.

## The Energy Metric

Your brain is truly a remarkable instrument from a computational point of view (as well as from many other points of view!). Although estimates (as well as individuals!) vary, it

is thought that your brain consists of about $10^{11}$ neurons, each interconnected (in certain regions of the brain) to $10^3$ - $10^4$ other neurons [Changeux, 1985; Dowling, 1987; Hubel, 1979]. The human brain exhibits both short and long term memory, performs sophisticated image analysis in fractions of a second, operates as an effective associative memory integrated over a whole lifetime of learning, and yet operates on a power budget that is only a fraction of the power dissipated by the average light bulb in your home [Iversen, 1979]. In order to accomplish this, the active switching elements, the neurons, operate at an average power level about seven orders of magnitude lower than that characteristic of VLSI logic circuits [Mead, 1989b]. If this were not possible, it's likely that you'd be running a temperature even *without* the flu!

This discussion points to one of many possible metrics by which computational systems can be judged: energy (or power) dissipation. In fact, many modern supercomputers are limited in performance *precisely* because of power dissipation boundaries, or the ability to extract the heat generated by the computational process from the volume used to perform the work. We can perhaps think of computation as broken down into three fundamental parts: *representation of information, implementation of computational complexity*, and *detection of the results.* From the energy metric point of view, *everything* costs energy: what goes in costs energy, what comes out costs energy, and what goes on in between costs energy too. The trick in building the computational engines of the future (neural or otherwise) will be to maximize the overall performance with a minimum expenditure of energy.

## Some Quantum Limitations

By *representation of information*, we mean the choice of data representation on which computations are performed. Some examples might include the binary representation, $M$-ary representations, an analog representation, or the residue representation [Huang, 1979]. This choice has implications at the fundamental level for the energy cost to represent a number within a given probability of error. For example, if we detect an optical signal bit that is binary encoded with a so-called "ideal" detector that can tell the difference

between receiving exactly zero photons and one or more photons, it only takes ten photons on the average to guarantee that the signal is received with a probability of error of one part in a billion, or a "bit error rate (BER)" of $10^{-9}$. The average photon at a typical optical communications wavelength of 1300 *nm* has an energy of only $1.5 \times 10^{-19}$ joules, so the total energy cost per bit is 1.5 attojoules ($1.5 \times 10^{-18}$ joules). For a communications channel operating at 10 gigabytes ($8 \times 10^9$ bits) per second, this implies a power dissipation due to representation cost alone (without worrying yet about the *transmission* or *detection* of the information) of only 0.12 microwatts. For currently available detectors, about a thousand photons are required to achieve the same BER, so the necessary representation cost increases to 12 microwatts. In most currently envisioned communications systems, this cost is overwhelmed by other factors.

But what if we chose to represent numbers in an *analog* representation instead? If we were to follow the same kinds of quantum statistical rules, we would find that to represent the number "1000", say, with an effective bit error rate of $10^{-9}$ requires about 150 *million* photons [Tanguay, 1988]. This is about 15 million times larger than the representation cost of a single binary bit, and about 1.5 million times larger than the binary representation cost of the number 1000.

If we assume that the analog representation need only cover numbers between 0 and 1000, then the dependence of the probability of error on the number of photons used to represent the highest number (1000) is given in Figure 15.33. Interestingly, even if we are willing to give up on a couple of orders of magnitude of error probability, our energy cost isn't reduced very much. In fact, it costs about 27 million photons to represent 1000 with 1% error, and about 11 million photons to represent it with as much as 10% error. These numbers can all be reduced by about two orders of magnitude if we are willing to give up a factor of ten in dynamic range, limiting the highest representable number to 100 instead of 1000, as shown in the Figure.

**FIGURE 15.33** The single pixel probability of error $P(Err)$ as a function of the number of photons detected within each pixel, for the cases of 100 and

1000 analog grey levels.

These errors arise fundamentally from the quantum statistical nature of light, and from the fact that we just can't guarantee the number of photons in a packet of light (without resorting to exotic things like "squeezed states", which have their own practical limitations as well as other costs). In the brain, of course, it is currently thought that many (but not *all*) of the quantities involved in signal transmission, both electrical and chemical, are analog in nature.

## The Incorporation of Computational Complexity

Given the fact that it is considerably more expensive to represent quantities in analog as opposed to binary form, why don't we always choose to compute in the binary representation? The answer is that many operations are less energy consumptive to perform in the binary representation, but others are not. The difference lies in the degree of computational complexity that can be implemented on a given representation for a particular computational operation within a chosen technological implementation. For our purposes here, we may define the computational complexity of a given operation as the minimum number of irreducible binary bit operations (over all possible computational algorithms and machine architectures) required to complete the calculation assuming that the data is represented in binary throughout.

For operations of low computational complexity such as transferring data from the CPU to memory or logic and control operations, computation in the binary representation tends to have a significant energy consumption advantage at the fundamental limits (as well as at the current technological limits for both electronic and photonic processors). On the other hand, for operations of high computational complexity such as the two-dimensional Fourier transform that require a very large number of irreducible binary operations to perform (for the optimum algorithm), computation in the analog representation tends to

exhibit lower overall energy consumption, particularly in photonic implementations.

## The Hybrid Representation Concept

In the case of neural networks, a number of characteristic types of computational operations are typically employed, including for example the calculation of weight updates and storage of updated weights, the fan-out of neuron unit outputs, the multiplication of fanned-out outputs by weights, the communication of weighted signals, the fan-in and summation (or differencing) of weighted inputs, and the thresholding of summed inputs to form neuron unit outputs. These operations span a wide gamut of computational as well as physical complexity. As such, we suggest that optimum neural network performance from an energy metric viewpoint may turn out to be best achieved with a hybrid representation, in which the signal representation is essentially binary for certain functions, and essentially analog for others.

An example of the use of this hybrid representation concept is the use of a hard threshold function within each neuron unit to create a two state output (*on* and *off*), and the use of analog holographic storage for all interconnection weights, as described in a previous section. In this case, the *inputs* to each neuron unit are analog, while the *outputs* from each neuron unit are binary. Multiplications are performed in a fully hybridized representation (multiplicands are analog while multipliers are binary), but summations are fully analog (the superposition of fanned-in input intensities). Given a particular choice of implementation technology, then, the overall power budget for a given hybrid representation can be established and compared with similar power budgets for fully analog and fully binary representations.

## The Inherent Costs of Interconnections

Regardless of the representation chosen, it is clear that any computational energy bud-

get must take into account the non-negligible cost of the interconnections themselves. Interconnections characteristic of neural networks are merely a form of weighted communication channels, characterized by a high degree of fan-out and fan-in. For both electronic and photonic implementations that have adaptive weights, it takes energy to calculate the weight updates, it takes energy to store the resultant updated weights, it takes energy to perform the multiplications implied by the weighting of the output signals, and it takes energy to communicate the various signals between layers.

In the electronics case, these energy costs derive from charging up the capacitances of switching devices in the various forms of memory, flipping switches in the various arithmetic operations (both addition and multiplication) for binary representations, operating linear and nonlinear devices for analog representations, and charging and discharging the capacitance associated with output line drivers as well as the capacitance associated with the physical interconnections (wires) among the various parts of the circuit. In the photonics case, the comparable energy costs derive from the generation of light by coherent optical sources, the holographic recording of weights and weight updates, the throughput losses engendered by readout of the holographically stored (and multiplexed) interconnection matrix, any throughput losses associated with the fan-out and fan-in processes, and the inter-and intralayer communication costs.

The bottom line is that in most cases, complex, highly multiplexed interconnection networks with a high degree of fan-in and fan-out are very energy consumptive, and for the neural network case may prove to be the largest energy sink.

## The Inherent Cost of Detection (Switching)

No useful computational system can avoid the costs of detection, both of intermediate results that are essential to following calculations, and of the sought-after answers or output states that initiate subsequent actions or analysis. And this is the cost above all costs that we are certainly willing to pay, as answers or outputs validate the usefulness of the system and its design. As pointed out earlier, there are three primary areas in which detections

are essential: in the generation of summed inputs prior to functional transformation into individual neuron unit outputs (usually on the input side of an optically addressed spatial light modulator), in the generation of specific system outputs, and in the holographic recording of interconnection weights.

The physical process of detection *inherently* involves the dissipation of finite energy (if accomplished in finite time within prescribed uncertainties, as is appropriate for computation), since it necessarily involves the irreversible switching of the state of a physical component, as well as the guarantee that the switched state will be maintained over the time period of measurement without fluctuations due, for example, to thermally induced statistical variations. This is particularly true in highly distributed computing systems such as neural networks that depend on a certain degree of predictability and synchronization of communicated results for progressive computation. As such this is not really a fundamental physical limitation, but rather is a technological limitation imposed by the system designer, who would really like to see some intelligible output from the system in the near future.

## Optimization of the Computational Architecture

The design of a computational architecture in many ways fixes the fundamental performance limitations of the system, as choices must be made about the representation of data within the architecture, the methods employed for the implementation of computational complexity, and the frequency and nature of the detections required for both intermediate and final results. For neural networks capable of sophisticated operations, optimization of the computational architecture against one or more metrics (such as total energy cost for a given computation, or total power at a given operating frequency) will necessitate an appropriate balance among the various representations employed, as well as among the physical mechanisms employed to accomplish the necessary computations. A further balance must be struck between the fraction of the computational burden that is assigned to interconnections, and the fraction that is accomplished by switching (whether logic,

arithmetic, or detection of results).

For many classes of computational problems, the neural network paradigm may prove to be nearly optimal even in the regime in which all of the individual components are assumed to be operating at their respective fundamental physical performance boundaries. Relative to a modern digital supercomputer, certainly, neural networks seemingly offer an unusual mix of hybrid representations (primarily analog), interconnections (highly multiplexed as well as weighted), and switching (infrequent relative to the rate at which interconnections are utilized). It is at the very least intriguing to imagine whether or not our biological heritage has stored within it a useful clue about highly efficient computation for truly sophisticated problems.

## Technological Limitations

The choice of a technological base (or bases) within which to design a neural network with a large number of neuron units and a high degree of connectivity implies yet another set of performance constraints above and beyond the fundamental physical boundaries refered to above. These *technological limitations* may not yet have been reached within the development of a given technology, but can at least be estimated given what we know about the physics of operation of the devices in question. One such technological limitation, for example, governs the total energy dissipation density that can be tolerated on a given semiconductor substrate without either an unacceptable temperature rise that affects device performance, or resort to extraordinary cooling measures (that may prove to be unfeasible in an optical path). As a second example, the energy required to represent a single bit using current digital logic circuits integrated in silicon is about seven orders of magnitude *above* the thermal fluctuation limit [Tanguay, 1988].

In a previous section of this chapter, we discussed a particular photonic implementation strategy for neural networks that involved specific technological choices for the various types of components required by the architecture. At the present state of development of photonic computational systems, we do not have the luxury of designing everything within

a single technological base, as is the case perhaps for computational subsystems based on VLSI chips. "Optical silicon" has not yet emerged, or at the very least has not yet been identified and recognized as such, even though numerous candidates have been intensively investigated.

Perhaps the leading candidate at the current time is the compound semiconductor system based on gallium arsenide ($GaAs$) and including related ternary compounds such as indium gallium arsenide ($In_xGa_{1-x}As$) and aluminum gallium arsenide ($Al_xGa_{1-x}As$). Within this system, at least, sources, source arrays, spatial light modulators, integrated electronic circuitry, volume holographic optical elements, detectors, and detector arrays have all been fabricated and evaluated with varying degrees of success. What has *not* been established to date is the mutual compatibility of all of these elements operating within a given systems context. This demonstration of mutual compatibility in all relevant performance specifications is essential, because in a highly interconnected system the overall performance achieved is often most strongly influenced by the component with the *least* desirable characteristics. An obvious example is that of a single channel optical communications link, for which the transmission bandwidth will be delimited by the lowest bandwidth component among the source/modulator, transmission medium, and detector/amplifier.

In the remainder of this section, we briefly discuss a number of the types of technological limitations that will impact the performance of currently envisioned photonic implementations of neural networks.

With regard to sources and source arrays, the principal technological issues are the minimization of laser thresholds to allow for parallel operation of a large number of sources on a single chip, the coherence length achievable with ultrashort cavity surface emitting lasers when fabricated in an array (which impacts the holographic recording process), the uniformity of wavelength across the array (particularly for parallel readout of wavelength sensitive devices such as multiple quantum well spatial light modulators), and both the short term (process-determined) yield and the long term reliability of individual sources within a large scale array.

For the case of spatial light modulators, key technological issues include the maximum

density of neuron units that can be integrated on a monolithic or hybrid chip with appropriate detectors, control circuitry, and modulators within each pixel; the sensitivity to input intensity; the neural unit functionality (and perhaps programmability) that can be achieved at the minimum cost in real estate and energy dissipation; the contrast ratio and uniformity of the contrast ratio across the array of pixels (neuron units); the achievable dynamic range of the input/output transfer function; and the operational bandwidth that can be reached assuming a 50% duty cycle for each neuron unit (which determines the total power dissipation of the chip).

The high degree of interconnectivity envisioned for photonic implementations of neural networks hinges primarily on the achievement of appropriate functionality in the volume holographic optical elements used to record and store interconnection weights, produce fan-in and fan-out from each neuron unit, and allow for highly parallel readout of the weighted interconnection network. Key technological limitations for currently investigated photorefractive materials include the optical quality routinely achievable in large (1 cubic inch) single crystal samples, the storage capacity of the medium as determined by the highest spatial frequency gratings that can be recorded, the sensitivity for recording of updated weights [Johnson, 1988] at the source wavelength (which in turn determines the source power necessary to initiate weight updates during the learning phase), the potential for "fixing" of the stored interconnection weights to allow for nondestructive readout during computation, and the capability of copying the contents of the stored interconnection matrix into another holographic medium (in order to provide the capacity for mass production of fixed pattern interconnections following a training sequence executed with a dynamic medium). Perhaps the most important technological limitation of a given holographic recording medium will prove to be the total number of weight update cycles that can be initiated without complete erasure of the weight updates recorded during the very first training cycle. This number in effect sets an upper bound on the learning capacity of the photonic neural network.

The primary technological limitations of importance to single pixel detectors as used, for example, on the input side of optically adressed spatial light modulators, include the sensitivity of the detector/amplifier combination (which together with the spatial light

modulator gain determines the overall loop gain for a single computational iteration), the modulation bandwidth in conjunction with following circuitry, and the chip area required to achieve the desired sensitivity and bandwidth tradeoff. Depending on the technological base within which the spatial light modulator is fabricated, the potential for integration with control circuitry and in some cases the modulation elements themselves provides an additional constraint.

For detector arrays, many of the same issues apply with the additional constraints of uniformity of each performance parameter across the array, and the reliability of the full array of pixels. Another important technological issue is the frame rate for readout of the entire array at a given pixel density, which is determined both by the technological base within which the array is fabricated, and by the physical structure of the array and its readout configuration. As discussed in an earlier section, charge-coupled-device (CCD) arrays are typically read out by temporally multiplexing the contents of a full frame onto one or a few high bandwidth serial outputs. The contents of each row of stored charge packets generated during the exposure cycle are shifted out to a high speed serial readout buffer, which reads out one entire column of pixels in between each lateral shift of the rows as shown schematically in Figure 15.34.

**FIGURE 15.34** Illustration of parallel-to-serial conversion in two-dimensional detector arrays such as the charge-coupled-device (CCD) array. Charge accumulated within each photosensitive region during exposure is transferred laterally by a set of row-parallel shift registers to a high speed parallel-to-serial shift register, which reads out the entire array one column at a time.

This parallel-to-serial conversion limits the frame readout rate to the maximum serial transfer rate achievable in the readout buffer, divided by the number of pixels in the array. For example, if the array is 2048 $\times$ 2048 pixels in size with a readout buffer operating at 200 *MHz*, the frame rate will be limited to about 60 frames per second. For many neural

network applications, this frame rate may be more than sufficient for access to the desired outputs (including the time required for temporal demultiplexing of the output). In cases that demand higher frame rates, the array can be segmented so that multiple readout buffers can be used, each accessing a fraction of the total number of rows in the array.

An unusual feature of identifying the technological constraints that bound *any* neural network implementation, whether it be electronic, photonic, chemical, mechanical, or all of the above, is the fact that we just don't know enough yet about the operation of highly interconnected nonlinear systems of large dimension to fully assess the impact of a *particular* constraint on the overall system operation. One example is the degree to which nonuniformities in the neuron units themselves (*e.g.* in their sensitivity, contrast ratio, or overall response function) or in the interconnection medium can be tolerated by an architecture that is to a large extent self-organizing. A second example is the necessity within a certain neural network paradigm of implementing precisely the right nonlinearity that translates summed neuron inputs to neuron outputs (or, for that matter, the very existence of nonlinearities in the recording and storage of weight updates, and in the readout of the full weighted interconnection pattern). Some of these types of questions may be amenable to simulation, but in other cases we may have to await the results from actual implementations to refine our understanding of the technological requirements.

## THE FUTURE OF NEURO-OPTICAL COMPUTATION

A wide variety of photonic architectures and components are currently under intensive investigation for neural network applications. A thorough discussion of these alternative strategies and their strengths and weaknesses is unfortunately beyond the scope of this chapter. In this final section, we address the future prospects of neuro-optical computation from the point of view of those critical issues that are common to all such proposed implementation strategies.

## The Critical Issues

As with any emerging technological breakthrough, successful advanced development of photonic neural networks will require that the technology prove to be *manufacturable*, in the sense that it is amenable to mass production techniques at reasonable cost; *flexible in design*, in that the technological base provides significant degrees of freedom for architectural and functional variations; and *leveraged* as much as possible by developments in related technologies that can offload a significant fraction of the development time and costs. Implicit in these three key features is the issue of the component uniformity that can be achieved over large array sizes, and the related issue of the scalability of the technology to large scale systems either by increases in the basic array sizes or by the incorporation of a modular design from the outset.

In all useful computational systems, the bottom line is to a large extent determined by the maximum amount of computational throughput capacity that can be squeezed into the smallest system volume within a tolerable energy dissipation constraint or power budget. In the case of photonic neural networks that utilize the implementation strategy outlined in a previous section of this chapter, the two most important factors that influence the computational throughput capacity are the storage capacity of the volume holographic optical element, and the operational bandwidth of the neuron units (nonlinear spatial light modulators).

The storage capacity of a holographic interconnection medium is in turn determined by the maximum number of independent weighted interconnections that can be recorded and retrieved per unit volume. Although we discussed this issue in an earlier section from a theoretical viewpoint (and in yet another section from an architectural viewpoint), we have not addressed herein the even more important question of the *actual* density of independent interconnections achievable with photorefractive (or other photosensitive real time materials) that contain a considerable number of scattering centers and exhibit diffraction efficiencies that depend strongly on the spatial frequency of the recorded grating. Demonstration of a high density of weighted interconnections with low interchannel crosstalk in a

real time holographic recording medium will provide a benchmark of achievement for photonic implementation strategies, as well as a metric by which one can more appropriately estimate eventual system performance. Demonstration of parallel weight updates at high sensitivity (requiring tolerable optical source intensities) during a complete training cycle is also necessary for a convincing proof of system feasibility.

The operational bandwidth of two-dimensional spatial light modulators that can be used as neuron unit arrays will prove to be orders of magnitude larger than the bandwidth characteristic of biological systems. Feasibility analyses, as well as preliminary device characterization studies, indicate no fundamental or technological barriers to operation of individual neuron units at bandwidths exceeding 100 *MHz* [Asthana, 1990b]. For a 100 $\times$ 100 element neuron unit to switch at 100 *MHz* with a 50% duty cycle, however, generates a watt of power dissipation for every 2 picojoules of switching energy required by an individual neuron unit, including the detector, amplifier, control circuitry, and modulator. Although the neuron unit arrays in most neural network architectures will not approach 50% duty cycles from full *off* to full *on* in actual operation, this still gives us a very tight energy dissipation budget, and may eventually force a lowering of the design bandwidth.

For photonic implementations of neural networks that use optical imaging and holographic interconnection systems extensively to increase the computational throughput capacity, an important question will continue to be the fraction of "unfilled" system volume dedicated to wavefront and beam propagation. Miniaturization of many sophisticated optical signal processing systems has been a focus of effort only recently, and can be expected to produce significant system volume reductions through clever (as well as careful) optomechanical engineering. For example, gradient index (GRIN) techniques have been used in conjunction with photolithographic planar processing to produce regular two-dimensional arrays of diffraction limited microlenses that can be incorporated in stacked planar optical modules with greatly reduced unfilled system volume [Iga, 1984]. There are, however, several inherent limits (both fundamental as well as technological) that provide lower bounds on the physical volume required to implement a high degree of interconnectivity among planes of neuron units.

When all is said and done, it is certainly a fair question to ask whether the physical

volume of a neuro-optical computer module would be better off densely packed with silicon chips that emulate the same functionality. All of the preliminary evidence gathered to date suggests that the answer to this question shifts rather dramatically from an emphatic "yes" in the limit of small numbers of neurons and required interconnections, to a more tenuous "no" as the number of neurons and density of required interconnections continues to increase. Perhaps the most interesting question for the future of neuro-optical computation is the clear identification of this performance boundary.

## The Incorporation of Neural Paradigms

In concluding this chapter on neuro-optical computation, we assert that although the majority of preliminary demonstrations of photonic neural network architectures have seemingly focused on associative memories in general, and on variations of the Hopfield-Amari [Hopfield, 1982; Amari, 1972] network in particular, it is essential that photonic implementations have the capacity for incorporation of a wide variety of neural network architectures, computational algorithms, and learning rules. This is particularly important in view of the early stage of development that characterizes our current understanding of the operational performance of even the most fashionable neural network models, when scaled up to large numbers of densely interconnected neuron units with realistic stochastic variations in individual neuron unit performance.

Furthermore, it is likely that useful neural networks incorporated in a systems framework may require either a number of layers with different characteristics, or considerable pre-and post-processing to achieve sophisticated functionality. One example of a neural paradigm that requires such additonal sophistication is the Dynamic Link Architecture of von der Malsburg discussed in Chapter 11 [Buhmann, 1991] as applied to pattern recognition problems by means of graph matching techniques. A photonic implementation of this architecture will likely require several interacting modules for complete functionality.

Currently investigated photonic architectures and components for neural network implementation do not yet enjoy the flexibility of full-fledged computer aided design and

computer aided manufacturing that is the hallmark of the silicon VLSI circuit industry. On the other hand, tremendous strides have been made in just the past few years in the simulation of even complex optical systems including the effects of both refractive and diffractive components. The next step, from extensive simulation capabilities to design automation, is under active investigation and may when taken herald the beginnings of a viable photonic-based neural network implementation technology.

## SUGGESTED FURTHER READING


Abu-Mostafa, Y. S., and Psaltis, D., "Optical Neural Computers", *Scientific American*, vol. 256, no. 3, 88-95, 1987.

*Applied Optics*, Special Issue on Neural Networks, vol. 26, no. 23, 1 December, 1987.

Arsenault, H., Szoplik, T., and Macukow, B., Eds., *Optical Processing and Computing*, Academic Press, New York, 1989.

Collier, R. J., Burckhardt, C. B., and Lin, L. H., *Optical Holography*, Academic Press, New York, 1971.

Feitelson, D. G., *Optical Computing: A Survey for Computer Scientists*, MIT Press, Cambridge, 1988.

Goodman, J. W., *Introduction to Fourier Optics*, McGraw-Hill Book Company, New York, 1968.

Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications I*, Vol-

ume 61 in Topics in Applied Physics Series, Springer-Verlag, New York, 1988; also Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications II*, Volume 62 in Topics in Applied Physics Series, Springer-Verlag, New York, 1989.

Haus, H. A., *Waves and Fields in Optoelectronics*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.

Ishihara, S., Ed., *Optical Computing in Japan*, Nova Science Publishers, Commack, New York, 1990.

Nussbaum, A., and Phillips, R. A., *Contemporary Optics for Scientists and Engineers*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1976.

*Optical Computing*, Volume 9 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D.C., 1989.

Pankove, J. I., *Optical Processes in Semiconductors*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

*Proceedings of the IEEE International Conference on Neural Networks*, vol. III, San Diego, 1987, pp. 549-648; *Proceedings of the IEEE International Conference of Neural Networks*, vol. II, San Diego, 1988, pp. 357-442; *Proceedings of the International Joint Conference on Neural Networks*, vol. II, Washington, D.C., 1989, pp. 457-494; (The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ).

Smith, H. M., *Holographic Recording Materials*, Volume 20 in Topics in Applied Physics Series, Springer-Verlag, New York, 1977.

Sze, S. M., *Physics of Semiconductor Devices, 2nd Ed.*, John Wiley and Sons, New York, 1981.

## REFERENCES

Abu-Mostafa, Y., "Complexity in Neural Systems", Appendix D in Mead, C.A., *Analog VLSI and Neural Systems*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1989, pp. 353-358.

Amari, S.-I., "Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements," *IEEE Transactions on Computers*, vol. C-21, 1197-1206, 1972.

Asthana, P., Chin, H., Nordin, G., Tanguay, A. R., Jr., Piazzolla, S., Jenkins, B. K., and Madhukar, A., "Photonic components for neural net implementations using incoherent/coherent holographic interconnections," *OC'90 Technical Digest, International Commission for Optics*, Kobe, Japan, 1990a.

Asthana, P., Chin, H., Nordin, G., Tanguay, A. R., Jr., Petrisor, G. C., Jenkins, B. K., and Madhukar, A., "Photonic components for neural net implementations using incoherent-coherent holographic interconnections", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990b, p. 57.

Asthana, P., Nordin, G., Piazzolla, S., Tanguay, A. R., Jr., and Jenkins, B. K., "Analysis of interchannel crosstalk and throughput efficiency in highly multipexed fan-out/fan-in holographic interconnections", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990c, p. 242.

Blouke, M. M., Corrie, B., Heidtmann, D. L., Yang, F. H., Winzenread, M., Lust, M. L., Marsh IV, H. H., and Janesick, J. R., "Large format, high resolution image sensors", *Optical Engineering*, vol. 26, no. 9, 837-843, 1987.

Buhmann, J., Lange, J., von der Malsburg, C., Vorbrüggen, J. C., and Würtz, R. P., "Object Recognition in the Dynamic Link Architecture–Parallel Implementation on a Transputer Network," in *Neural Networks: A Dynamical Systems Approach to Machine Intelligence*, B. Kosko, Ed., Prentice-Hall, Englewood Cliffs, NJ, 1991; Chapt. 11, pp. xxx-xxx.

Carpenter, G. A. and Grossberg, S., "ART 2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol. 26, no. 23, 4919-4930, 1987.

Casey, H. C., Jr., and Panish, M. B., *Heterostructure Lasers. Part A: Fundamental Principles*, in Quantum Electronics–Principles and Applications Monograph Series, P. F. Liao and P. Kelley, Eds., Academic Press, Inc., New York, 1978a.

Casey, H. C., Jr., and Panish, M. B., *Heterostructure Lasers. Part B: Materials and Operating Characteristics*, in Quantum Electronics–Principles and Applications Monograph Series, P. F. Liao and P. Kelley, Eds., Academic Press, Inc., New York, 1978b.

Chang-Hasnain, C. J., Maeda, M. W., Stoffel, N. G., Harbison, J. P., and Florez, L. T., "Surface Emitting Laser Arrays with Uniformly Separated Wavelengths", *Electronics Letters*, vol. 26, no. 13, 940-942, 1990.

Changeux, J.-P., *Neuronal Man: The Biology of Mind*, Pantheon Books, New York, 1985.

Dammann, M., and Gortler, K., "High-Efficiency In-Line Multiple Imaging by Means of

Multiple Phase Holograms", *Optics Communications*, vol. 3, no. 5, 312-315, 1971.

Drabik, T. J., and Handschy, M. A., "Silicon VLSI/ferroelectric liquid crystal technology for micropower optoelectronic computing devices", *Applied Optics*, vol. 29, no. 35, 5220-5223, 1990.

Dowling, J. E., *The Retina: An Approachable Part of the Brain*, The Belknap Press of Harvard University Press, Cambridge, Massachusetts, 1987.

Farhat, N. H., "Optoelectronic analogs of self-programming neural nets: architecture and methodologies for implementing fast stochastic learning by simulated annealing," *Applied Optics*, vol. 26, no. 23, 5093-5103, 1987.

Farhat, N. H., Psaltis, D., Prata, A., and Paek, E., "Optical implementation of the Hopfield model," *Applied Optics*, vol. 24, no. 10, 1469-1475, 1985.

Fossum, E. R., "Charge-coupled computing for focal plane image preprocessing", *Optical Engineering*, vol. 26, no. 9, 916-922, 1987.

Goodman, J. W., *Introduction to Fourier Optics*, McGraw-Hill Book Company, New York, 1968, Chapt. 4.

Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications I*, Volume 61 in Topics in Applied Physics Series, Springer-Verlag, New York, 1988.

Gunter, P., and Huignard, J.-P., *Photorefractive Materials and Their Applications II*, Volume 62 in Topics in Applied Physics Series, Springer-Verlag, New York, 1989.

Haus, H. A., *Waves and Fields in Optoelectronics,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984, pp. 63-72.

Herbulock, E. J., Garrett, M. H., and Tanguay, A. R., Jr., "Electric field profile effects on photorefractive grating formation in bismuth silicon oxide", *OSA Annual Meeting Technical Digest 1988*, Vol. 11 of the 1988 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1988, p. 143.

Hopfield, J. J., "Neural Networks and Physical Systems with Emergent Collective Computational Activity," *Proceedings of the National Academy of Sciences, USA*, vol. 79, 2554-2558, 1982.

Huang, A., Tsunoda, Y., Goodman, J. W., and Ishihara, S., "Optical computation using residue arithmetic", *Applied Optics*, vol. 18, no. 2, 149-162, 1979.

Hubel, D. H., "The Brain", *Scientific American*, vol. 241, no.3, 44-53, 1979.

Iga, K., Kokubun, Y., and Oikawa, M., *Fundamentals of Microoptics: Distributed-Index, Microlens, and Stacked Planar Optics*, Academic Press, Inc., Tokyo, 1984.

Iversen, L. L., "The Chemistry of the Brain", *Scientific American*, vol. 241, no. 3, 134-149, 1979.

Jackel, L. D., "Electronic Neural Networks," in *OSA Annual Meeting Technical Digest, 1988*, Vol. 11 of the 1988 OSA Technical Digest Series, Optical Society of America, Washington, D.C., 1988, p. 146.

Jenkins, B. K., Chavel, P., Forchheimer, R., Sawchuk, A. A., and Strand, T. C., "Architectural Implications of a Digital Optical Processor," *Applied Optics*, vol. 23, no. 19, pp. 3465-3474, 1984.

Jenkins, B. K., Petrisor, G. C., Piazzolla, S., Asthana, P., and Tanguay, A. R., Jr., "Photonic architecture for neural nets using incoherent/coherent holographic interconnections," in *OC'90 Technical Digest, International Commission for Optics*, Kobe, Japan, 1990a.

Jenkins, B. K., Tanguay, A. R., Jr., Piazzolla, S., Petrisor, G. C., and Asthana, P., "Photonic neural network architecture based on incoherent-coherent holographic interconnections", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990b, p. 56.

Jenkins, B. K. and Tanguay, A. R., Jr., "Incoherent/coherent multiplexed holographic recording for photonic interconnections and holographic optical elements," United States Patent Application USC-2254, University of Southern California, Los Angeles, California, 1990c.

Jewell, J. L., Lee, Y. H., Scherer, A., McCall, S. L., Olsson, N. A., Harbison, J. P., and Florez, L. T., "Surface-emitting microlasers for photonic switching and interchip connections", *Optical Engineering*, vol. 29, no. 3, 210-214, 1990.

Johnson, R. V., and Tanguay, A. R., Jr., "Optical beam propagation method for birefringent phase grating diffraction," *Optical Engineering*, vol. 25, no. 2, 235-249, 1986.

Johnson, R. V., and Tanguay, A. R., Jr., "Fundamental Physical Limitations of the Photorefractive Grating Recording Sensitivity", Chapter 3 in *Optical Processing and Computing*, H. Arsenault, T. Szoplik, and B. Macukow, Eds., Academic Press, New York, 1989.

Jones, W. B., Jr., *Introduction to Optical Fiber Communication Systems*, Holt, Rinehart and Winston, Inc., New York, 1988.

Kaminow, I. P., *An Introduction to Electrooptic Devices*, Academic Press, New York, 1974.

Karim, Z., Garrett, M. H., and Tanguay, A. R., Jr., "Bandpass AR coating design for bismuth silicon oxide", in *OSA Annual Meeting Technical Digest 1988*, Vol. 11 of the 1988 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1988, p. 125.

Karim, Z., and Tanguay, A. R., Jr., "Bandpass AR coating for the photorefractive materials $LiNbO_3$, $BaTiO_3$, CdTe, and PLZT", in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1989a, p. 78.

Karim, Z., Kyriakakis, C., and Tanguay, A. R., Jr., "Improved two beam coupling gain and diffraction efficiency in bismuth silicon oxide crystals using a bandpass AR coating", in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1989b, p. 29.

Kogelnik, H., "Coupled Wave Theory for Thick Hologram Gratings", *Bell System Technical Journal*, vol. 48, no. 9, 2909-2947, (1969).

Kressel, H., and Butler, J. K., *Semiconductor Lasers and Heterojunction LEDs*, in Quantum Electronics–Principles and Applications Monograph Series, Y.-H. Pao and P. Kelley, Eds., Academic Press, Inc., New York, 1977.

Kyriakakis, C., Karim, Z., Jung, J. J., Tanguay, A. R., Jr., and Madhukar, A., "Funda-

mental and Technological Limitations of Asymmetric Cavity MQW InGaAs/GaAs Spatial Light Modulators", in *Proceedings of the Optical Society of America Topical Conference on Spatial Light Modulators, Incline Village, Nevada*, Optical Society of America, Washington, D.C., 1990.

Lee, B. W. and Sheu, B. J., "Designs and Analysis of VLSI Neural Networks", in *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Bart Kosko, Ed., Prentice-Hall, Englewood Cliffs, New Jersey, 1990; Chapt. 14, pp. xxx-xxx.

Lentine, A. L., Hinton, H. S., Miller, D. A. B., Henry, J. E., Cunningham, J. E., and Chirovsky, L. M. F., "Symmetric self-electro-optic effect device: optical set-reset latch," *Applied Physics Letters*, vol. 52, 1419-1421, 1988.

Lentine, A. L., Chirovsky, L. M. F., and D'Asaro, L. A., "Photonic Ring Counter Using Batch-Fabricated Symmetric Self-Electro-Optic-Effect Devices," *Optics Letters*, vol. 16, no. 1, 36-38, 1991.

Marrakchi, A., Hubbard, W. M., Habiby, S. F., and Patel, J. S., "Dynamic holographic interconnects with analog weights in photorefractive crystals", *Optical Engineering*, vol. 29, no. 3, 215-224, 1990.

McCormick, F. B., "Generation of large spot arrays from a single laser beam by multiple imaging with binary phase gratings", *Optical Engineering*, vol. 28, no. 4, 299-304, 1989.

McCormick, F. B., Lentine, A. L., Morrison, R. L., Walker, S. L., Chirovsky, L. M. F., and D'Asaro, L. A., "Simultaneous parallel operation of an array of symmetric self-electrooptic effect devices," in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA

Technical Digest Series, Optical Society of America, Washington, D.C., 1989b, pp. 60-61.

Mead, C. A., *Analog VLSI and Neural Systems*, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1989.

Mead, C. A., *op. cit.*, 1989b, p. 3.

Mead, C. A. and Mahowald, M. A., "A silicon model of early visual processing," *Neural Networks*, vol. 1, 91-97, 1988.

Miller, D. A. B., "Quantum-well self-electro-optic effect devices," *Optical and Quantum Electronics*, vol. 22, S61-S98, 1990.

Milonni, P. W. and Eberly, J. H., "Specific Lasers and Pumping Mechanisms", Chapter 13 in *Lasers*, John Wiley and Sons, New York, 1988, pp. 411-468.

Morrison, R.L., and Walker, S.L., "Binary phase gratings generating even numbered spot arrays", in *OSA Annual Meeting Technical Digest 1989*, Vol. 18 of the 1989 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1989, p. 111.

*Optical Engineering*, Special Issue on Charge-Coupled-Device Manufacture and Application, vol. 26, no. 9, 827-943, 1987a.

*Optical Engineering*, Special Issue on Charge-Coupled-Device and Charge-Injection-Device Theory and Application, vol. 26, no. 10, 963-1076, 1987b.

Orenstein, M., von Lehmen, A. C., Chang-Hasnain, C., Stoffel, N. G., Harbison, J. P.,

Florez, L. T., Wullert, J. R., and Scherer, A., "Matrix addressable surface emitting laser array", in *Proceedings of the 1990 Conference on Lasers and Electro-Optics*, Vol. 7 of the 1990 Technical Digest Series, Optical Society of America, Washington, D. C., 1990a, p. 88.

Orenstein, M., von Lehmen, A. C., Stoffel, N. G., Chang-Hasnain, C., Harbison, J. P., Florez, L. T., Clausen, E., and Jewell, J. L., "Lateral definition of high performance surface emitting lasers by planarity preserving ion implantation processes", in *Proceedings of the 1990 Conference on Lasers and Electro-Optics*, Vol. 7 of the 1990 Technical Digest Series, Optical Society of America, Washington, D. C., 1990b, p. 504.

Petrisor, G. C., Jenkins, B. K., Chin, H., and Tanguay, A. R., Jr., "Dual function adaptive neural networks for photonic implementation", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990, p. 56.

Psaltis, D., and Farhat, N. H., "Optical information processing based on an associative-memory model of neural nets with thresholding and feedback," *Optics Letters*, vol. 10, no. 2, 98-100, 1985.

Psaltis, D., Brady, D., and Wagner, K., "Adaptive optical networks using photorefractive crystals", *Applied Optics*, vol. 27, no. 9, 1752-1759, 1988.

Psaltis, D., Brady, D., Gu, X.-G., and Hsu, K., "Optical Implementation of Neural Computers," in *Optical Processing and Computing*, H. H. Arsenault, T. Szoplik, and B. Macukow, Eds., Academic Press, New York, 1989, pp. 251-276.

Shirouzu, S., Tsuji, T., Harada, N., Sado, T., Aihara, S., Tsunoda, R., and Kanno, T., "64 × 64 InSb Focal Plane Array with Improved Two Layer Structure", *Proceedings of*

*the SPIE*, vol. 661, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, 1986.

Smith, H. M., *Holographic Recording Materials*, Volume 20 in Topics in Applied Physics Series, Springer-Verlag, New York, 1977.

*Spatial Light Modulators and Applications*, Volume 14 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D.C., 1990.

*Spatial Light Modulators for Optical Information Processing*, Special Issue of *Applied Optics*, vol. 28, no. 22, 1989, pp. 4739-4913.

Streetman, B. G., *Solid State Electronic Devices*, Second Edition, Prentice-Hall, Englewood Cliffs, New Jersey, 1980.

Sze, S. M., *Physics of Semiconductor Devices, 2nd Ed.*, John Wiley and Sons, New York, 1981a, pp. 312-361.

Sze, S. M., *op. cit.*, 1981b, pp. 783-784.

Sze, S. M., *op. cit.*, 1981c, pp. 407-427.

Tai, K., Fischer, R. J., Wang, K. W., Chu, S. N. G., and Cho, A. Y., "Use of Implant Isolation for Fabrication of Vertical Cavity Surface-Emitting Laser Diodes", *Electronics Letters*, vol. 25, no. 24, 1644-1645, 1989a.

Tai, K., Fischer, R. J., Seabury, C. W., Olsson, N. A., Huo, T.-C. D., Ota, Y., and Cho,

A. Y., "Room-temperature continuous-wave vertical-cavity surface-emitting GaAs injection lasers", *Applied Physics Letters*, vol. 55, no. 24, 2473-2475, 1989b.

Tanguay, A. R., Jr., "Physical and Technological Limitations of Optical Information Processing and Computing", *Materials Research Society Bulletin*, Special Issue on Photonic Materials, vol. XIII, no. 8, 36-40, 1988.

Tanguay, A. R., Jr., "Materials requirements for optical processing and computing devices," *Optical Engineering*, vol. 24, no. 1, 2-18, 1985.

von der Malsburg, C., "Goal and Architecture of Neural Computers", in *Neural Computers*, R. Eckmiller and C. von der Malsburg, Eds., Volume 41 of NATO Advanced Science Institutes Series F: Computer and Systems Sciences, Springer-Verlag, New York, 1987.

von Lehmen, A., Orenstein, M., Chang-Hasnain, C., Banwell, T., Wullert, J., Stoffel, N., Florez, L., and Harbison, J., "Rastered operation of row-column addressed vertical-cavity surface-emitting laser array", in *OSA Annual Meeting Technical Digest 1990*, Vol. 15 of the 1990 OSA Technical Digest Series, Optical Society of America, Washington, D. C., 1990, p. 15.

Wagner, K., and Psaltis, D., "Multilayer Optical Learning Networks," *Applied Optics*, vol. 26, no. 23, 5061-5076, 1987.

Warde, C., and Fisher, A. D., "Spatial Light Modulators: Applications and Functional Capabilities", Chapter 7.2 in *Optical Signal Processing*, J. L. Horner, Ed., Academic Press, Inc., New York, 1987, pp. 477-523.

Whitehead, M., and Parry, G., "High-contrast reflection modulation at normal incidence

in asymmetric multiple quantum well Fabry-Perot structure", *Electronics Letters*, vol. 25, 566-568, 1989a.

Whitehead, M., Parry, G., and Wheatley, P., "Investigation of etalon effects in GaAs-AlGaAs multiple quantum well modulators", *IEE Proceedings*, vol. 136, pt. J, no. 1, 52-58, 1989b.

Whitehead, M., Rivers, A., Parry, G., Roberts, J. S., and Button, C., "Low-voltage multiple quantum well reflection modulator with on:off ratio > 100:1" *Electronics Letters*, vol. 25, no. 15, 984-985, 1989c.

Yan, R. H., Simes, R. J., and Coldren, L. A., "Wide-bandwidth, high-efficiency reflection modulators using an unbalanced Fabry-Perot structure," *Applied Physics Letters*, vol. 55, no. 19, 1946-1948, 1989.

## ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions to this effort provided by our faculty colleagues Christoph von der Malsburg, Joachim Buhmann, and Anupam Madhukar, and by our graduate research assistants Greg Nordin, Praveen Asthana, Howard Chin, Sabino Piazzolla, Greg Petrisor, Chris Kyriakakis, Zaheed Karim, John Rilum, and Ed Herbulock. Special thanks are also due to Gloria Bullock and Delsa Tan for their help in the preparation of this manuscript. Funding for the materials, device, and systems aspects of our research program on photonic implementations of neural networks has been provided

by the Defense Advanced Research Projects Agency, the University Research Initiative "Center for the Integration of Optical Computing" (sponsored by the Air Force Office of Scientific Research), the National Center for Integrated Photonic Technology (sponsored by the Defense Advanced Research Projects Agency), the Joint Services Electronics Program, and NTT Corporation.

## PROBLEMS

1. Consider a system of beamsplitters arranged to combine a set of $N$ input beams to form a single, collinear output beam with an intensity proportional to an equally weighted sum of all of the inputs. Choose a particular architecture for the beamsplitter arrangement, and justify it in terms of efficiency, simplicity, or minimization of component count. For the case of incoherent illumination, derive the optimal transmissivities of the beamsplitters in your arrangement, and prove that the chosen architecture generates an input-output relationship in the form of Equation (2). Repeat the analysis for the case of coherent illumination, and derive the equivalent of Equation (8) that characterizes the chosen architecture.

2. Two mutually coherent beams of intensities $|a|^2$ and $|b|^2$ are incident on a detector. The coherent superposition of the beams is given by, in one dimension,

$$A(x) = ae^{jk_1x} + be^{jk_2x}.$$

   (a) Plot the resulting intensity, $|A(x)|^2$, as a function of $x$.

   (b) Show that the integral of $|A(x)|^2$ over an integral number of its periods is equal to $|a|^2+|b|^2$, thus proving that the detector's response is equal to the *incoherent* sum of the individual beams.

3. For the case of a thin phase grating with a sinusoidal index modulation given by $n(x) = n_0 + n_1 sink_G x$ with $n_1 < n_0$, calculate the value of $n_1 d$ that maximizes

the diffraction efficiency into the first diffracted order, and the maximum diffraction efficiency achievable. Assume that the grating is read out by a semiconductor laser with a wavelength of 850 $nm$. For a given incident intensity of the optical readout beam, calculate the ratio of the intensity diffracted into the $0^{th}$ , $2^{nd}$, and $3^{rd}$ diffracted orders to that diffracted into the $1^{st}$ order.

4. Consider the process of diffraction from a thin *amplitude* grating with a spatially varying transmissivity and negligible phase modulation. For a sinusoidal transmittance modulation given by $t(x) = t_0 + t_1 sink_G x$ with $t_1 < t_0$, calculate the value of $t_1$ that maximizes the diffraction efficiency into the first diffracted order, and the maximum diffraction efficiency achievable. Assume that the grating is read out by a semiconductor laser with a wavelength of 850 $nm$. For a given incident intensity of the optical readout beam, calculate the ratio of the intensity diffracted into the $0^{th}$ , $2^{nd}$, and $3^{rd}$ diffracted orders to that diffracted into the $1^{st}$ order. Discuss the essential differences observed between the amplitude and phase grating cases.

5. Consider the process of holographic grating recording in a thick holographic recording medium. Assume that the grating is recorded and read out by a semiconductor laser with a wavelength of 850 $nm$, and that the angle included between the two recording beams is 30 degrees. What is the spatial frequency of the recorded grating? How thick must the hologram be in order to generate a grating parameter $Q$ (as defined in the text) of 1,000? What is the approximate angular width of this recorded grating (as measured, for example, by varying the angle of incidence of the readout beam)?

6. For the thick holographic grating described in Problem 4, calculate the amplitude of the refractive index modulation that is necessary to achieve 100% diffraction efficiency on readout. How large an absorption coefficient can be tolerated in the holographic recording medium at the readout wavelength if the absorption loss in diffraction efficiency is to be kept below 5% ?

7. Consider the holographic interconnection scheme depicted in Figure 15.9. First, derive the basic relationship for a lens that associates a given point $p_1$ in the input

plane with a resulting beam angle $\theta_1$. Given a focal length of 5 centimeters for lenses $L_1$ and $L_2$, what is the minimum spacing required between nearest neighbor points in the input plane for a grating $Q$ of 1,000? What $Q$ will be required to accommodate of order $10^4$ input positions?

8. Design a differentiating circuit for incorporation in an optically addressed spatial light modulator, using the principles of Chapters 14 and 15. Assume that the detector is a *p-i-n* photodiode, and that the modulator can be treated as a purely capacitive load. If the modulator can be modeled as a parallel plate capacitor of dimensions $30 \times 50$ $\mu m$, with a thickness of 1 $\mu m$ and a relative dielectric constant typical of gallium arsenide multiple quantum well devices ($\epsilon = 13$), estimate the bandwidth over which the differentiator is operational.

9. If storage of one synaptic weight in VLSI requires a memory element 15 $\mu m \times 15$ $\mu m$ in size, what is the maximum number of synaptic weights that can be implemented on a 1 $cm \times 1$ $cm$ chip? If a VLSI neural network is fully connected using one such memory element for each synapse, how many pins are required for input to, and output from, the set of neuron units? How many pins would be required for parallel input of the weights? If an optical synaptic weight can be implemented in an effective volume of 5 $\mu m \times 5$ $\mu m \times 5$ $\mu m$, what is the maximum number of synaptic weights that can be implemented in a volume 1 $cm \times 1$ $cm \times 1$ $cm$?

10. (a) If a neural network is simulated on a digital, sequential machine, how many multiply operations and add operations are required to simulate the computational process of a single-layer feedforward network with $N$ neuron units and a connectivity (number of connections per neuron unit) of $M$? If a multiply operation can be performed in 100 $ns$ and an add operation in 25 $ns$, what is the minimum time it will take for one pass through the network if $N = 10^6$ and $M = 10^4$?

    (b) For the outer product learning of Equation (24), neglecting the decay term, how many multiply operations and add operations are required for one iteration of weight updates, in terms of $M$ and $N$? For a two-layer network (1 hidden

layer), with each layer having $N = 10^6$ and $M = 10^4$, and assuming $10^4$ different patterns presented 1,000 times each, how long will the network take to be trained (assuming 1 forward pass and 1 update of all weights per presentation, 100 $ns$ per multiply operation and 25 $ns$ per add operation)?

(c) For the same numbers of (b), assuming each layer of a photonic system can perform a forward pass in 100 $ns$ and a set of parallel weight updates in 1 $\mu s$, how long will it take to be trained?

11. (a) Design an algorithm that uses only nonnegative signals outside of each neuron unit, nonnegative weights, and allows two separate inputs to and two separate outputs from, each neuron unit. It should be able to perform neural computation and weight updates for learning. Your answer should be in the form of a flow chart. You may perform subtraction and division only within each neuron unit; only addition, multiplication, interconnection and storage can be performed external to the neuron units. (No need to simulate.)

(b) After many iterations during learning, might there be a problem with weights saturating or going out of bounds? If not, why not? If so, conjecture as to how this problem might be avoided.

12. In regard to Equation (24), find an expression for $\delta_i$ for the following algorithms:

(a) Perceptron

(b) Widrow-Hoff

(c) Least minimum square (back propagation), for a multi-layer net for

(i) output layer

(ii) hidden layers

Assume that $\beta = 0$ for this problem.

(Note: This problem requires familiarity with neural networks not discussed in this chapter.)

13. Referring to the system of Figure 15.21, if the interconnection mask SLM were electronically addressed with a serial line, capable of transmitting analog values at 50 *MHz*, and there are $10^6$ pixels (analog weights), what is the maximum frame rate? If there are $10^3$ parallel lines addressing? If the SLM is optically addressed, what limits the frame rate?

14. (a) Referring to Figure 15.23, show how the following network can be drawn as a single layer network with feedback.

    << Insert Figure for Problem 14(a) >>

    (b) How can the architecture shown in Figure 15.26 be modified to include feedback connections within the module? Sketch the resulting architecture.

15. Consider a charge-coupled-device array as shown schematically in Figure 15.34, of dimension 1000 × 1000 pixels with a serial readout buffer that operates at a clock frequency of 100 *MHz*. Calculate the maximum frame rate achievable, and the speed required of the row shift registers. Calculate the ratio between the total storage time required of the first pixel read out to that of the last in each frame.

## FIGURE CAPTIONS

**Fig. 15.1** Illustration of optical addition utilizing a 50/50 beamsplitter: (a) collinear *incoherent* beam geometry; (b) collinear *coherent* beam geometry, showing input and output *amplitudes*; (c) collinear *coherent* beam geometry, showing input and output *intensities*.

**Fig. 15.2** Illustration of optical addition utilizing mirrors: (a) angularly multiplexed *incoherent* beam geometry; (b) angularly multiplexed *coherent* beam geometry.

**Fig. 15.3** Illustration of optical multiplication utilizing a medium with variable transparency.

**Fig. 15.4** Fundamental principles of spatial light modulator function: (a) block diagram of the principal functions of an optically-addressed spatial light modulator, including the detection, functional implementation, and modulation functions; (b) schematic diagram of an $N \times N$ array of spatial light modulator pixels, in which three pixels are shown in different transmission states; (c) expanded view of the pixel array, showing an incomplete fill factor within each pixel; (d) expanded view of a single pixel within the array, illustrating one possible pixel configuration that incorporates two detector elements $D_1$ and $D_2$, control electronics for impedance matching and functional implementation, and two modulator elements, shown here in different transmittance states.

**Fig. 15.5** Examples of monolithically-integrated spatial light modulators. The chosen examples incorporate photodetectors, control circuitry, and multiple quantum well modulators within each pixel on a single gallium arsenide ($GaAs$) substrate. In (a), the control electronics and photodetector elements are fabricated following the photolithographic definition and physical isolation of the modulator elements, while in (b) a buffer (isolation) layer is used to allow fabrication and interconnection of all of the elements without chemical or ion beam etching.

**Fig. 15.6** Example of a hybrid spatial light modulator, in which the photodetectors and control electronics are fabricated on a silicon substrate, and the multiple quantum well modulator elements are fabricated on a gallium arsenide ($GaAs$) substrate. The two sets of devices are bump contacted on a pixel-by-pixel basis to provide parallel electrical continuity.

**Fig. 15.7** VLSI layout of a generalizable silicon-based spatial light modulator structure: (a) neuron pixel layout; (b) photograph of a single neuron unit in VLSI implementation, with probe pads substituted for the two detectors (bottom) and for contact to the two modulation elements (top); (c) photograph of a $6 \times 6$ array of neuron units on a VLSI chip that incorporates additional test circuitry.

**Fig. 15.8** A simplified holographic recording configuration: case of plane wave signal and

reference beams, and a *thin* holographic recording medium; (a) recording, and (b) reconstruction with a plane wave readout beam.

**Fig. 15.9** A point-to-point interconnection system, using a holographic optical element (HOE) for interconnection routing, and lenses as position-to-angle and angle-to-position encoders. In this example, the holographic optical element effectively performs an input angle to output angle transformation, such that light emitted (or transmitted) at point $p_1$ in the input plane ($P_1$) is detected at point $p_2$ in the output plane ($P_2$).

**Fig. 15.10** Volume holographic recording with plane wave signal and reference beams; (a) recording, and (b) reconstruction, showing the elimination of the higher diffracted orders.

**Fig. 15.11** The angular alignment sensitivity of a volume holographic optical element, as a function of the dimensionless $Q$-parameter defined in the text. The grating strength for all of the curves (3.14 radians) is optimized to produce 100% diffraction efficiency in the limit of large $Q$ (Bragg diffraction regime), and is not optimized for low $Q$ gratings. Note that the diffraction efficiency is essentially independent of angle for low $Q$ gratings, and is very strongly peaked at the Bragg angle (7.5 degrees in this case) for high $Q$ gratings.

**Fig. 15.12** The diffraction efficiency of thin (Raman-Nath diffraction regime) and thick (Bragg diffraction regime) holographic gratings as a function of the grating strength.

**Fig. 15.13** Schematic representation of a 4 input, 4 output holographic interconnection, showing 4 coherent input beams $x_1$-$x_4$ and 4 coherent recording beams $y_1$-$y_4$, each of which corresponds to a desired output $y_1'$-$y_4'$. In (a), the sets $\{x_j\}$ and $\{y_i\}$ interfere within the volume holographic medium, recording the desired interconnection diffraction gratings. In (b), a new set of input beams $\{x_j\}$ illuminates the volume holographic medium, reading out the weighted interconnection pattern and forming appropriately weighted sums at each of the outputs $\{y_i'\}$.

**Fig. 15.14** Schematic representation of the fan-out process for optical beams, for the case of one input and three outputs: (a) with beamsplitters ($BS_1$ - $BS_3$); (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings.

**Fig. 15.15** Schematic representation of the fan-in process for optical beams, for the case of three angularly distinct inputs and one combined collinear output beam: (a) with beamsplitters, showing the unavoidability of a throughput loss associated with the set of transmitted (and multiply reflected) beams; (b) with a single holographic optical element containing three multiplexed (spatially superimposed) diffraction gratings, showing an analogous throughput loss.

**Fig. 15.16** Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: recording/readout configuration. The input beams $\{x_j\}$ are assumed to have interfered within the volume holographic medium with the set of recording beams $\{y_i\}$, producing the desired set of interconnection gratings with weights $w_{ij}$. Illumination of the volume holographic medium with beam $x_1$ produces a 1 to 4 fanout into the output beams $\{y_i'\}$, as well as the zeroth order beam $x_1'$. Due to the effects of beam degeneracy, power is also coupled into the zeroth order beams $x_2'$-$x_4'$, and crosstalk terms $\{c_i\}$ are introduced into the outputs.

**Fig. 15.17** Illustration of the generation of crosstalk in holographic optical interconnections due to beam degeneracy: diffraction efficiency as a function of grating strength for the readout configuration of Figure 15.16. Shown are the depletion of the zero order beam $x_1'$ and the rise of the desired output beams $y_i'$, accompanied by a strong buildup of the cross-coupled beams $x_2'$-$x_4'$.

**Fig. 15.18** Illustration of a surface-emitting laser diode source array [after Jewell, 1990]. In this example, the individual semiconductor laser diodes are isolated by chemically assisted ion beam etching techniques, must be individually contacted, and emit *through* the *GaAs* substrate.

Fig. **15.19** Schematic diagram of a photodarlington pair utilized as a high gain detector/amplifier combination.

Fig. **15.20** Schematic diagram of a charge coupled device (CCD) photodetector array fabricated on a silicon substrate. Electrostatic potential wells are created by application of appropriate voltages to the three phase bias electrode structure, with electrical isolation provided by the gate oxide layer. Light incident through the transparent electrodes creates stored charge that can be transferred to an output signal terminal by proper sequential phasing of the bias voltages ($P_1$ - $P_3$).

Fig. **15.21** Block diagram of a 1-D to 1-D photonic neural network, in which a one-dimensional neuron array is fully interconnected to a one-dimensional detector array by means of a two-dimensional interconnection mask.

Fig. **15.22** Block diagram of a 2-D to 2-D photonic neural network, in which a two-dimensional neuron array is fully interconnected to a two-dimensional output array by means of a three-dimensional volume holographic optical interconnection mask. The input plane, output plane, and optional training plane are shown. Many variants of this geometry with similar properties are possible.

Fig. **15.23** A single layer physical neural network with feedback, used to implement a multilayer recurrent functional network. The solid boxes indicate feedforward connections, and the broken boxes indicate lateral connections.

Fig. **15.24** Incoherent/coherent technique for recording and reconstructing multiple holograms simultaneously, in which all solid lines represent mutually coherent beams, and all broken lines represent a separate set of mutually coherent beams: (a) recording; (b) reconstruction; and (c) holographic representation, in which each hologram represents the fanout from a given neuron unit.

Fig. **15.25** Doubly angularly multiplexed volume holographic optical interconnection, designed to circumvent the effects of beam degeneracy. The mutually incoherent input

beams ($\{x_j\}$) are angularly multiplexed over $j$, as are the corresponding sets of output beams from the training plane ($\{\delta_i^{(j)}\}$) generated by the coherent sources $S_j$, to produce an angularly multiplexed fan-in at each summed output, thus yielding the neuron activation potentials $\{\rho_i\}$.

Fig. **15.26** Photonic architecture for neural network implementation that incorporates a parallel source array, double angular multiplexing, and incoherent/coherent recording and reconstruction; the Hebbian case is depicted.

Fig. **15.27** Photonic architecture for neural network implementation: recording configuration. This configuration implements the learning function in the photonic architecture of Figure 15.26. The sets of beams emitted from the source array (two are shown) interfere in the volume holographic medium to update the weights stored in the interconnection holograms.

Fig. **15.28** Photonic architecture for neural network implementation: reconstruction configuration. This configuration implements a single forward pass of the computing function in the photonic architecture of Figure 15.26. The lower set of beams acts as a set of reference beams, and generates a set of weighted output arrays that are imaged onto the detector array. Each stored hologram is reconstructed by a single neuron unit $x_j$, and fans out with appropriate weights to illuminate the detector array. The full set of reconstructed holograms sums within each pixel to yield the neuron activation potentials $\{\rho_i\}$.

Fig. **15.29** Generalized photonic architecture for neural network implementation, including provision for the generation of arbitrary training terms ($\delta_i$) .

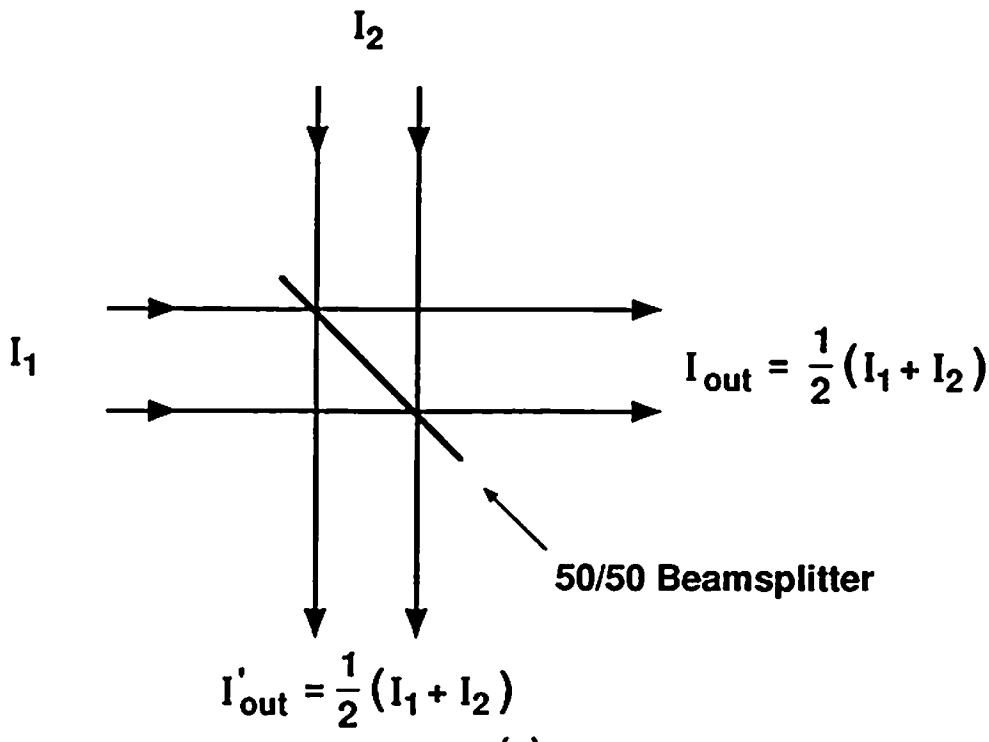
Fig. **15.30** Schematic diagram of a dual-input, dual-output differential amplifier that effects a sigmoid-like transfer characteristic.

Fig. **15.31** Experimentally obtained transfer characteristics from the circuit shown in Figure 15.30, showing the output voltage in both channels ($V_{out1}$ and $V_{out2}$) as a function of one input voltage ($V_{in1}$), with the other input voltage ($V_{in2}$) as a parameter.

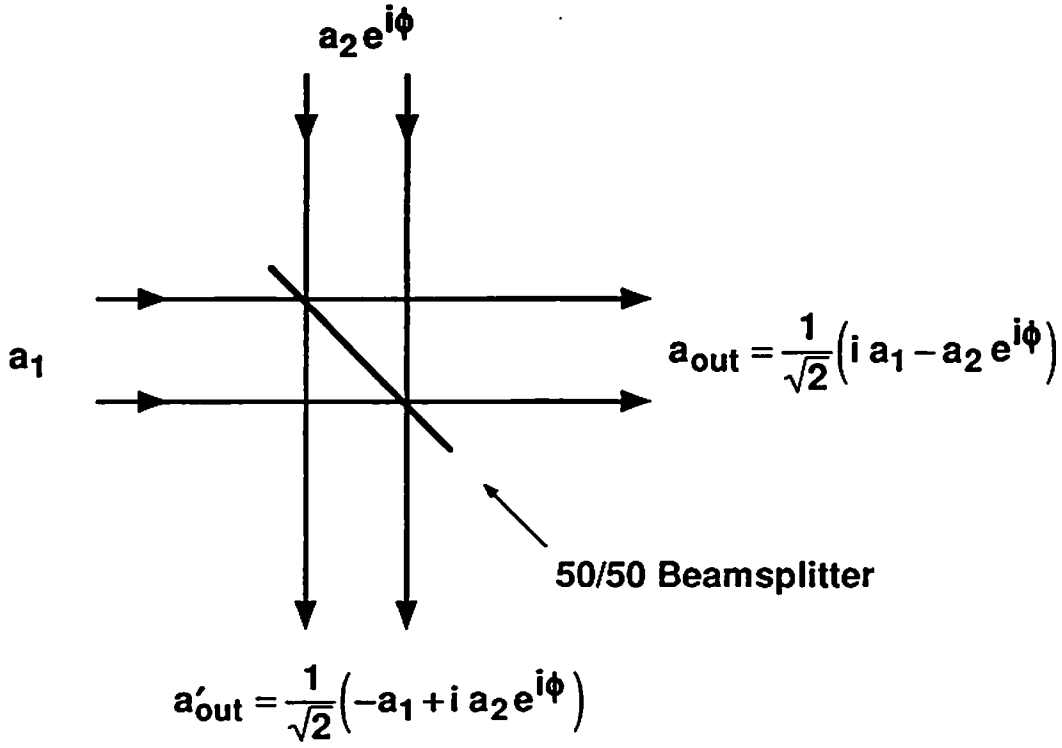
**Fig. 15.32** Optical layout for copying the entire contents of a three-dimensional volume holographic optical element (VHOE) into a second VHOE, utilizing a two-dimensional array of individually coherent, but mutually incoherent sources.

**Fig. 15.33** The single pixel probability of error $P(Err)$ as a function of the number of photons detected within each pixel, for the cases of 100 and 1000 analog grey levels.

**Fig. 15.34** Illustration of parallel-to-serial conversion in two-dimensional detector arrays such as the charge-coupled-device (CCD) array. Charge accumulated within each photosensitive region during exposure is transferred laterally by a set of row-parallel shift registers to a high speed parallel-to-serial shift register, which reads out the entire array one column at a time.




(a)




(b)

Figure 15.1 (a), (b)

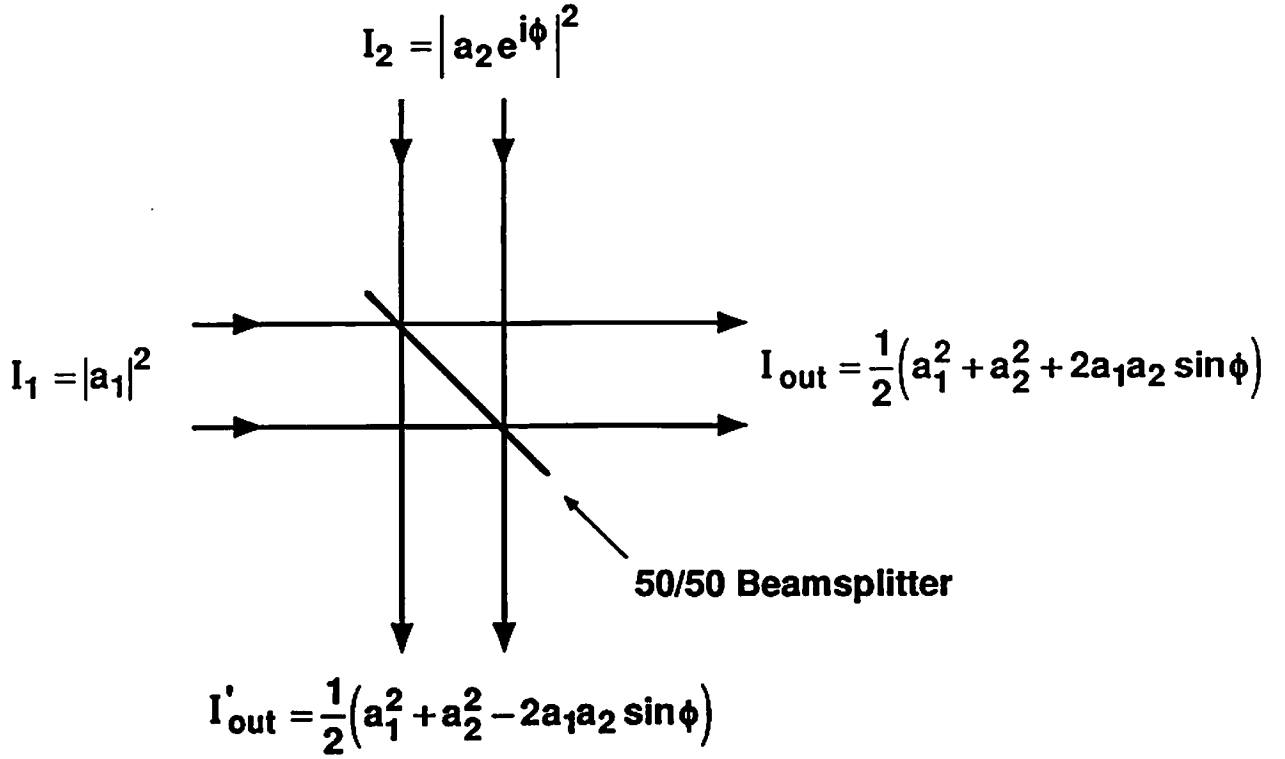
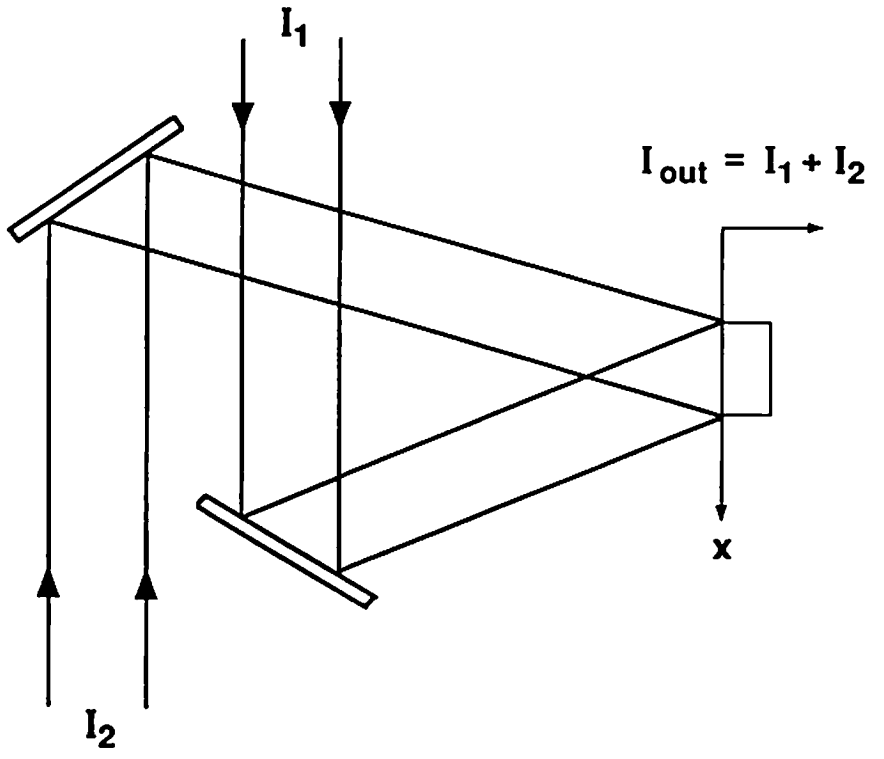

$I_2 = |a_2 e^{i\phi}|^2$
$I_1 = |a_1|^2$
$I_{out} = \frac{1}{2}(a_1^2 + a_2^2 + 2a_1 a_2 \sin\phi)$
50/50 Beamsplitter
$I'_{out} = \frac{1}{2}(a_1^2 + a_2^2 - 2a_1 a_2 \sin\phi)$

Figure 15.1 (c)

(a)




(b)

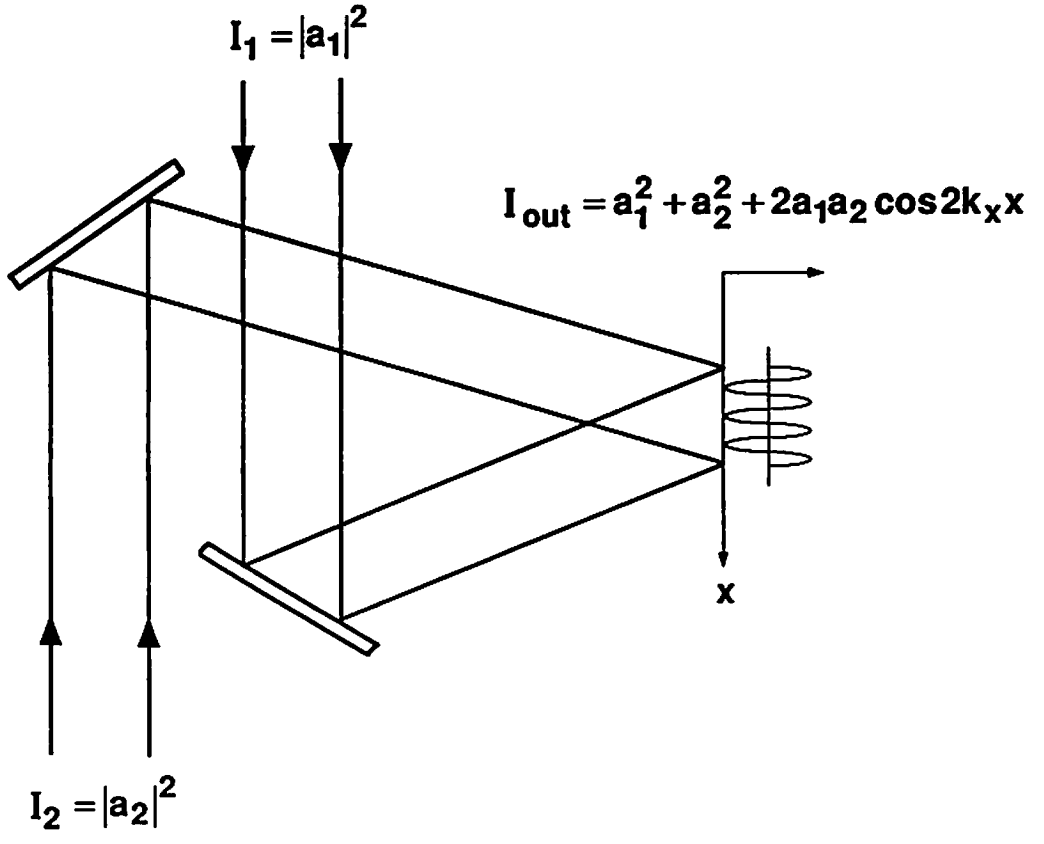
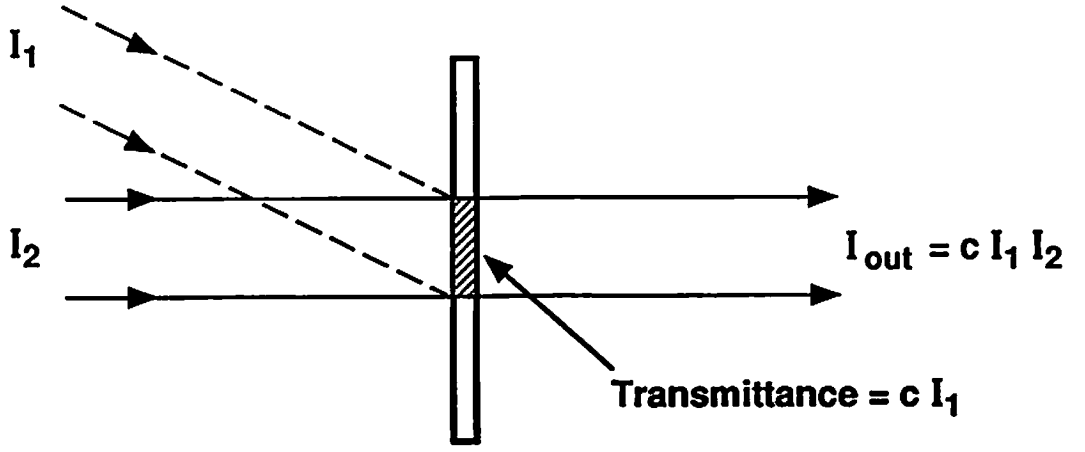



Figure 15.2




Transparency

Figure 15.3

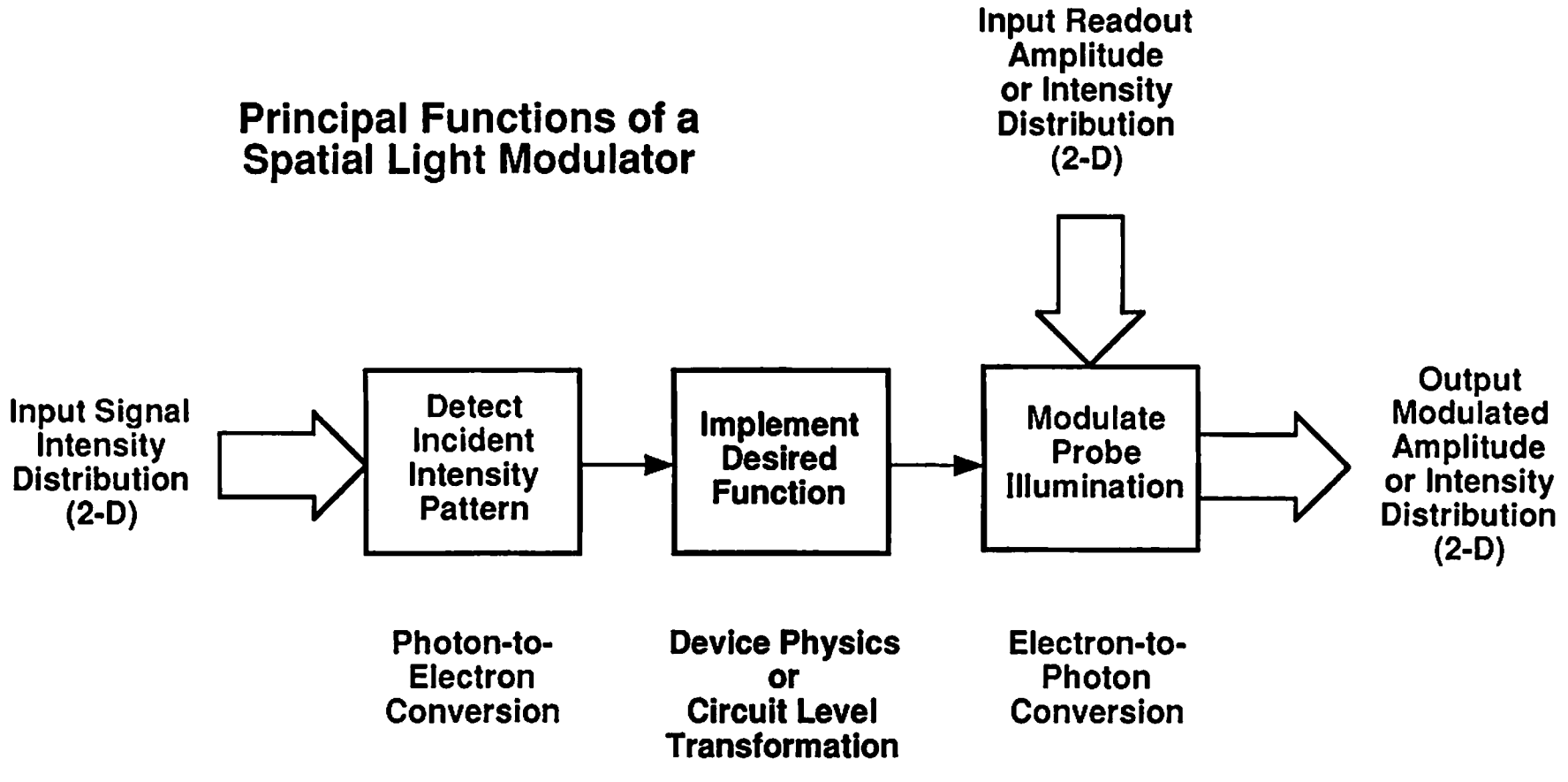

Principal Functions of a Spatial Light Modulator
Input Readout Amplitude or Intensity Distribution (2-D)
Input Signal Intensity Distribution (2-D)
Detect Incident Intensity Pattern
Implement Desired Function
Modulate Probe Illumination
Output Modulated Amplitude or Intensity Distribution (2-D)
Photon-to-Electron Conversion
Device Physics or Circuit Level Transformation
Electron-to-Photon Conversion
(a)

Figure 15.4(a)

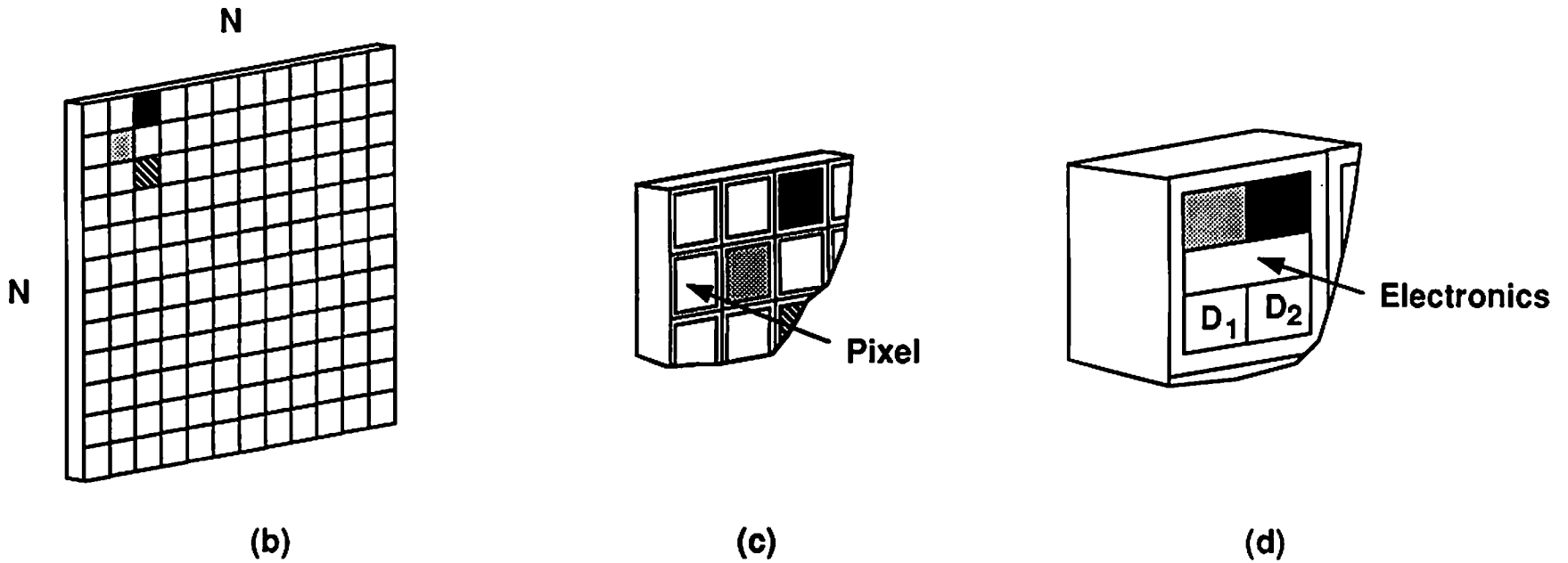
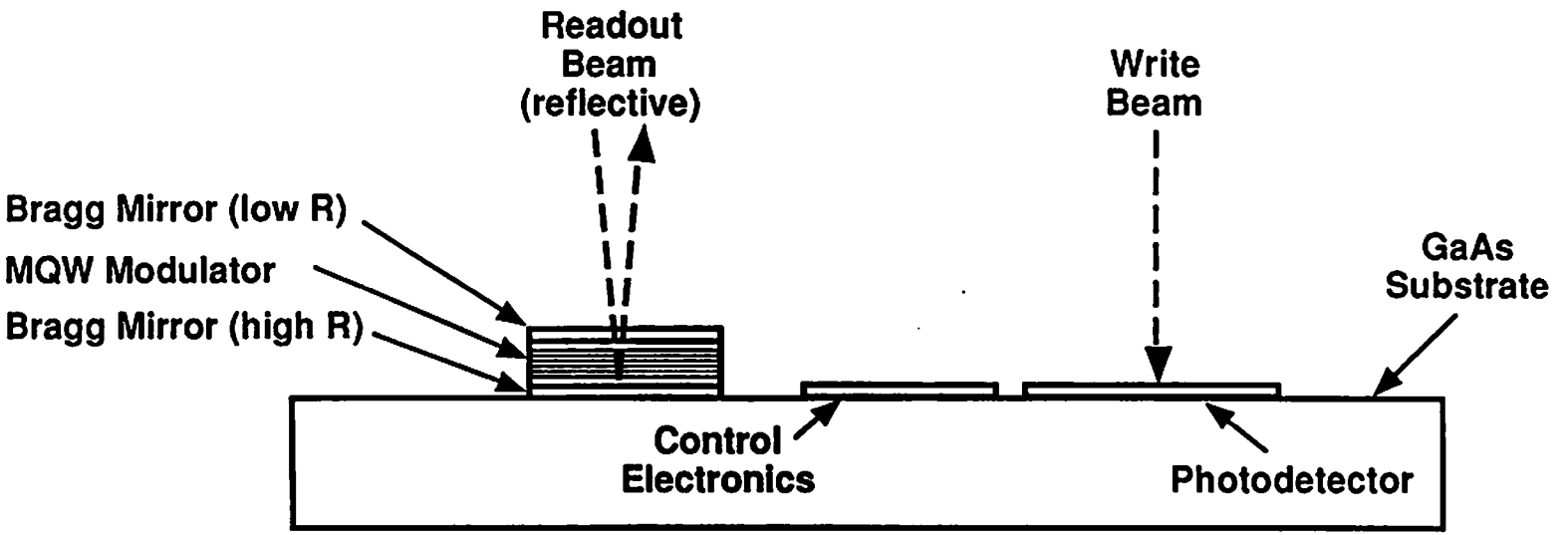
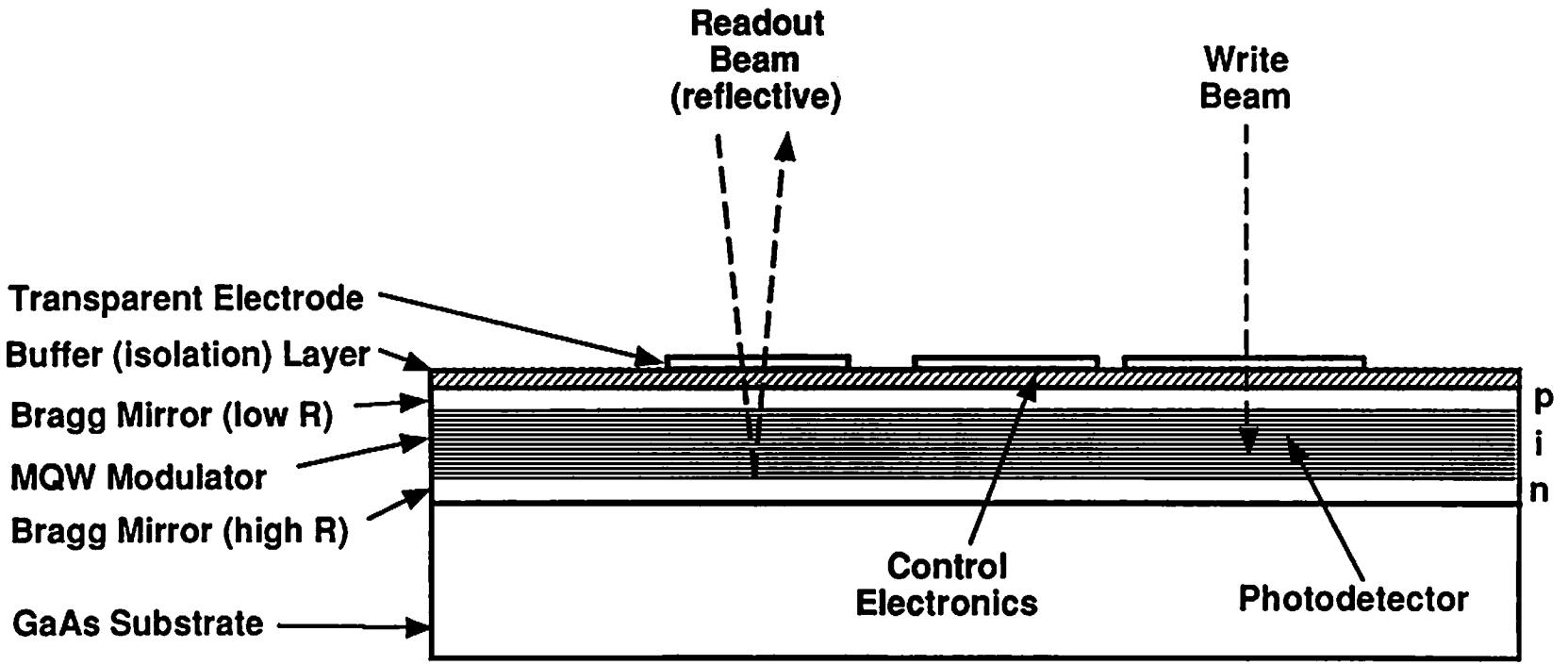



Figure 15.4 (b), (c), (d)




(a)

Figure 15.5(a)




(b)

Figure 15.5(b)

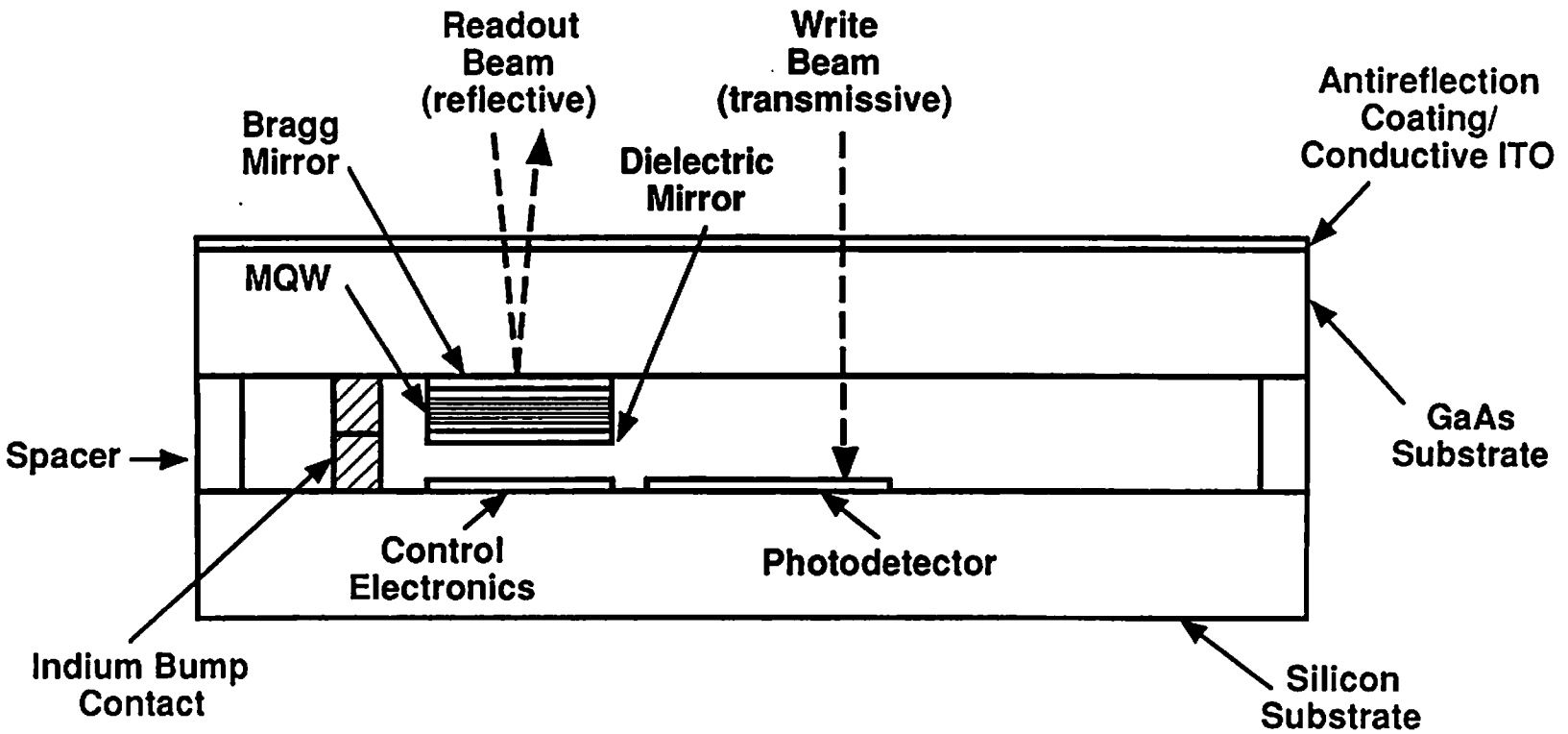

Readout
Beam
(reflective)
Write
Beam
(transmissive)
Antireflection
Coating/
Conductive ITO
Bragg
Mirror
Dielectric
Mirror
MQW
Spacer
GaAs
Substrate
Control
Electronics
Photodetector
Indium Bump
Contact
Silicon
Substrate

Figure 15.6

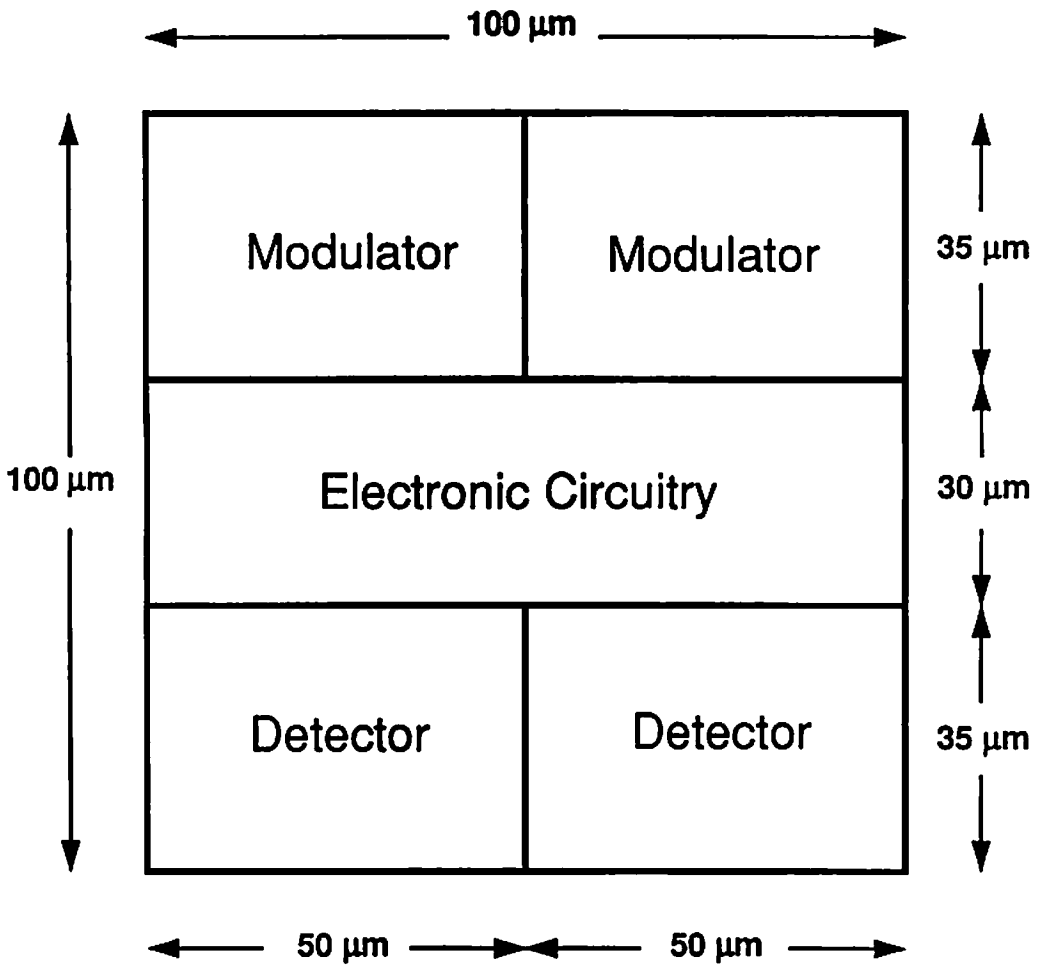

100 µm
Modulator
Modulator
Electronic Circuitry
Detector
Detector
100 µm
35 µm
30 µm
35 µm
50 µm
50 µm

**Figure 15.7 (a)**

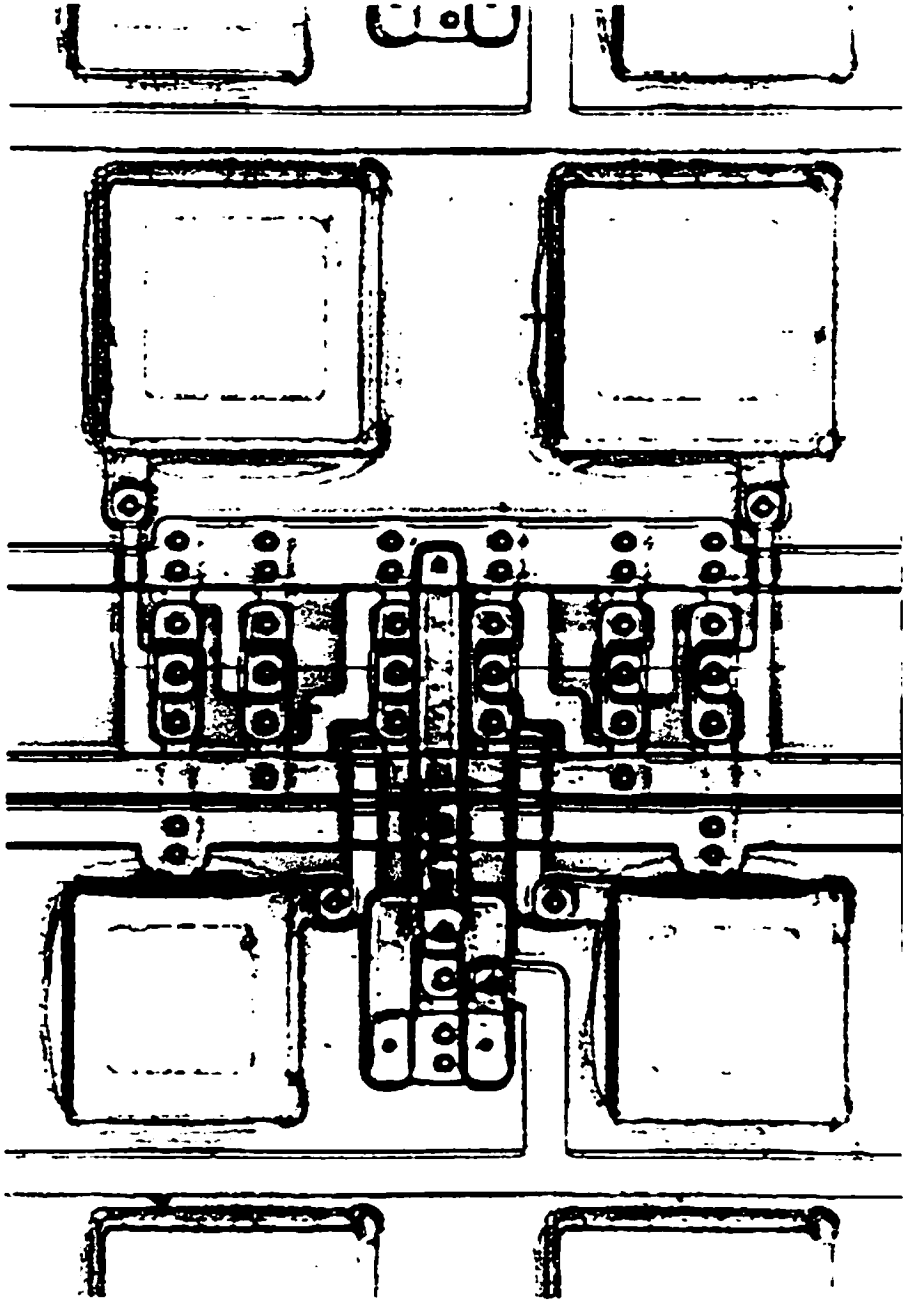


Figure 15.7 (b)

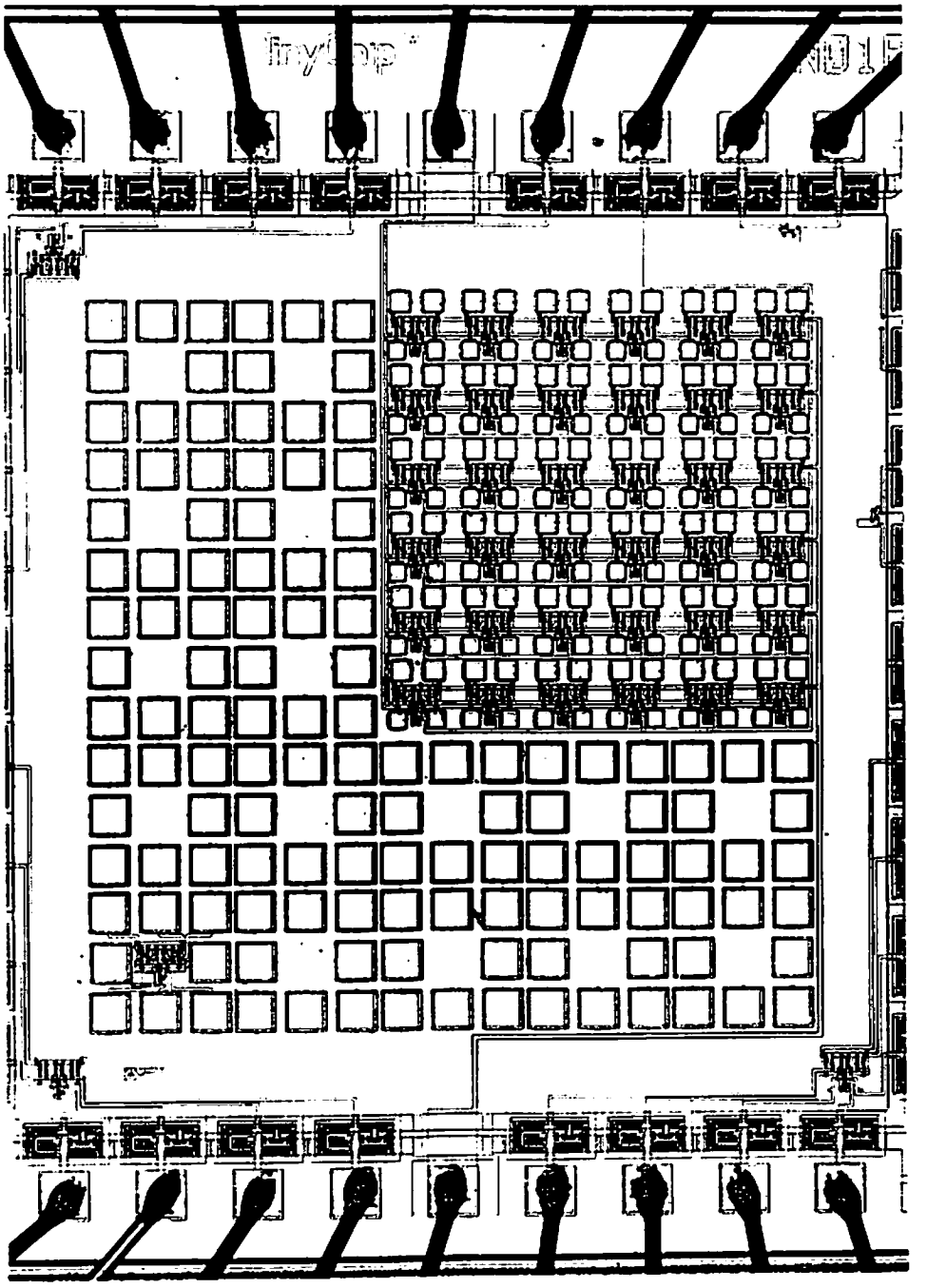


Figure 15.7 (c)

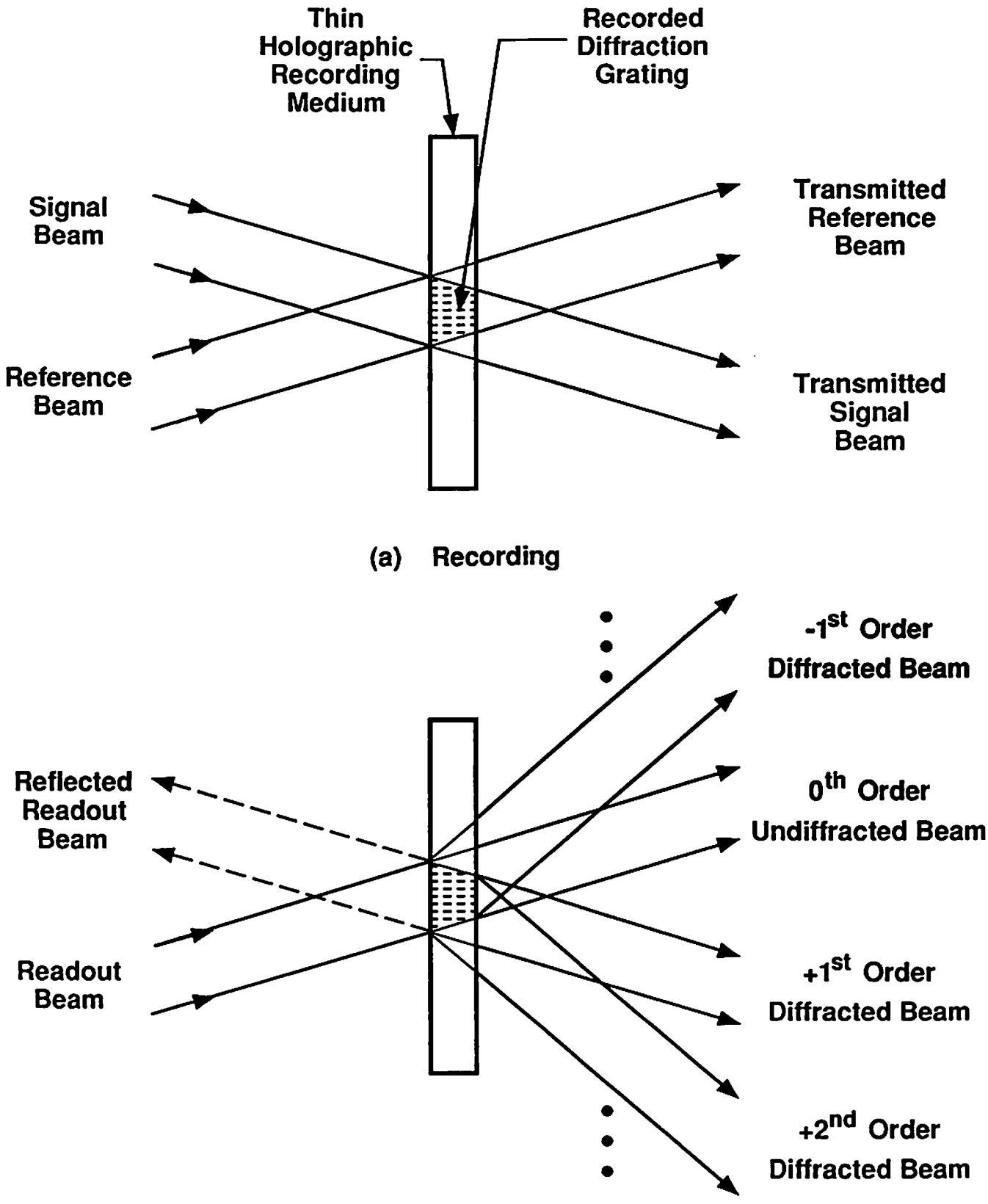



Figure 15.8 (a), (b)

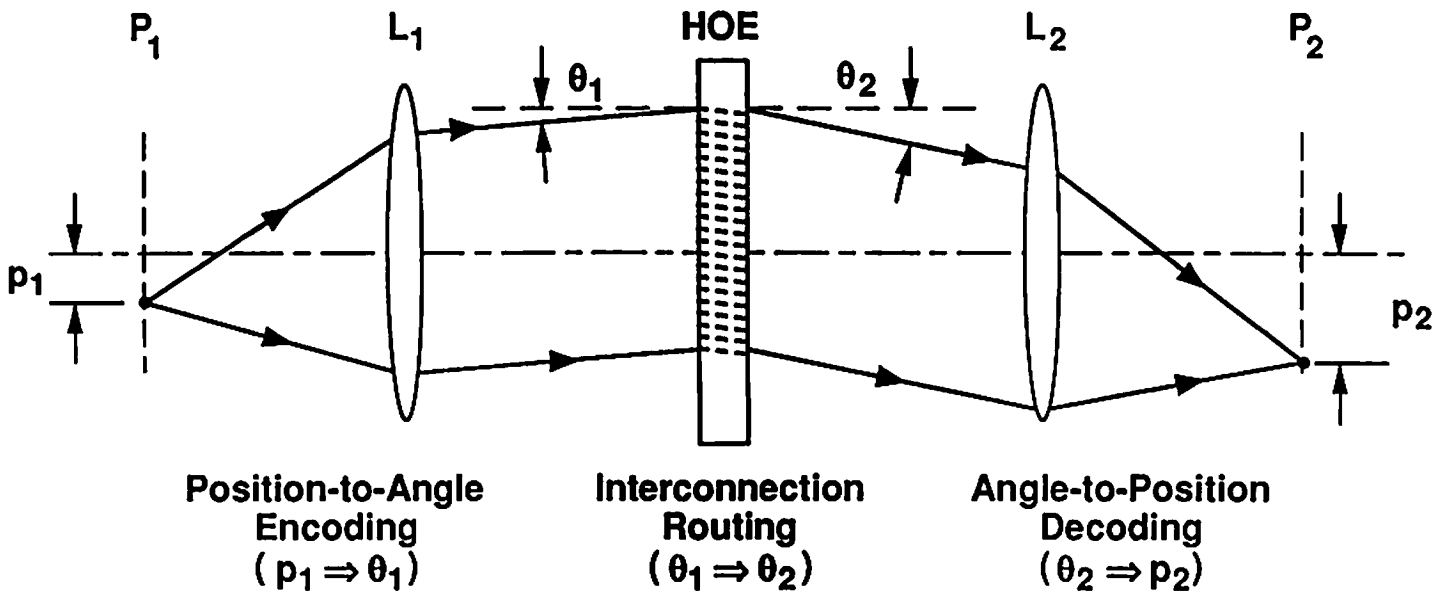
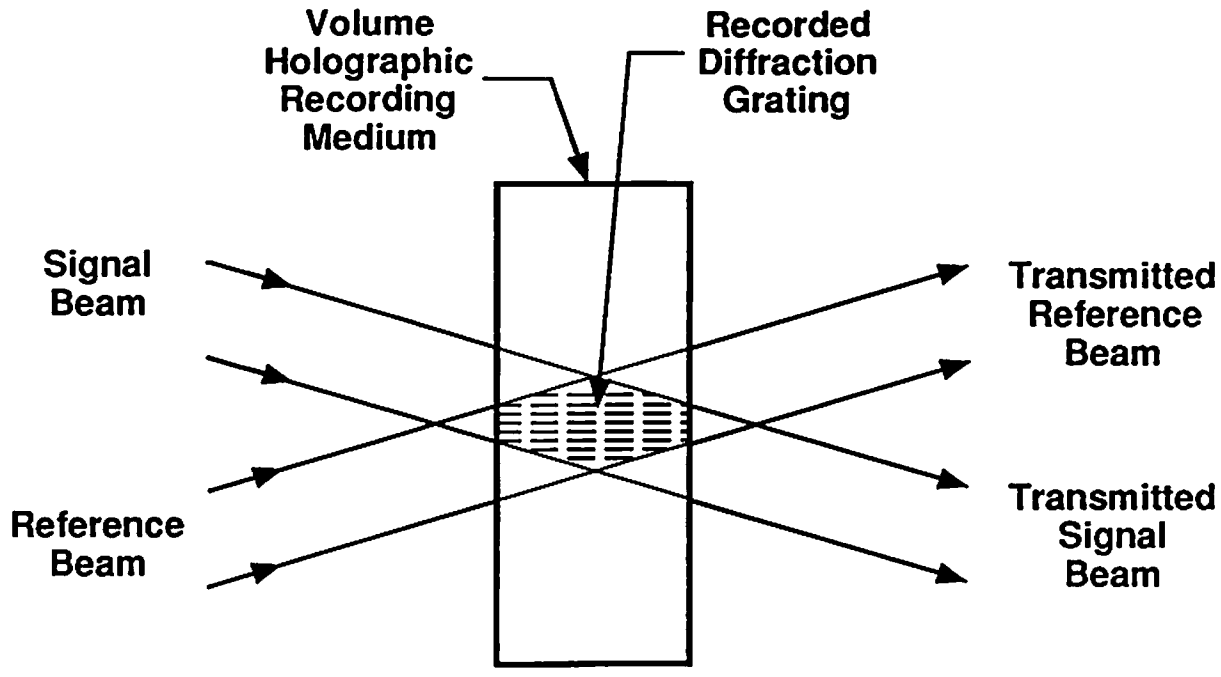
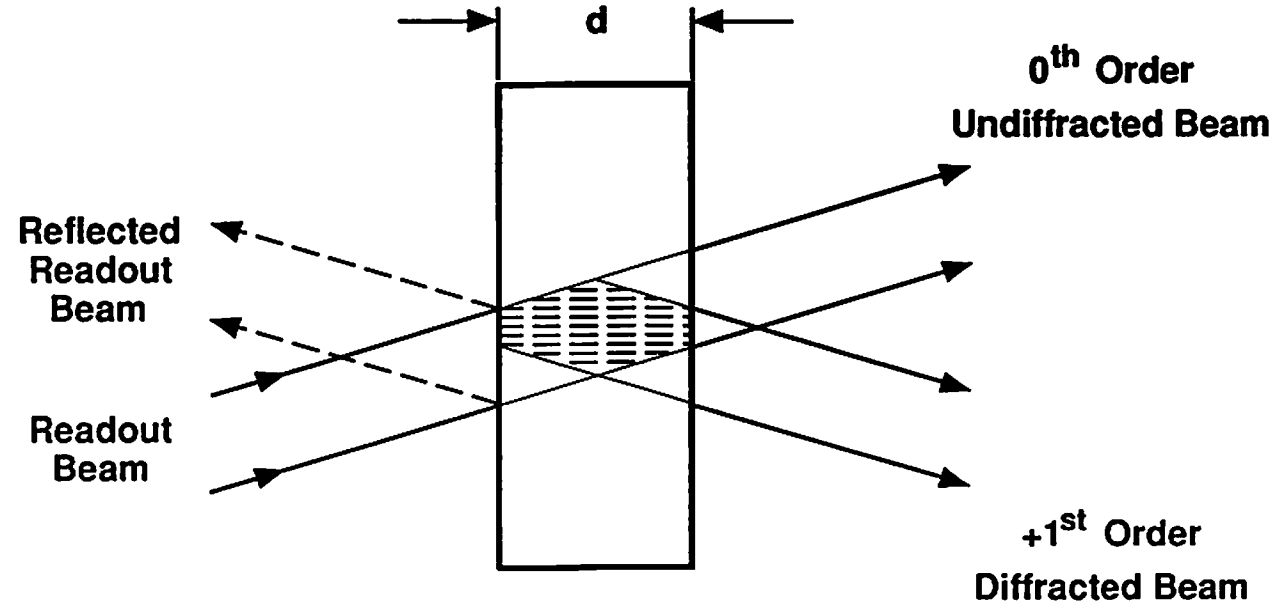

P1
L1
HOE
θ1
θ2
L2
P2
p1
p2
Position-to-Angle
Encoding
( p1 ⇒ θ1 )
Interconnection
Routing
( θ1 ⇒ θ2 )
Angle-to-Position
Decoding
( θ2 ⇒ p2 )

Figure 15.9




(a) Recording




(b) Reconstruction (readout)

Figure 15.10 (a), (b)

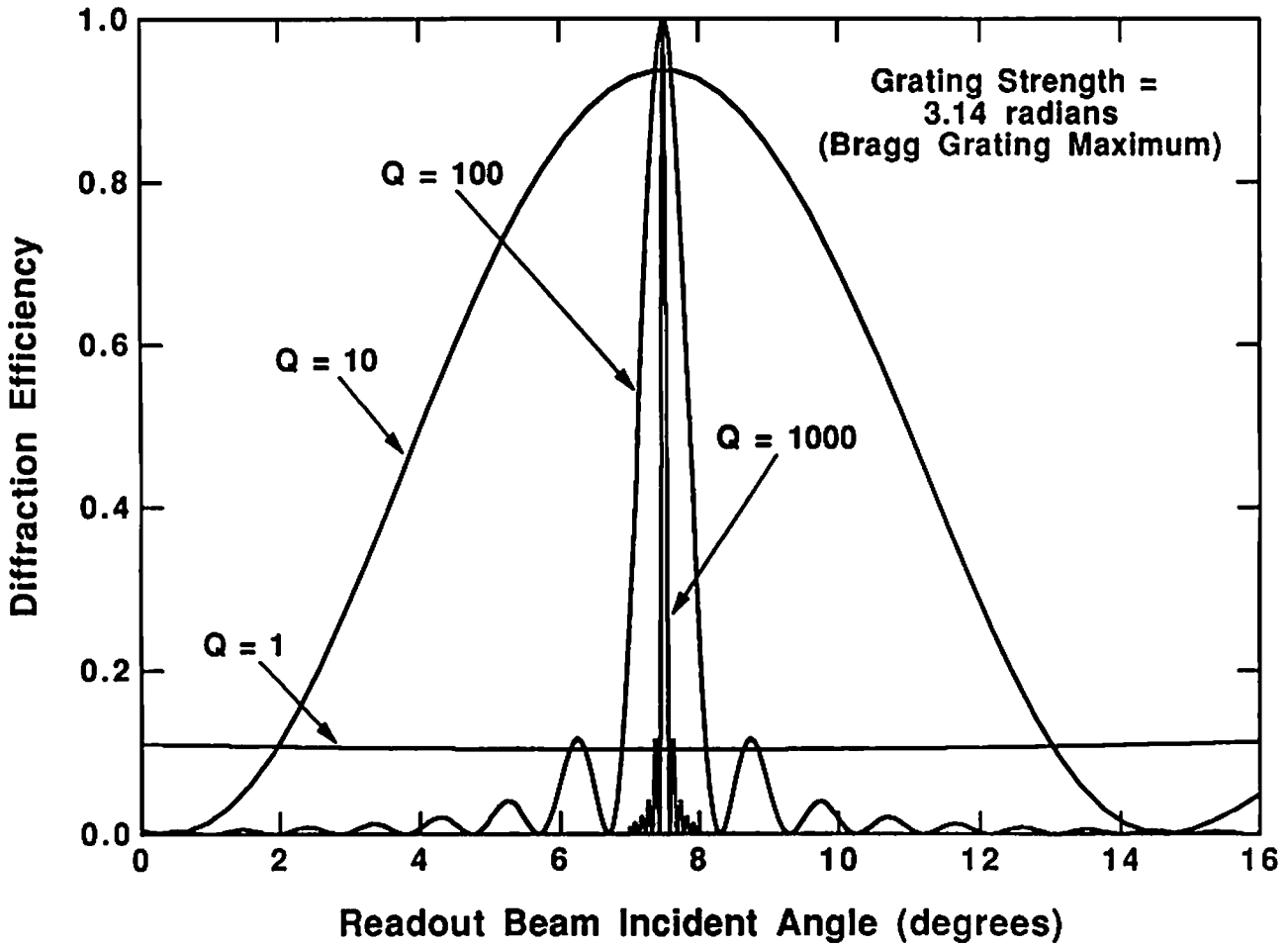

1.0
0.8
0.6
0.4
0.2
0.0
Diffraction Efficiency
Grating Strength =
3.14 radians
(Bragg Grating Maximum)
Q = 100
Q = 10
Q = 1000
Q = 1
0
2
4
6
8
10
12
14
16
Readout Beam Incident Angle (degrees)

Figure 15.11

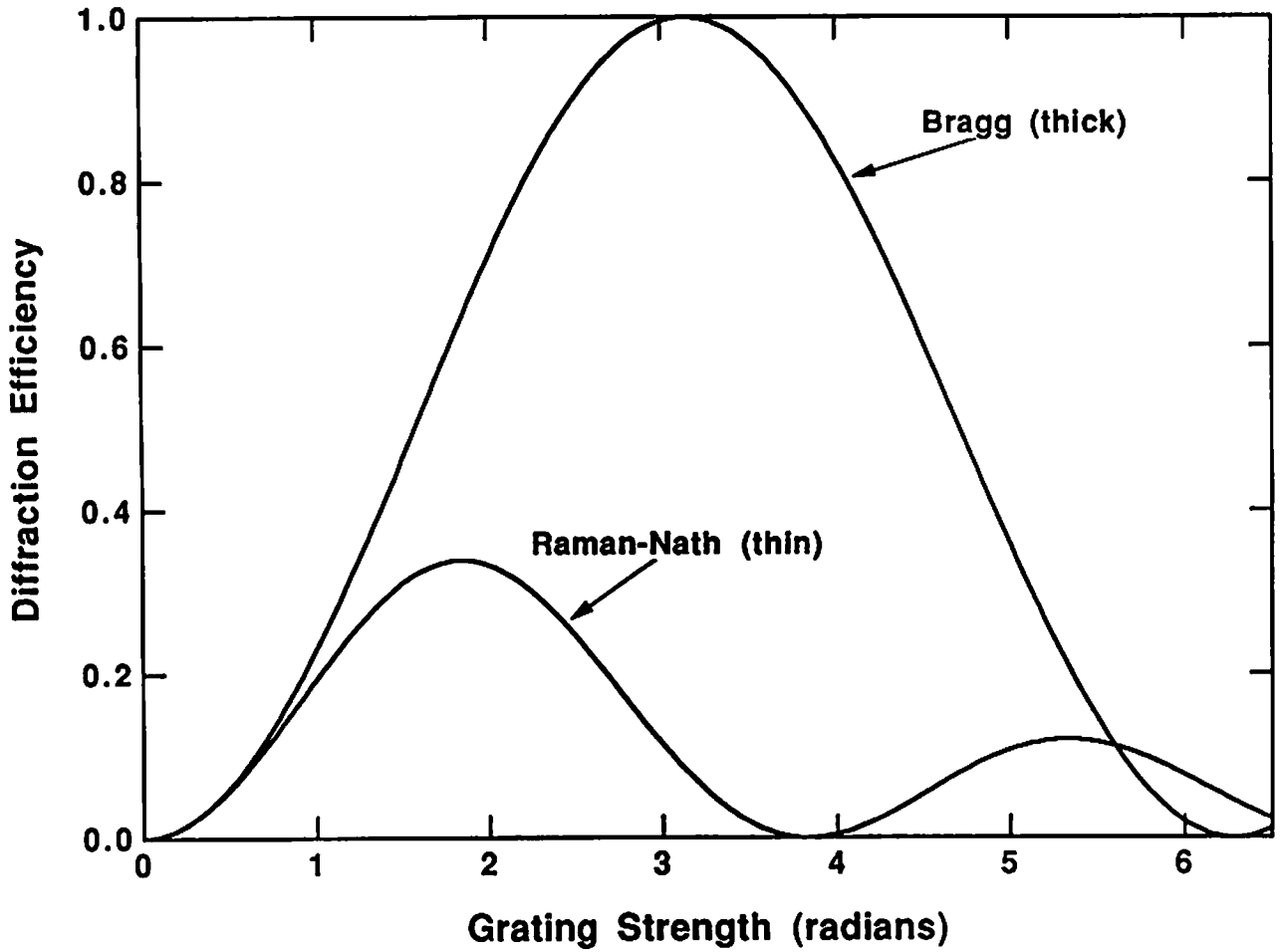

1.0
0.8
0.6
0.4
0.2
0.0
Diffraction Efficiency
Bragg (thick)
Raman-Nath (thin)
0
1
2
3
4
5
6
Grating Strength (radians)

Figure 15.12

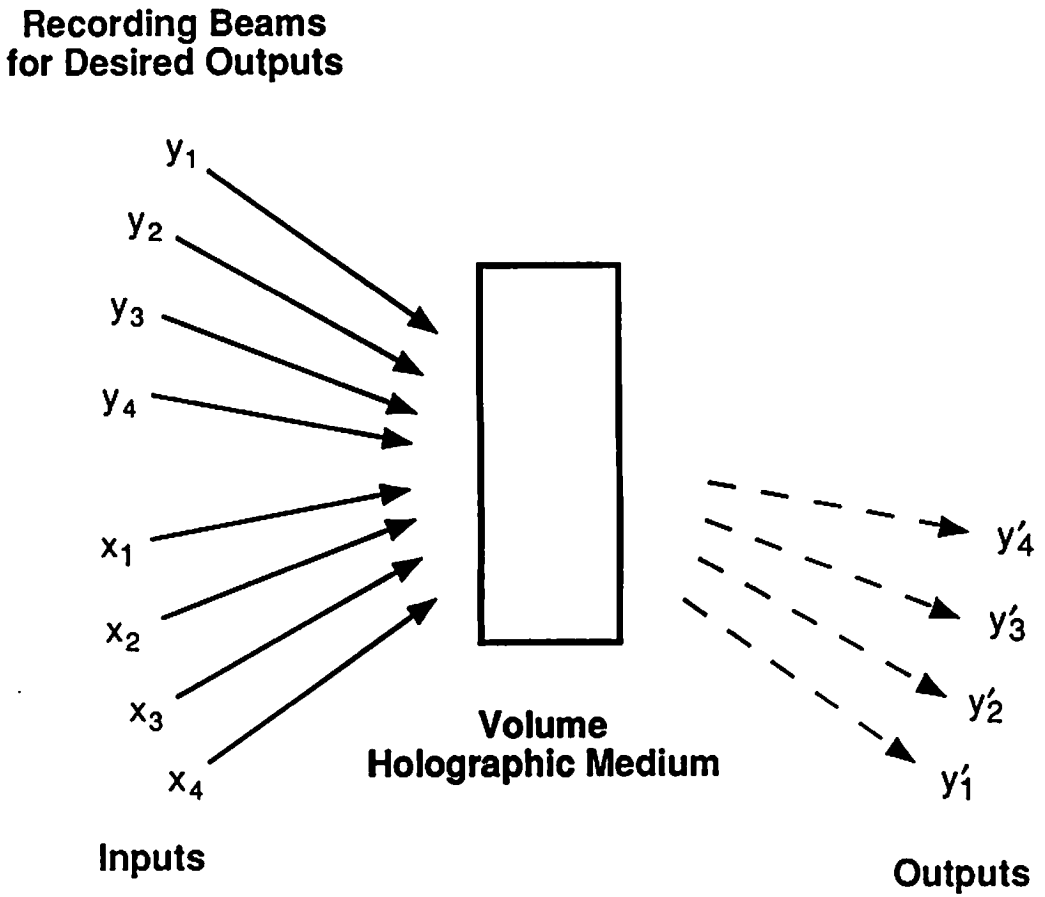



Figure 15.13(a)

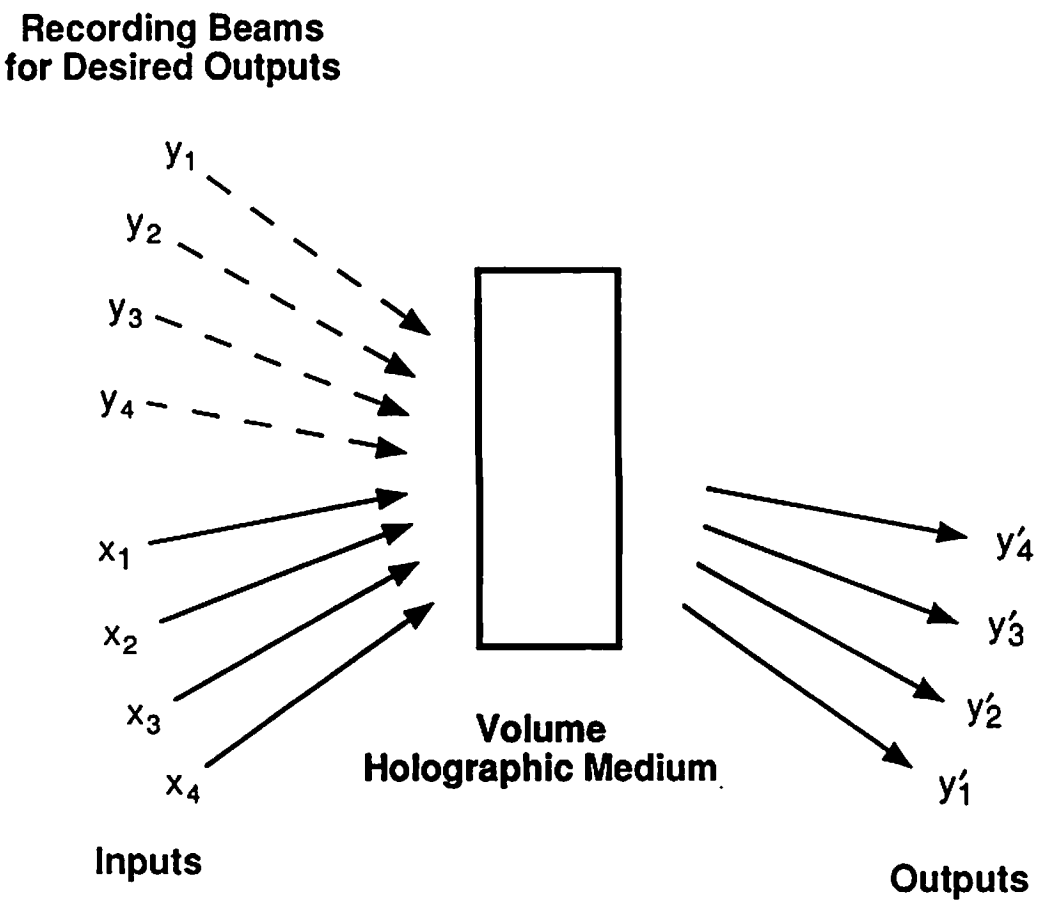



Figure 15.13(b)

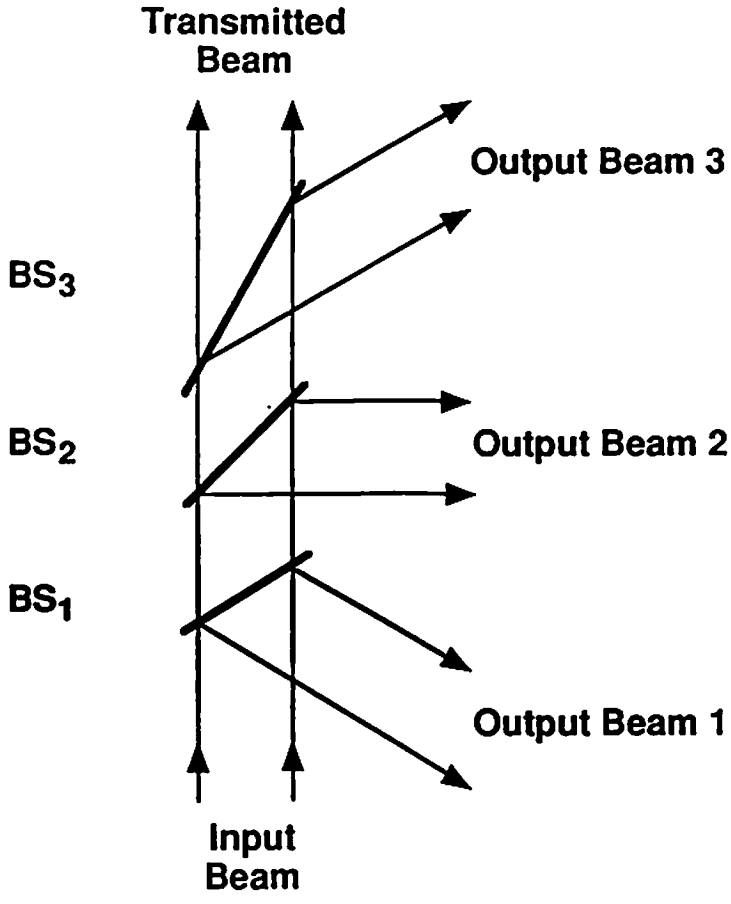



Figure 15.14 (a)

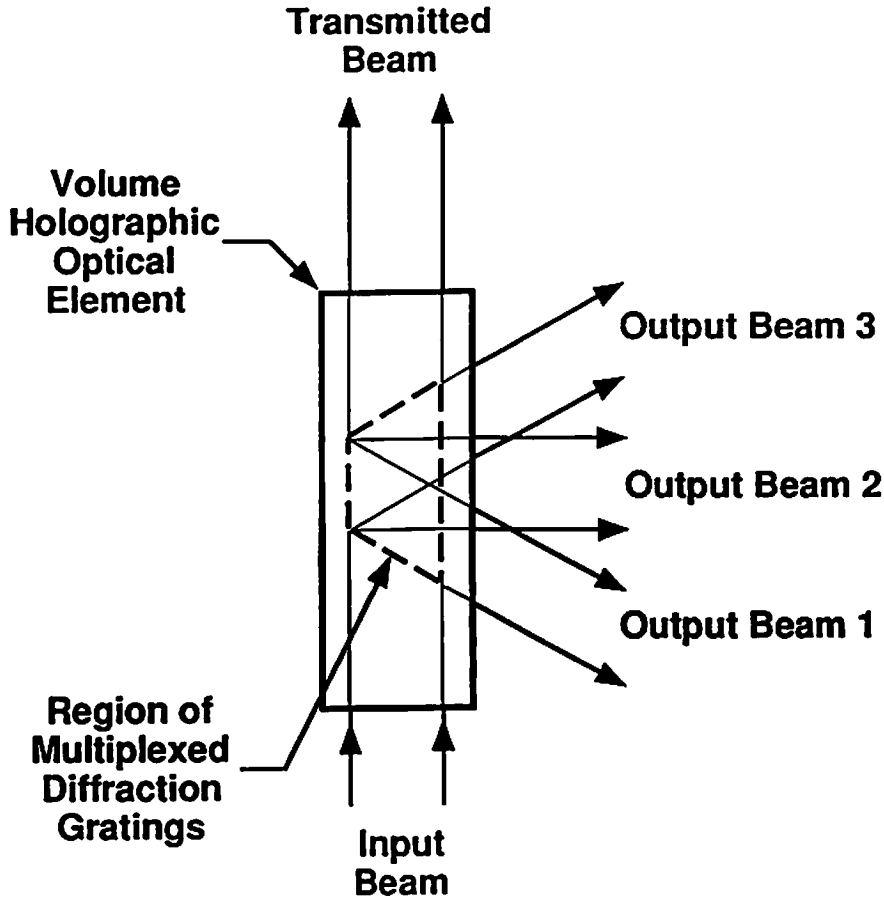



Figure 15.14 (b)

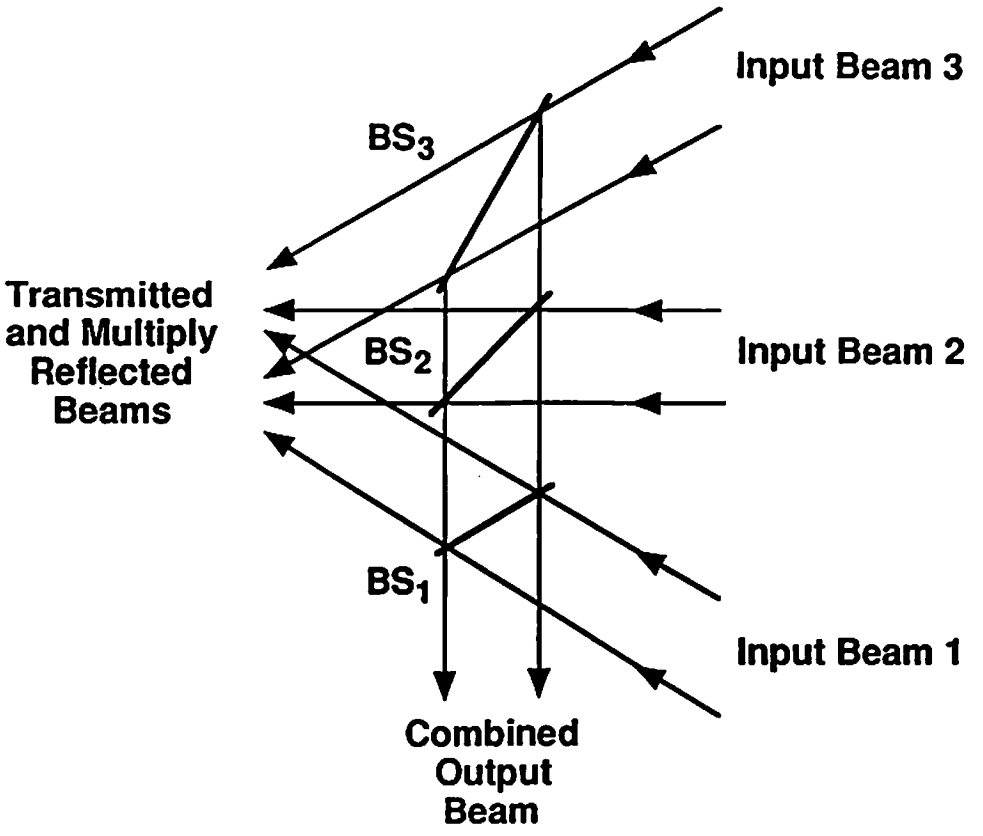



Figure 15.15 (a)

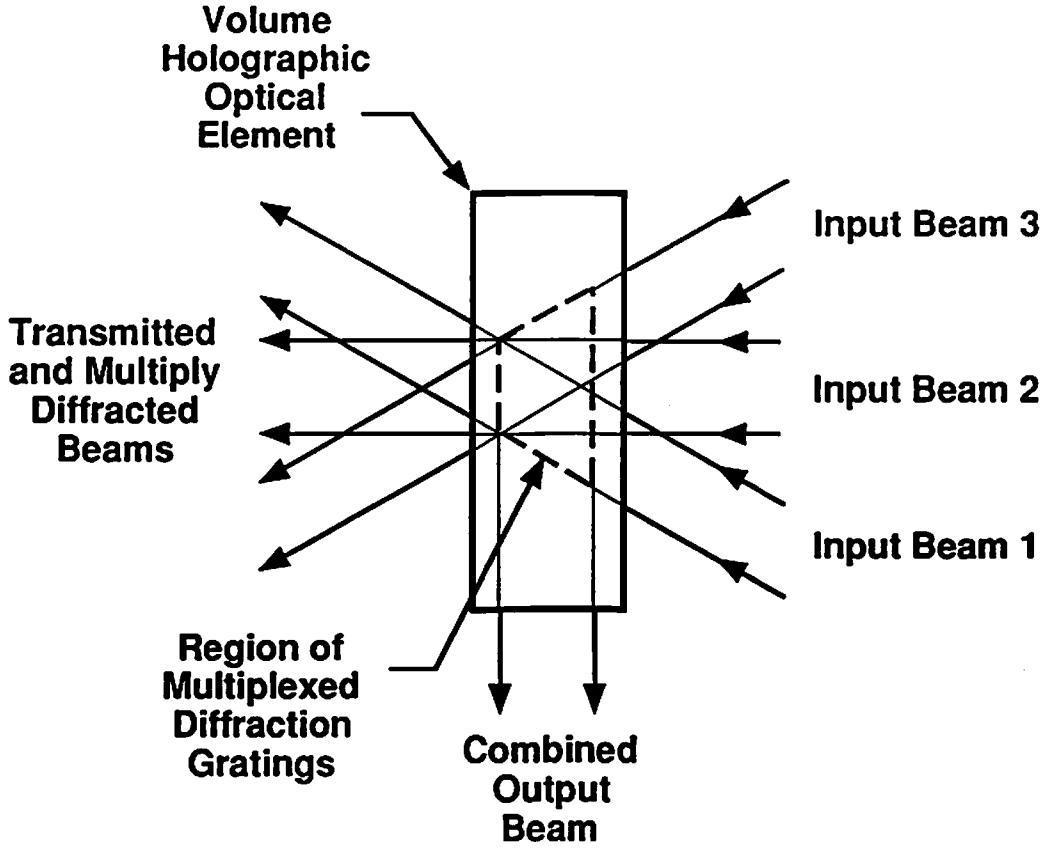



Figure 15.15 (b)

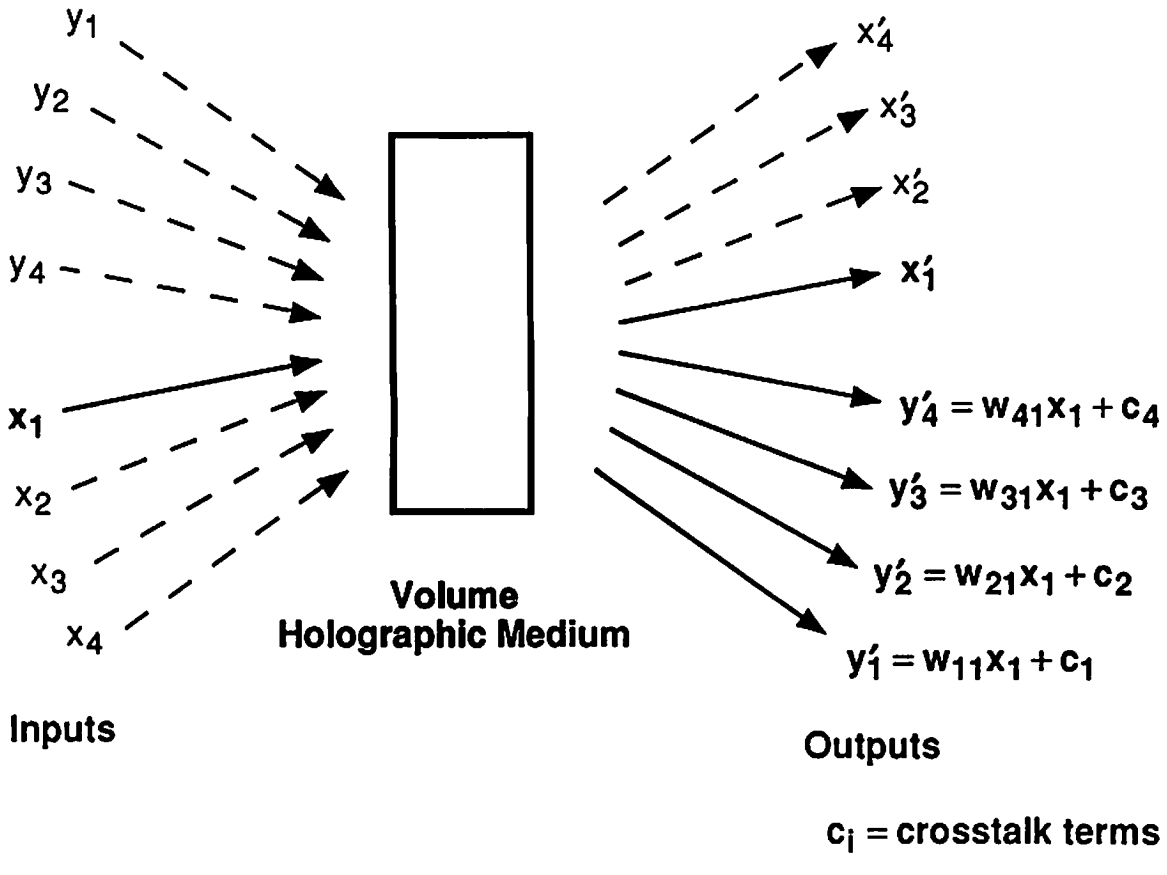

$y_1$
$y_2$
$y_3$
$y_4$
$x_1$
$x_2$
$x_3$
$x_4$
Inputs
Volume
Holographic Medium
$x'_4$
$x'_3$
$x'_2$
$x'_1$
$y'_4 = w_{41}x_1 + c_4$
$y'_3 = w_{31}x_1 + c_3$
$y'_2 = w_{21}x_1 + c_2$
$y'_1 = w_{11}x_1 + c_1$
Outputs
$c_i$ = crosstalk terms

Figure 15.16

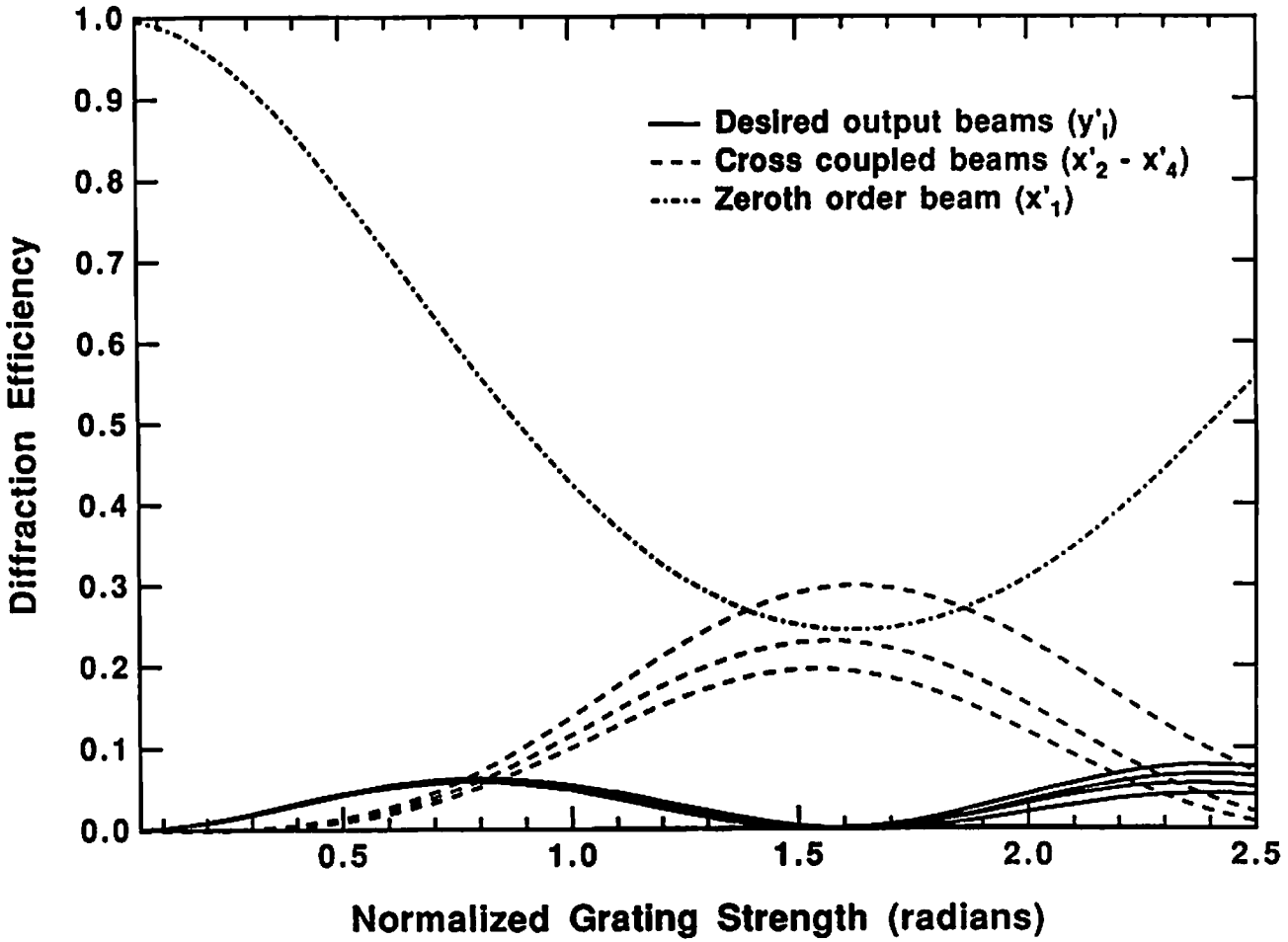

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Diffraction Efficiency
0.5
1.0
1.5
2.0
2.5
Normalized Grating Strength (radians)
Desired output beams (y'1)
Cross coupled beams (x'2 - x'4)
Zeroth order beam (x'1)

Figure 15.17

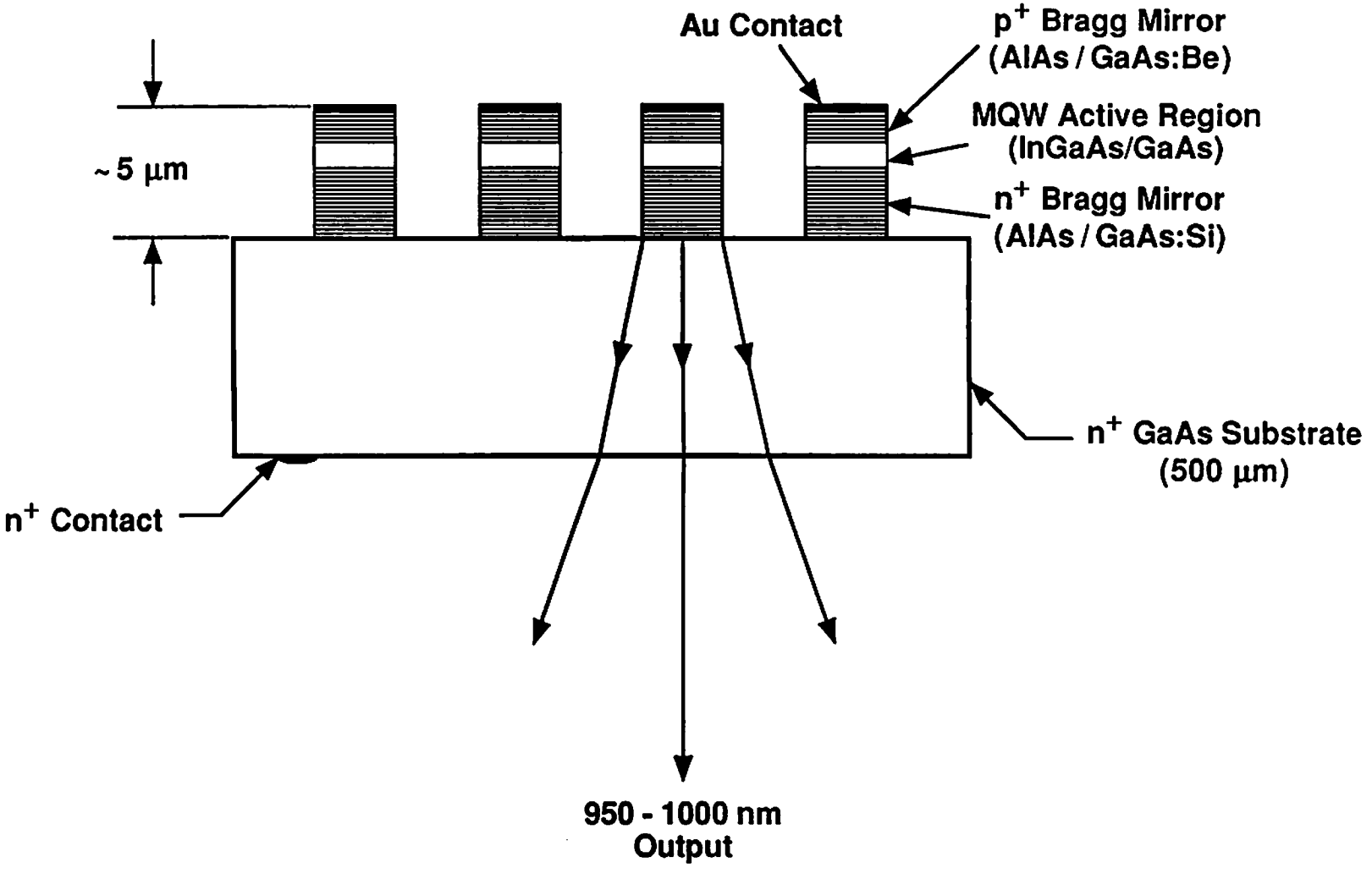

Au Contact
p+ Bragg Mirror
(AlAs / GaAs:Be)
MQW Active Region
(InGaAs/GaAs)
n+ Bragg Mirror
(AlAs / GaAs:Si)
~ 5 μm
n+ GaAs Substrate
(500 μm)
n+ Contact
950 - 1000 nm
Output

Figure 15.18

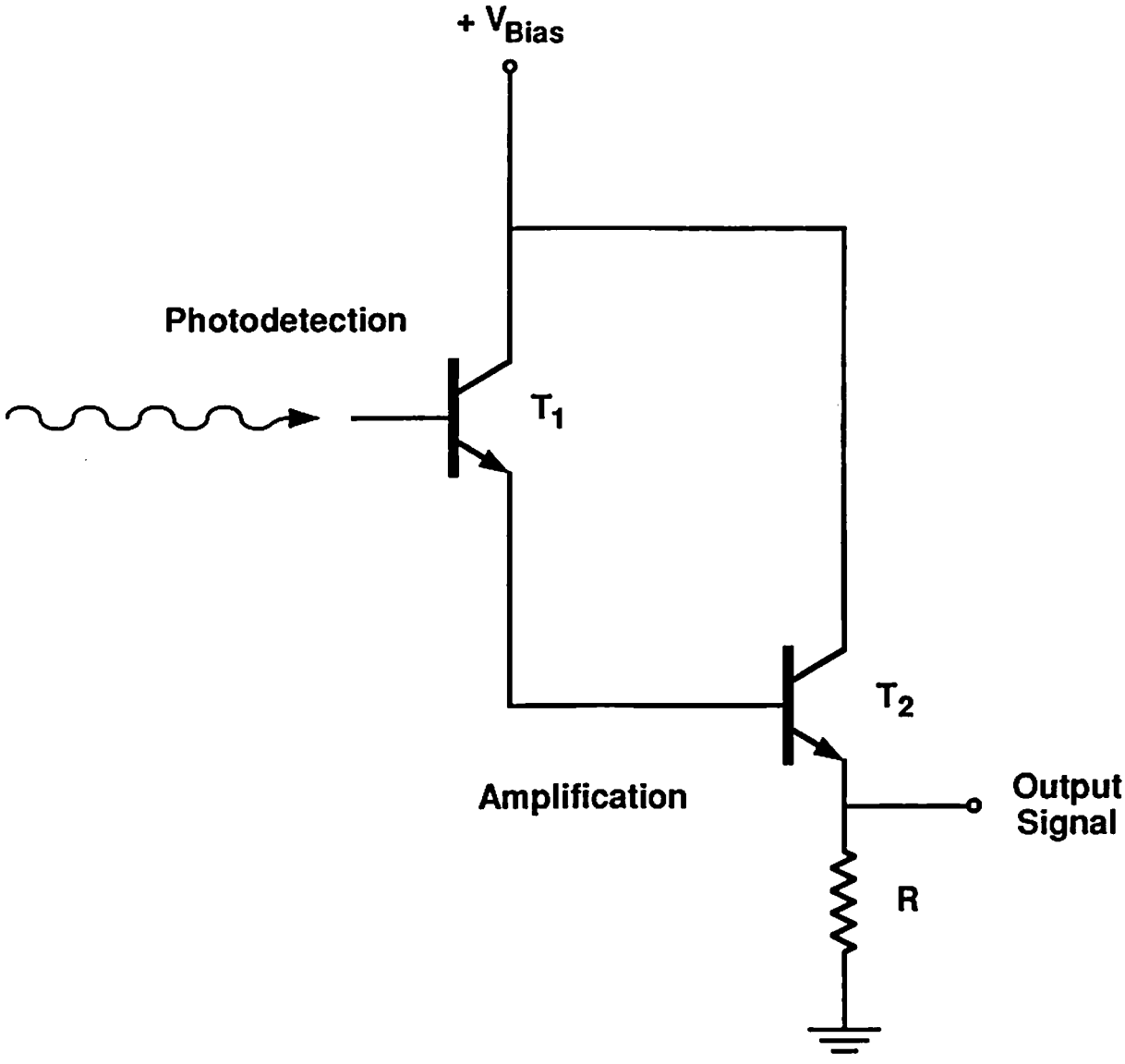



Figure 15.19

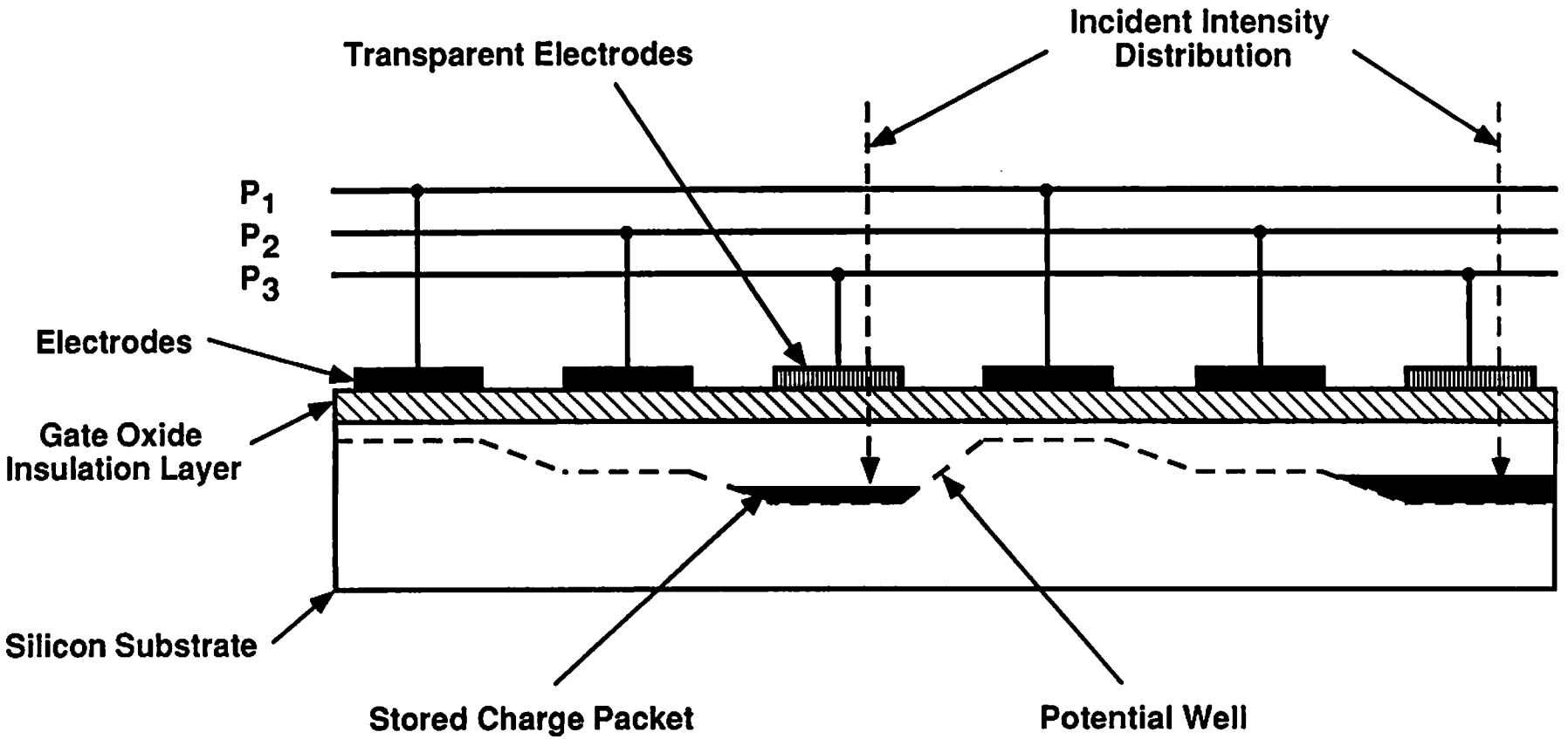

Transparent Electrodes
Incident Intensity Distribution
P1
P2
P3
Electrodes
Gate Oxide Insulation Layer
Silicon Substrate
Stored Charge Packet
Potential Well

Figure 15.20

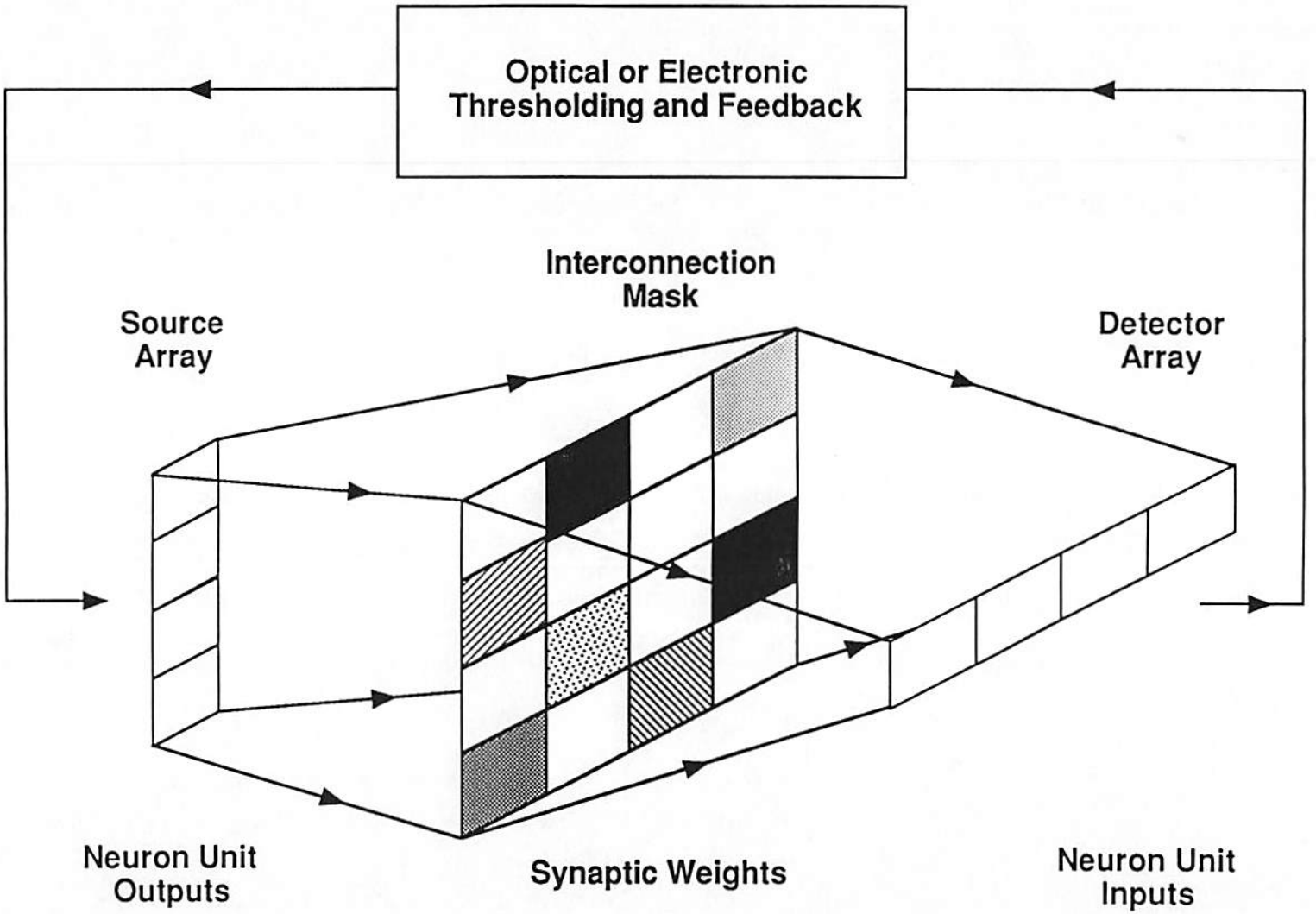

Optical or Electronic
Thresholding and Feedback
Interconnection
Mask
Source
Array
Detector
Array
Neuron Unit
Outputs
Synaptic Weights
Neuron Unit
Inputs

Figure 15.21

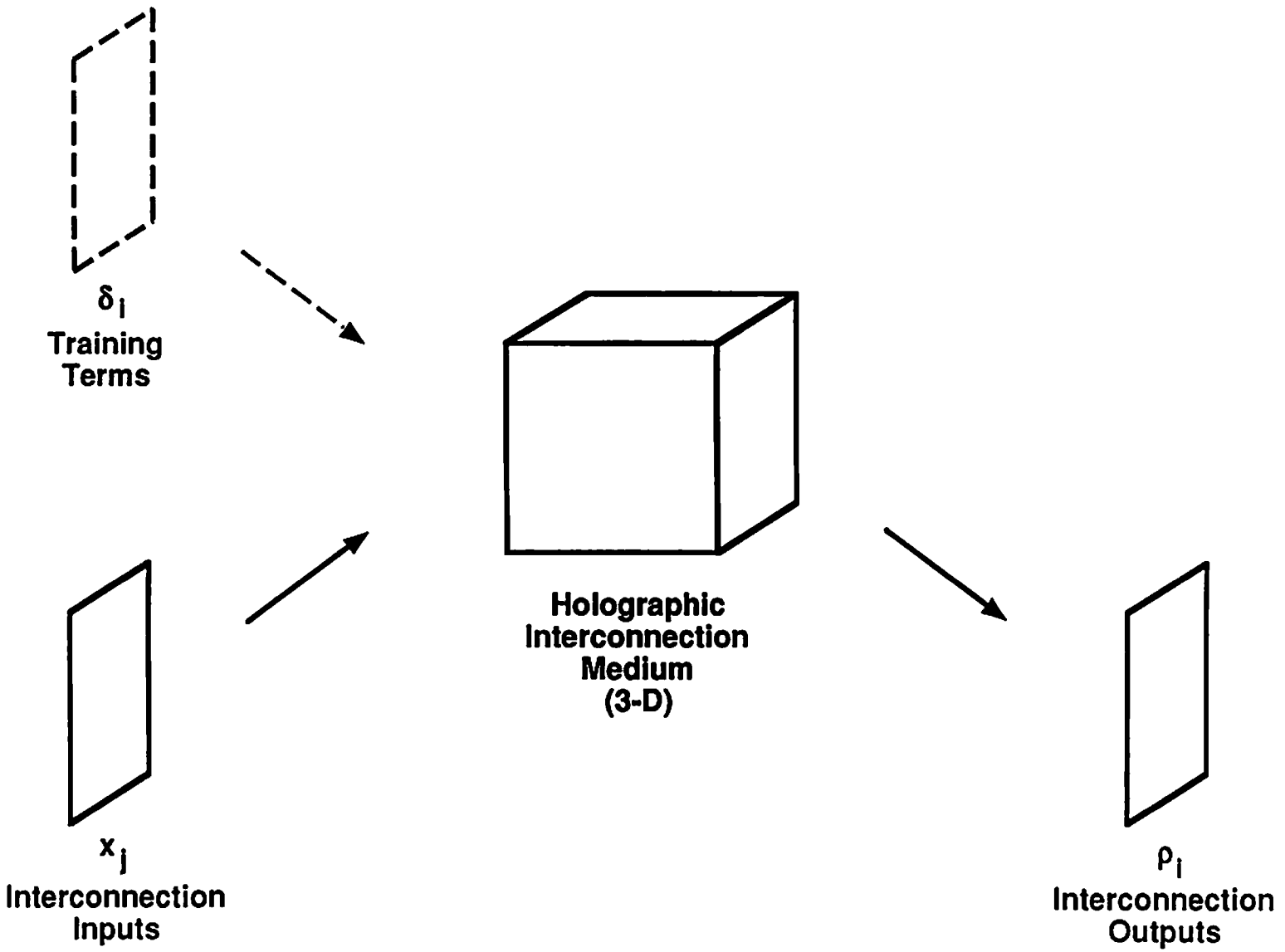
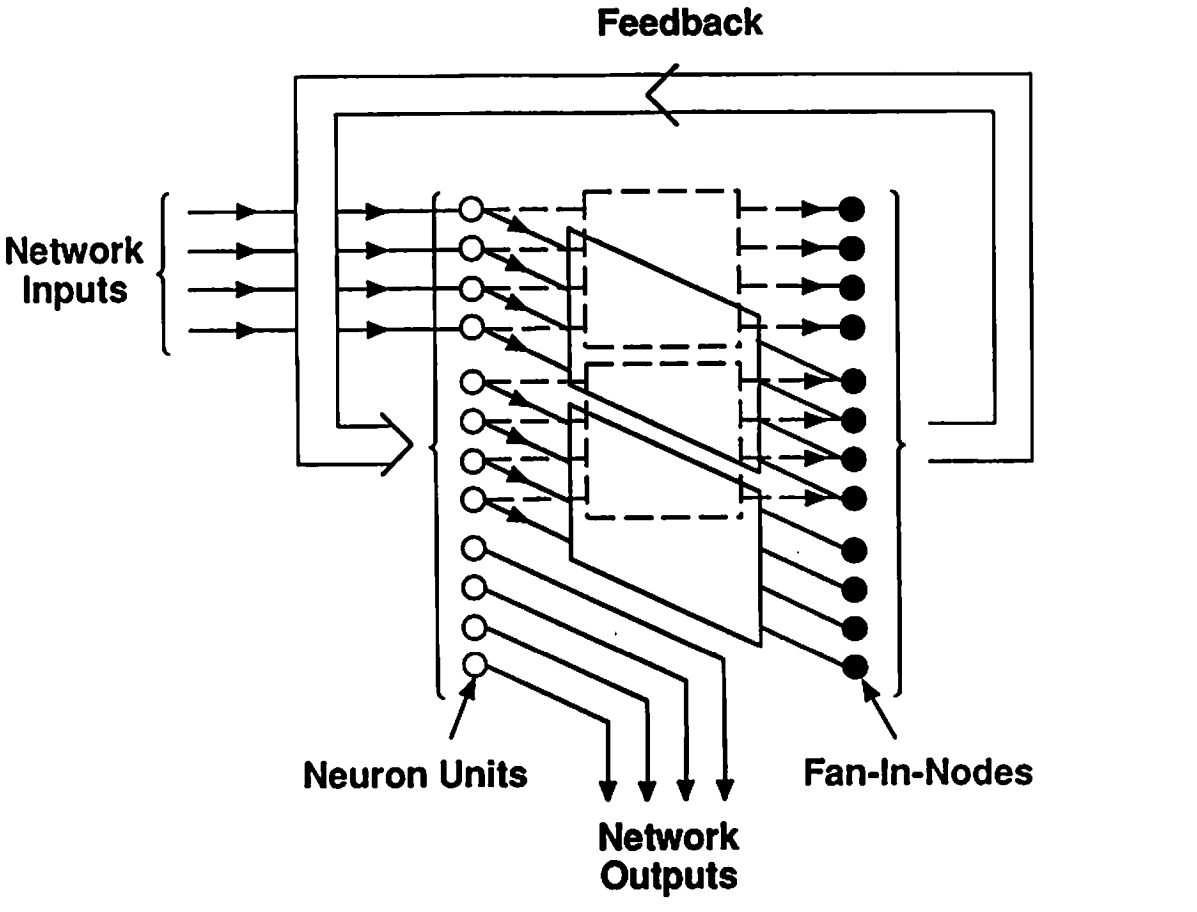
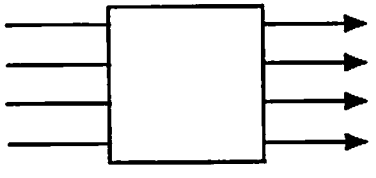
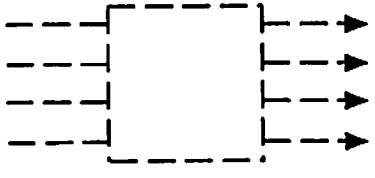

$\delta_i$
Training
Terms
$x_j$
Interconnection
Inputs
Holographic
Interconnection
Medium
(3-D)
$\rho_i$
Interconnection
Outputs

Figure 15.22



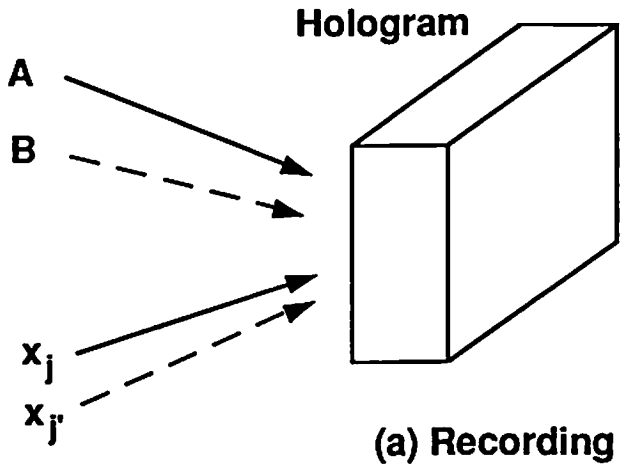



Arbitrary Feedforward Connections

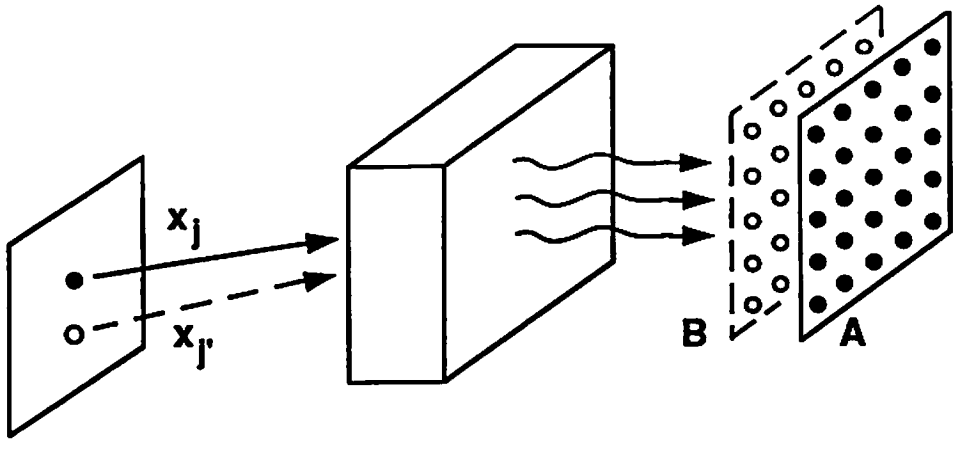


Arbitrary Lateral Connections

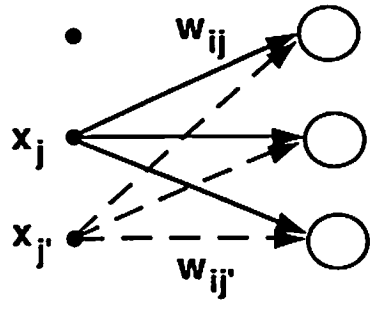
Figure 15.23




(a) Recording




(b) Reconstruction




(c) Fan-out/Fan-in Interconnections

Figure 15.24

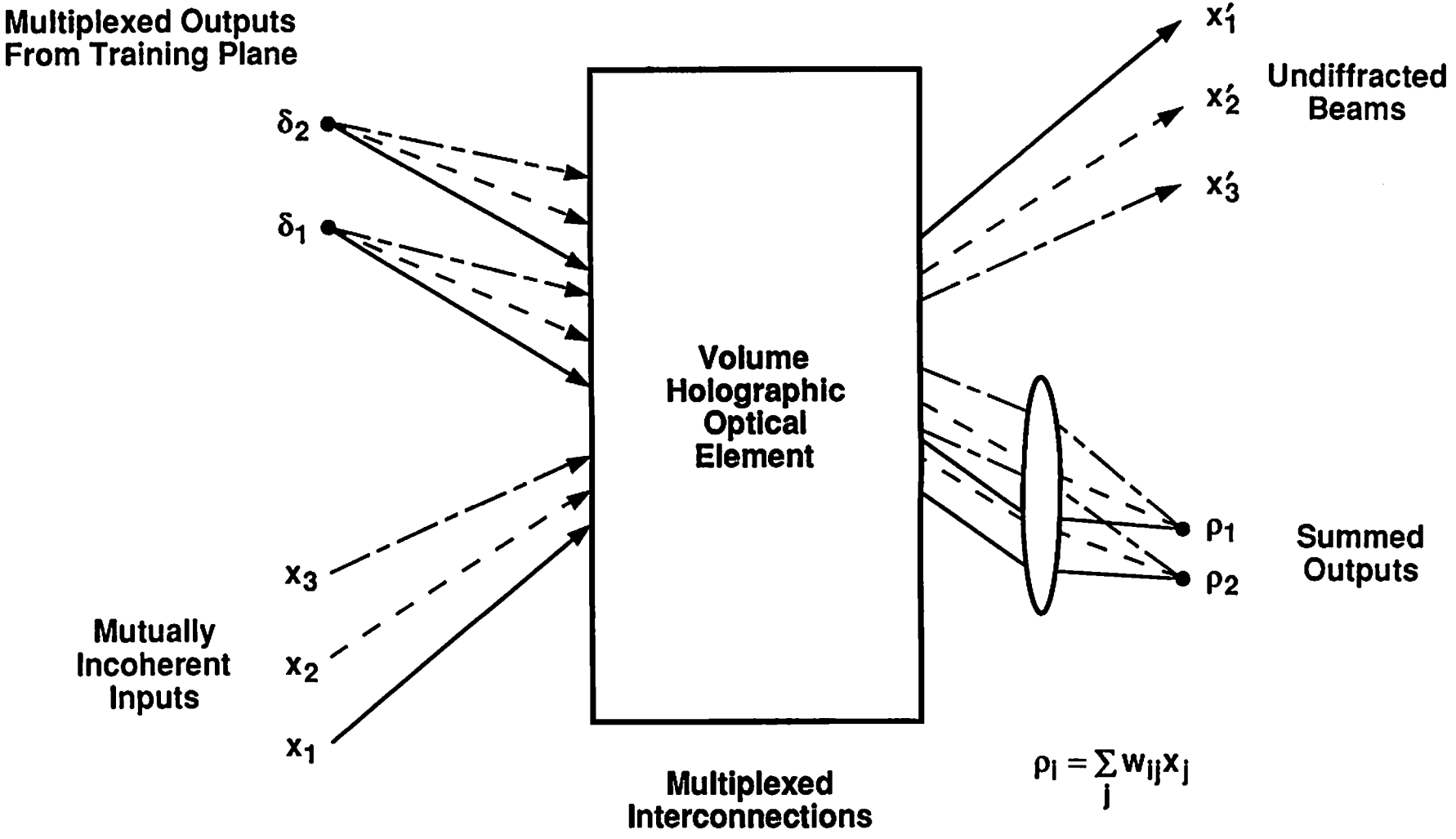

Multiplexed Outputs
From Training Plane
$\delta_2$
$\delta_1$
$x_3$
$x_2$
$x_1$
Mutually
Incoherent
Inputs
Volume
Holographic
Optical
Element
Multiplexed
Interconnections
$x'_1$
$x'_2$
$x'_3$
Undiffracted
Beams
$\rho_1$
$\rho_2$
Summed
Outputs
$\rho_i = \sum_j w_{ij} x_j$

Figure 15.25

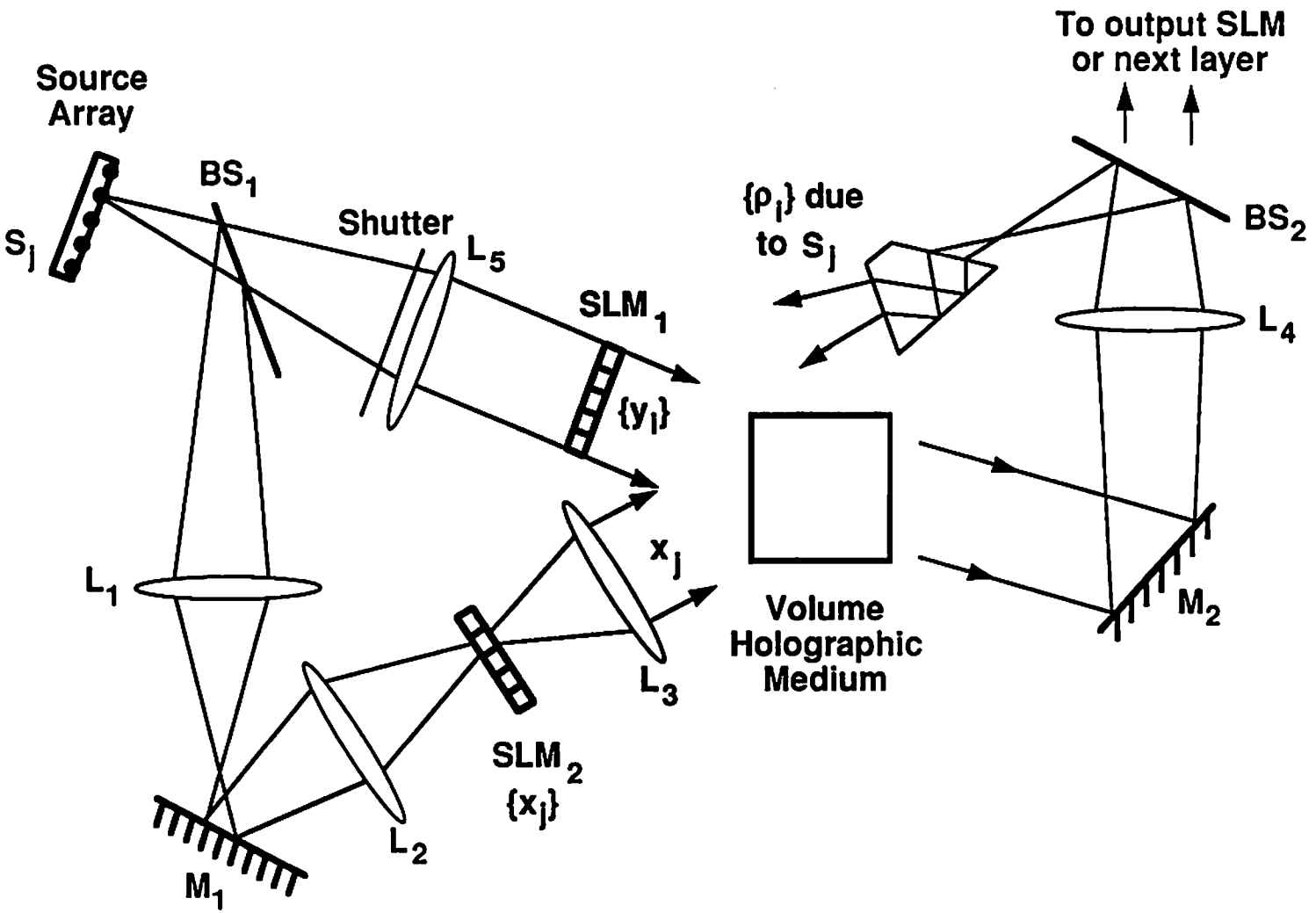

To output SLM
or next layer
Source
Array
BS1
Sj
Shutter
L5
SLM 1
{yj}
{ρj} due
to Sj
BS2
L4
xj
L1
Volume
Holographic
Medium
M2
L3
SLM 2
{xj}
L2
M1

Figure 15.26

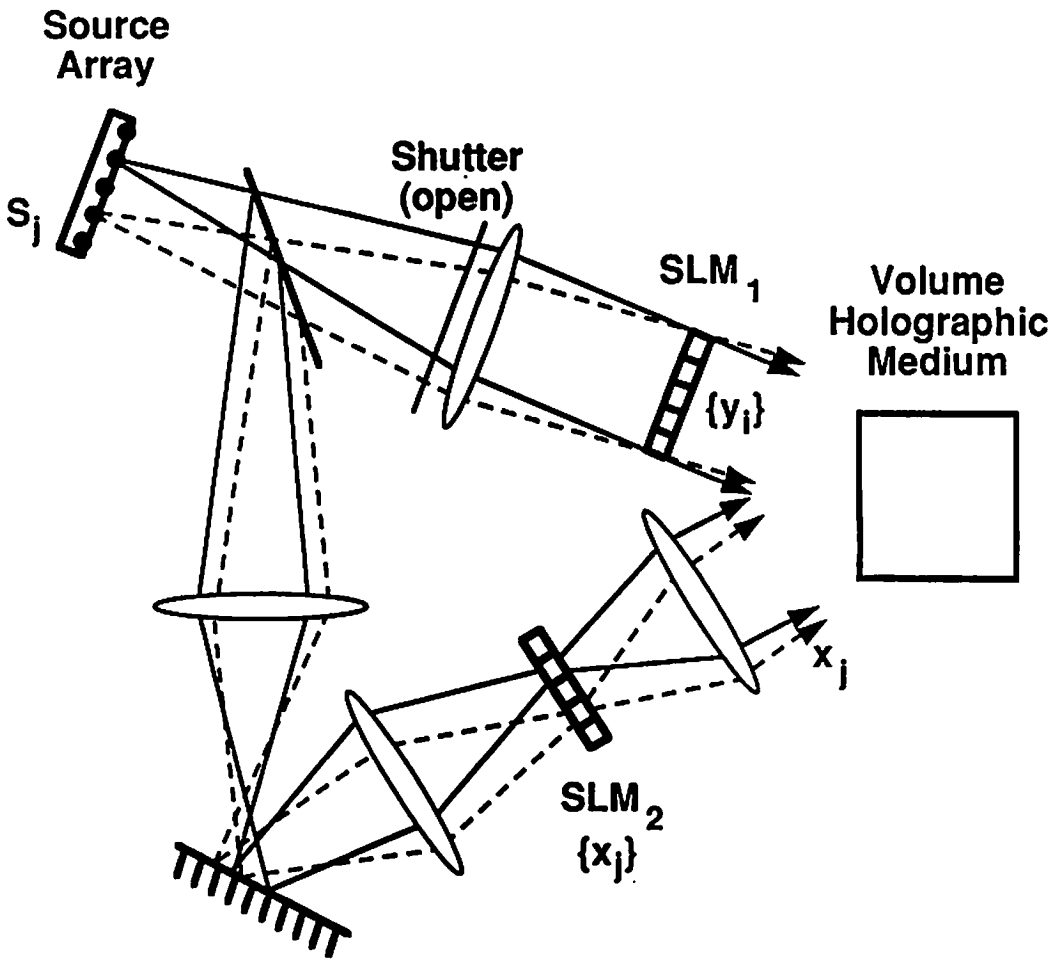

Source
Array
$S_j$
Shutter
(open)
$SLM_1$
$\{y_i\}$
Volume
Holographic
Medium
$x_j$
$SLM_2$
$\{x_j\}$

Figure 15.27

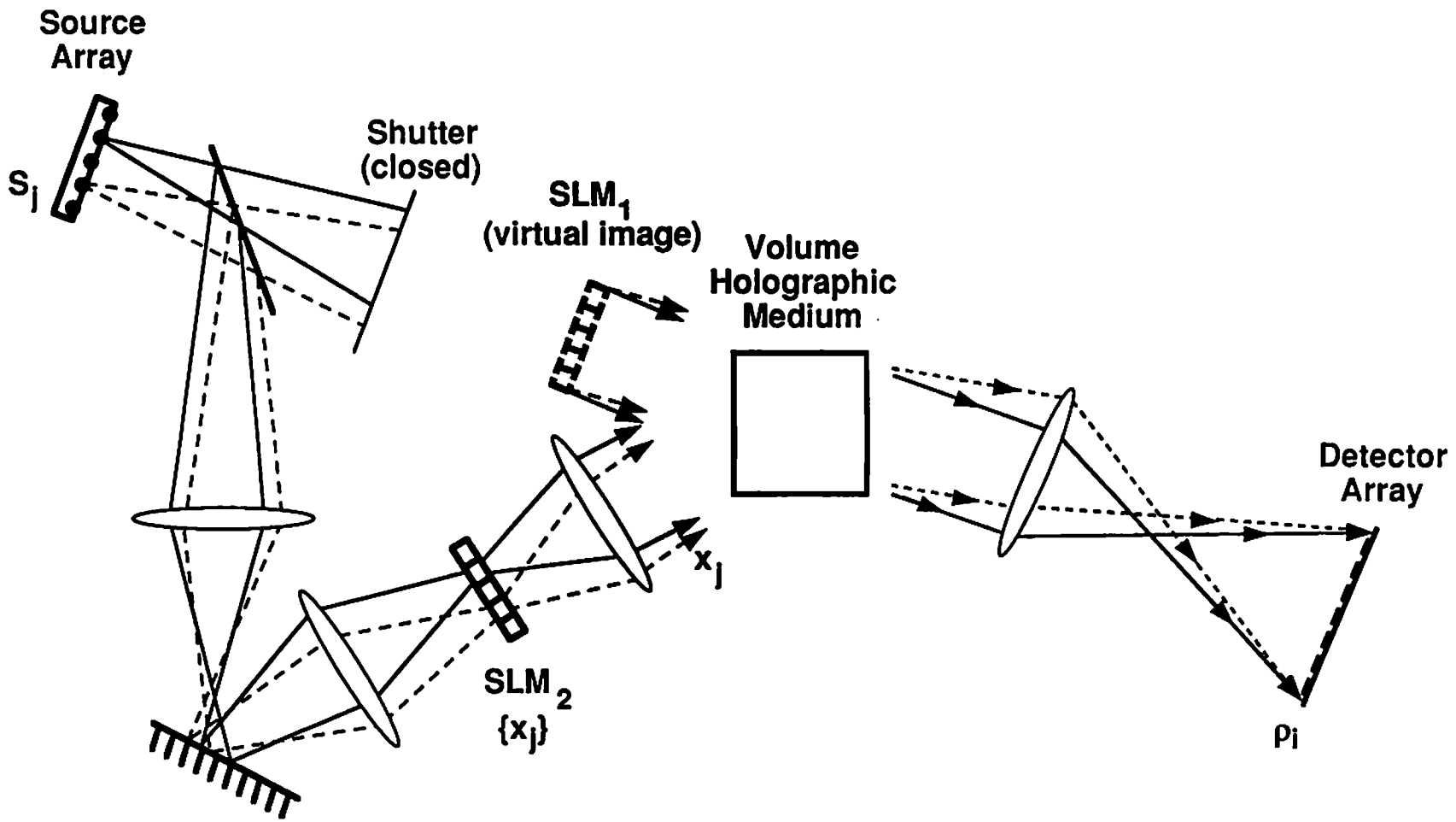

Source
Array
$S_j$
Shutter
(closed)
$SLM_1$
(virtual image)
Volume
Holographic
Medium
Detector
Array
$x_j$
$SLM_2$
$\{x_j\}$
$\rho_i$

Figure 15.28

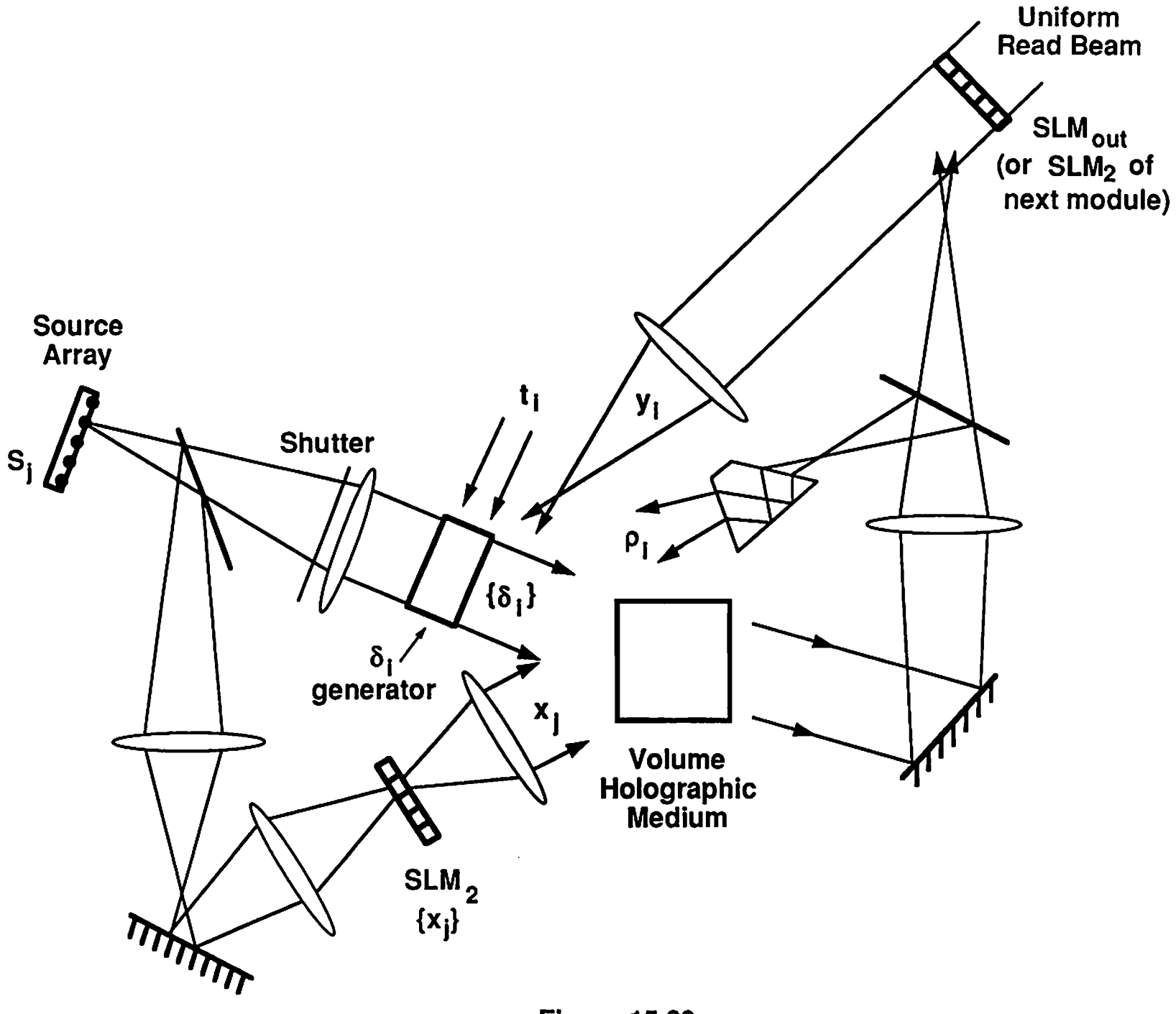

Uniform
Read Beam
SLM out
(or SLM2 of
next module)
Source
Array
Sj
Shutter
tl
yl
ρl
{δl}
δl
generator
xl
Volume
Holographic
Medium
SLM 2
{xj}

Figure 15.29

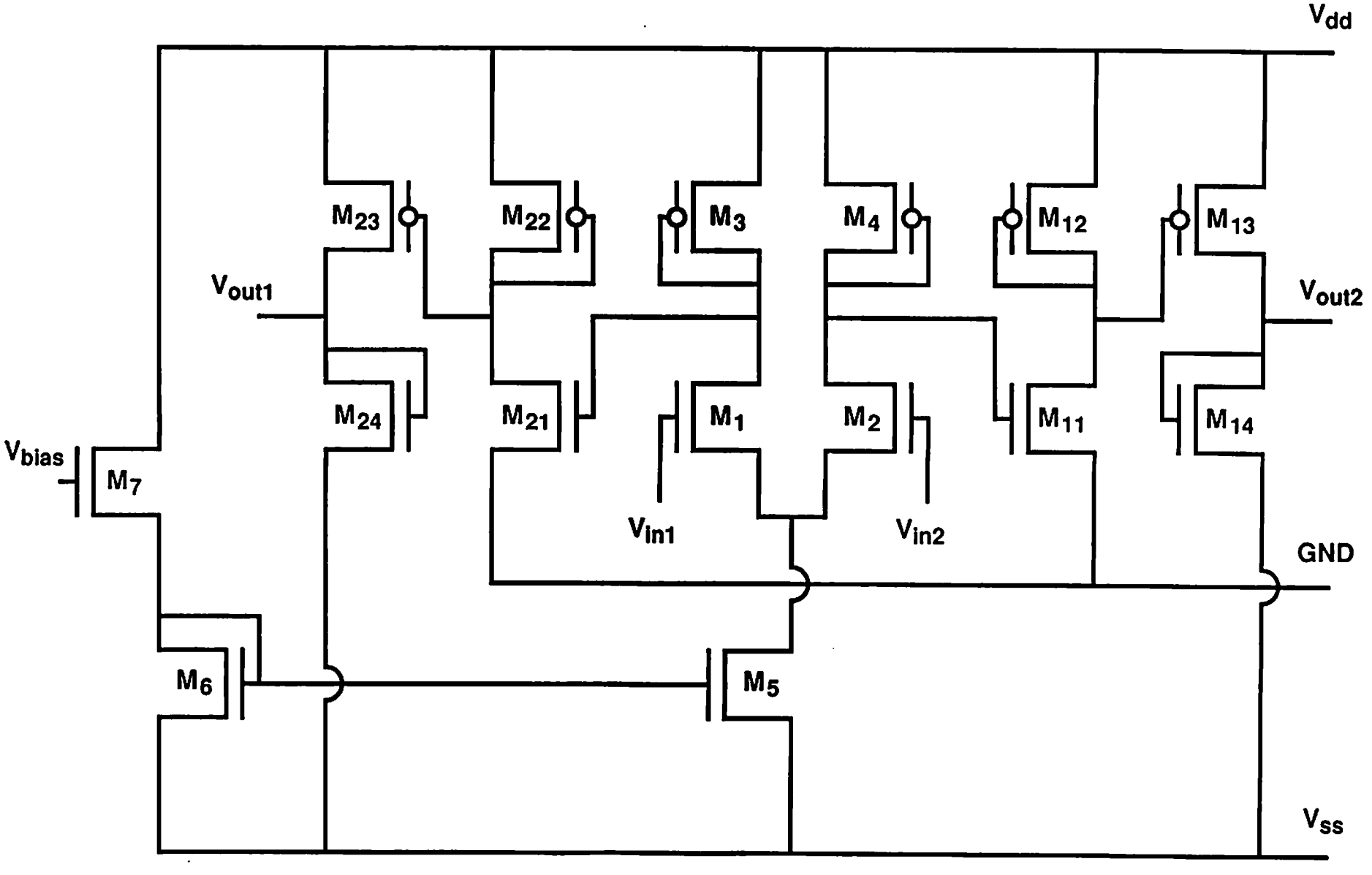

Vdd
M23
M22
M3
M4
M12
M13
Vout1
Vout2
M24
M21
M1
M2
M11
M14
Vbias
M7
Vin1
Vin2
GND
M6
M5
Vss

Figure 15.30

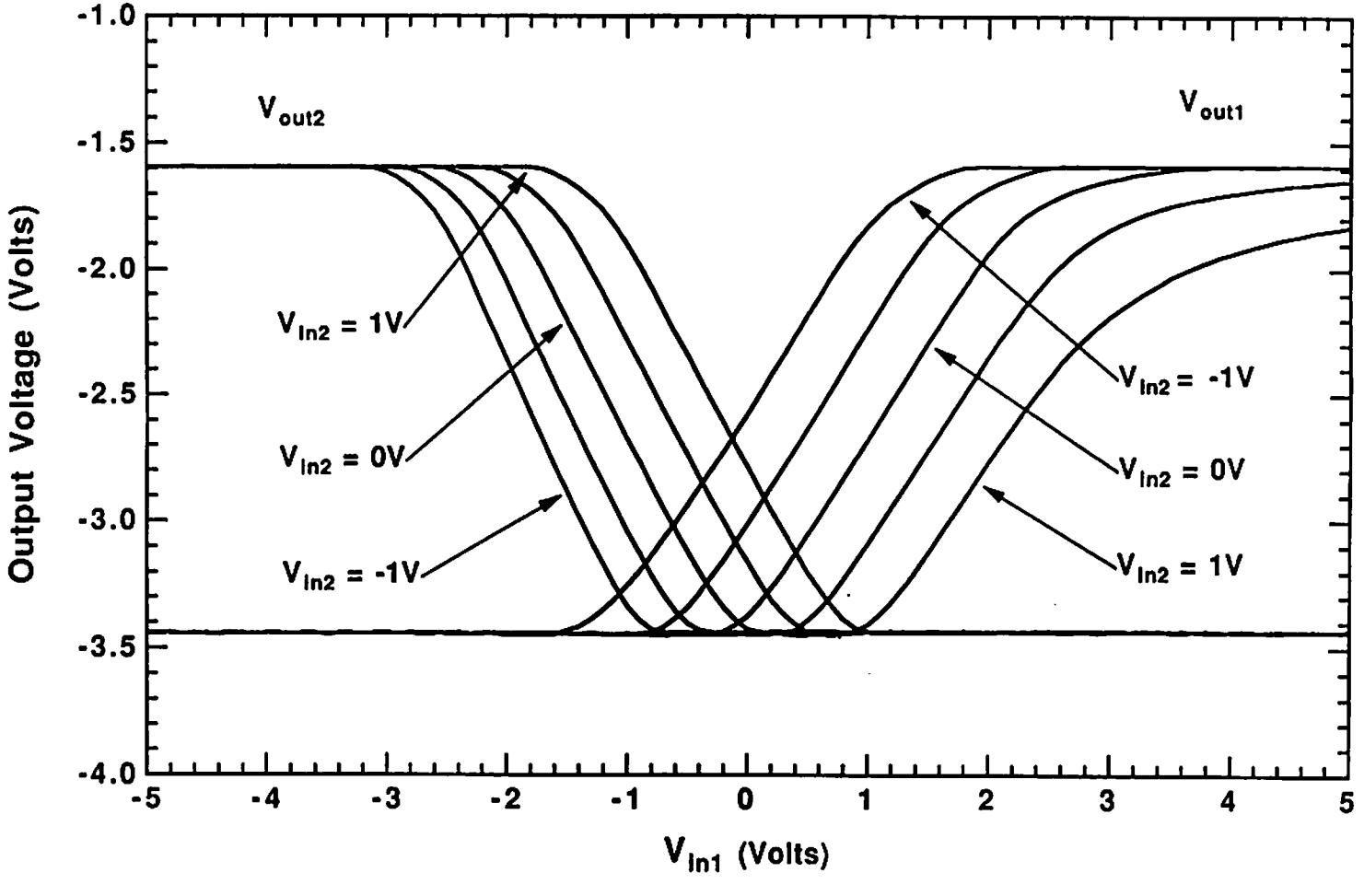

-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
-4.0
Output Voltage (Volts)
$V_{out2}$
$V_{out1}$
$V_{In2}$ = 1V
$V_{In2}$ = 0V
$V_{In2}$ = -1V
$V_{In2}$ = -1V
$V_{In2}$ = 0V
$V_{In2}$ = 1V
-5
-4
-3
-2
-1
0
1
2
3
4
5
$V_{In1}$ (Volts)

Figure 15.31

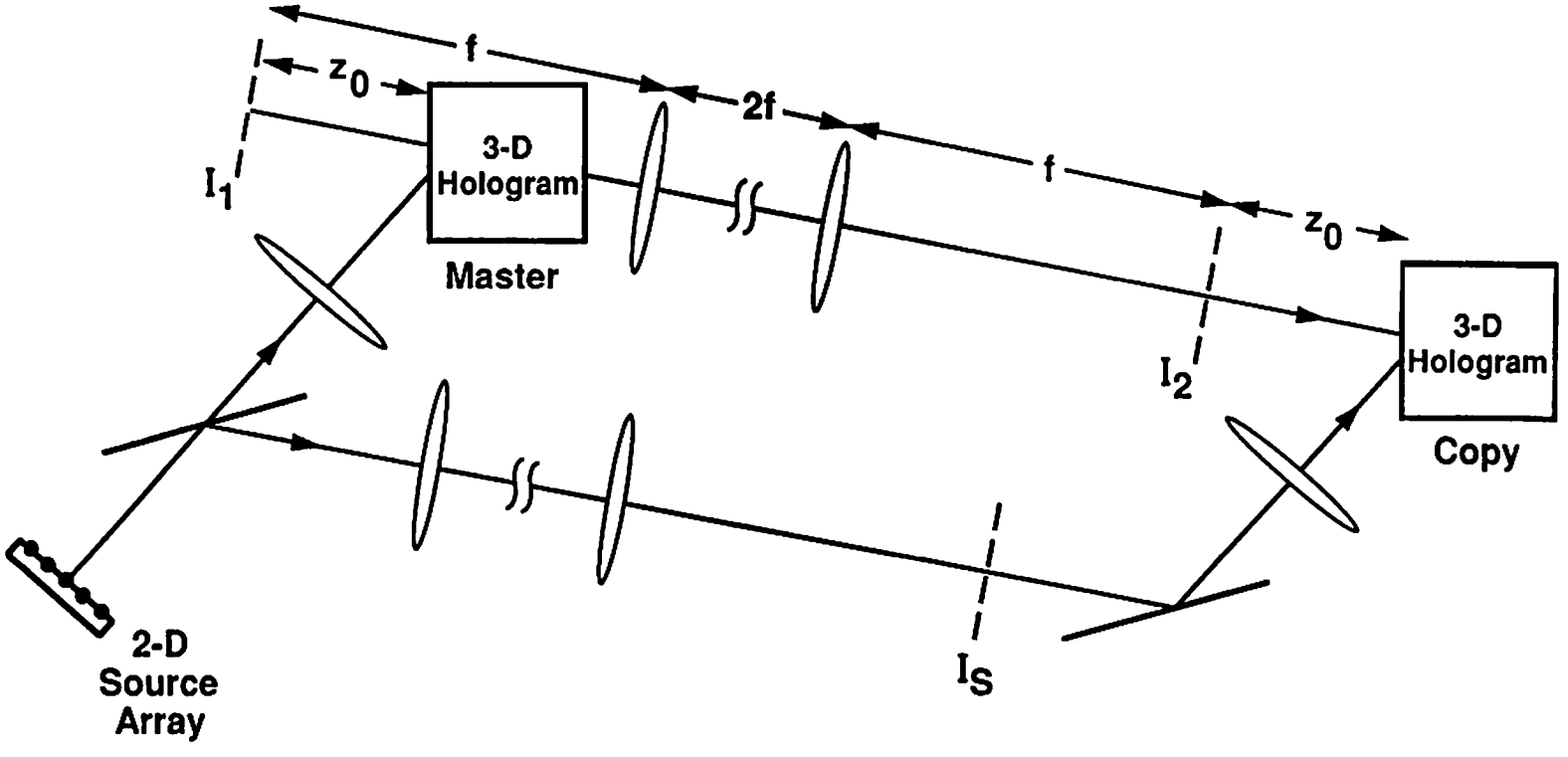

f
$z_0$
2f
f
$z_0$
3-D
Hologram
Master
$I_1$
$I_2$
3-D
Hologram
Copy
2-D
Source
Array
$I_S$

Figure 15.32

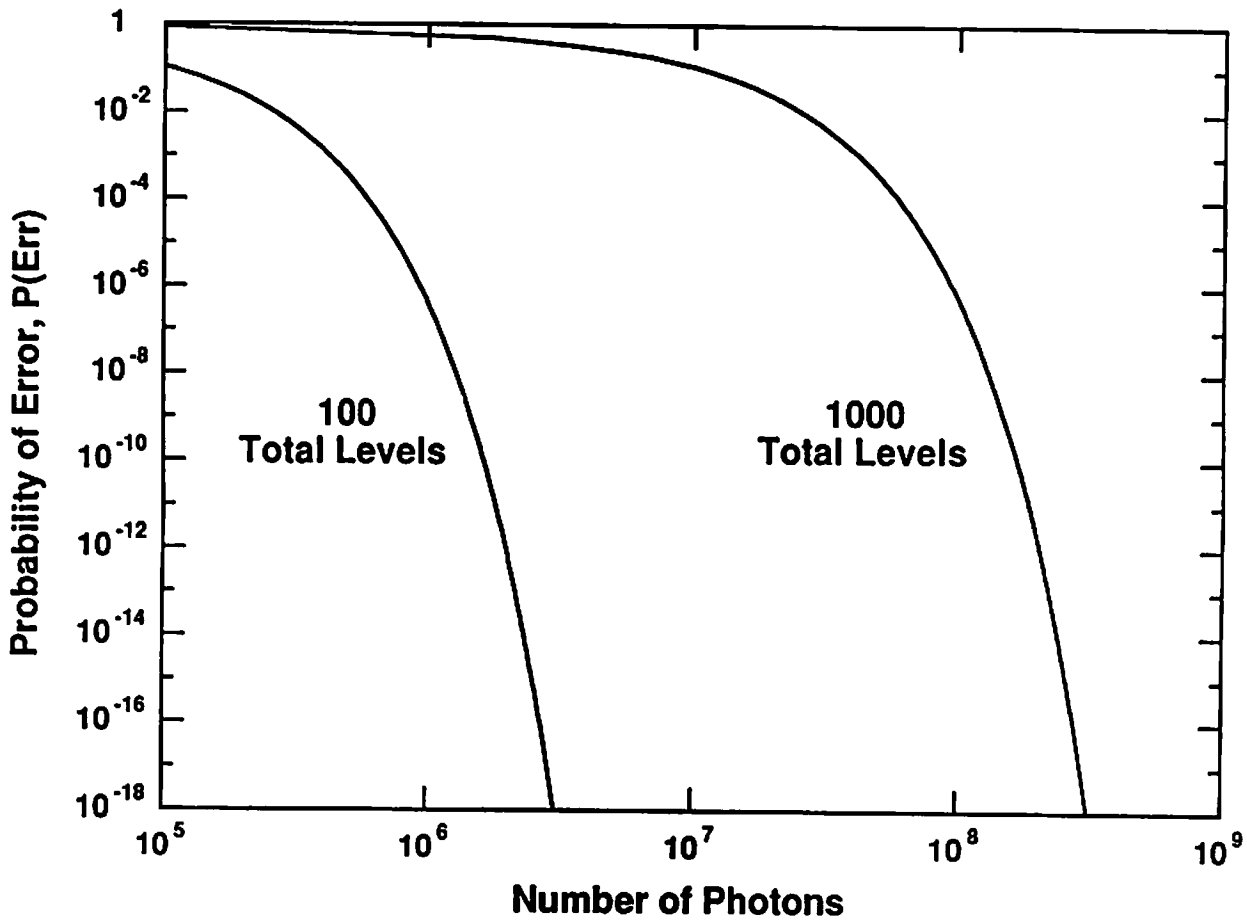

1
$10^{-2}$
$10^{-4}$
$10^{-6}$
$10^{-8}$
$10^{-10}$
$10^{-12}$
$10^{-14}$
$10^{-16}$
$10^{-18}$
Probability of Error, P(Err)
100
Total Levels
1000
Total Levels
$10^{5}$
$10^{6}$
$10^{7}$
$10^{8}$
$10^{9}$
Number of Photons

Figure 15.33

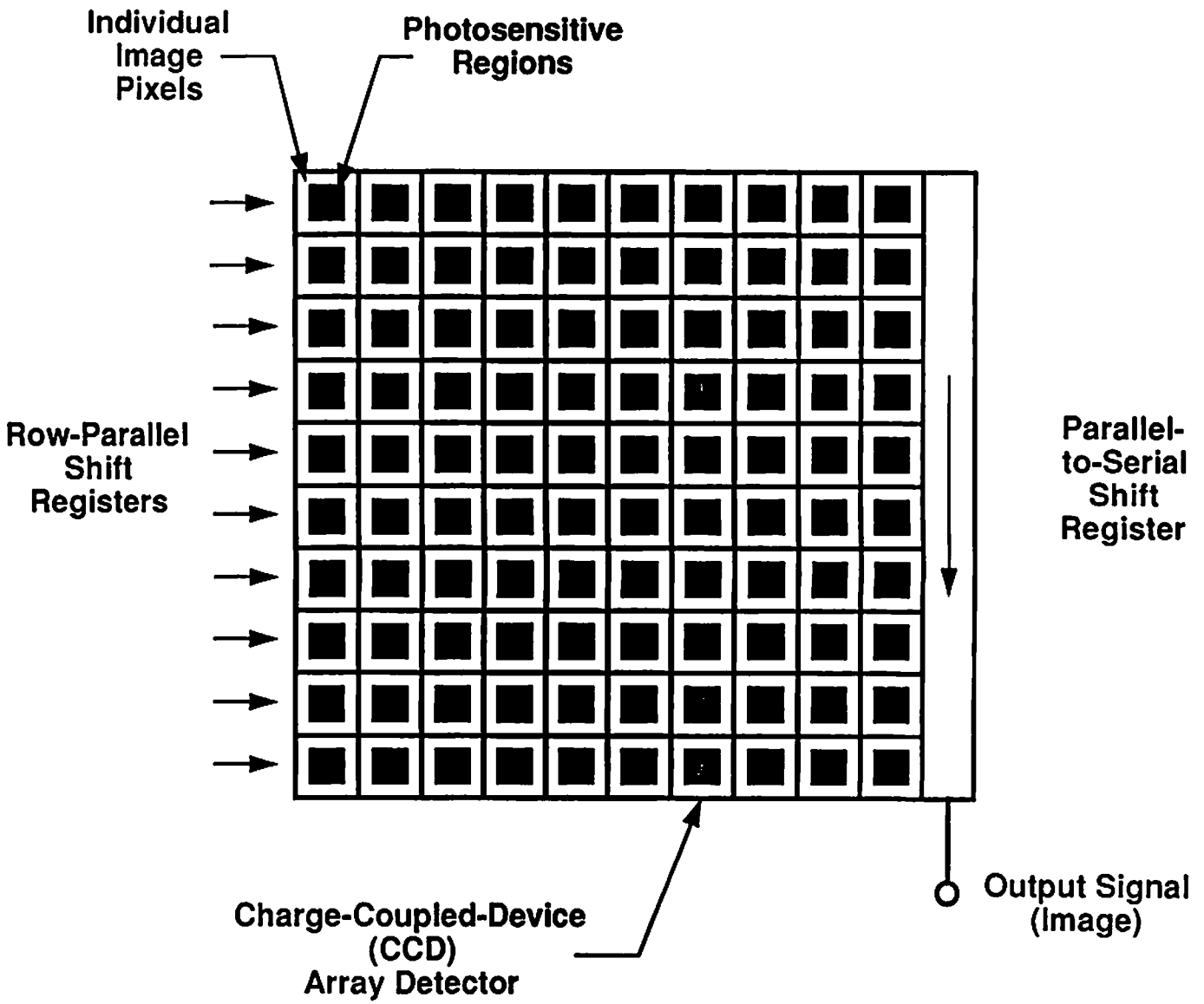

Individual
Image
Pixels
Photosensitive
Regions
Row-Parallel
Shift
Registers
Parallel-
to-Serial
Shift
Register
Charge-Coupled-Device
(CCD)
Array Detector
Output Signal
(Image)

Figure 15.34

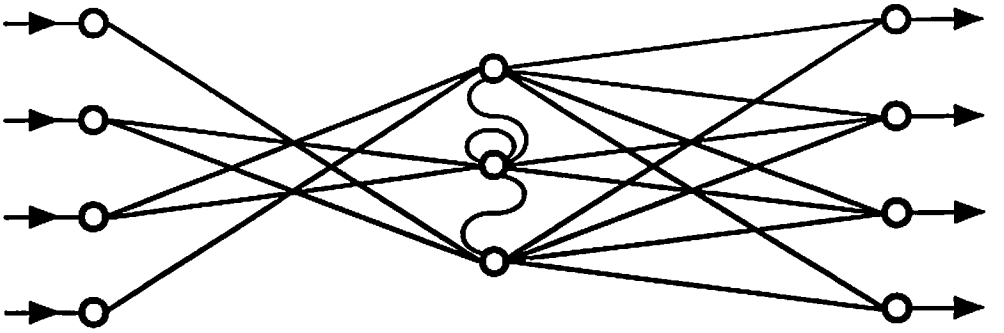


Figure for Problem 14(a)