USC–SIPI REPORT #107

Optical Computing and Interconnections

by

B. Keith Jenkins

October 1985

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 404
Los Angeles, CA 90089-2564 U.S.A.

Table of Contents

Optical Computing and Interconnections

B. K. Jenkins

ABSTRACT

This report describes research in digital optical computing and optical interconnections. It assesses the current state-of-the-art, discusses future prospects and worthwhile directions, and identifies areas that need further research in order to make digital optical computing feasible. Digital optical computing systems are compared to electronic systems, acoust-optics and integrated optics. Optical fixed and reconfigurable interconnections are considered for use in optical computers and electronic computers.

1. INTRODUCTION

This report presents an overview of digital optical computing and optical interconnections research. The architectural characteristics of digital optical computers, and optical interconnections in electronic computers, are discussed, examples are drawn from recent and on-going research, and classifications and possible applications are given. The different components needed for digital optical computers and optical interconnections are identified, and the present state of development of each is given. In addition, some specific aspects of digital optical computers are compared to alternative technologies, i.e. electronics, integrated optics, and acousto-optics, to the degree that is possible at the very early stage of research and development that digital optical computing is in. The remainder of this section defines digital optical computing and its objective.

While there has not been a consensus on one definition of optical computing among researchers in the field, a definition is given in [1]:

> **Optical computing:** "The use of optical systems to perform numerical computations on one-dimensional or multidimensional data that are generally not images."

Here we define digital optical computing:

> **Digital optical computing:** The use of optical systems to perform computations on digital (base 2) numbers. The numbers are represented primarily by photons.

With this definition we include computations on digitally represented images and do not require the data to be 1-D or multidimensional. (Although, as will be pointed out in this report, in most cases the data *will* be at least 1-D.) It should be pointed out that in this report the term "digital optical computing" is meant to exclude integrated optic implementations. Also, in the optical computing field the terms *computing* and *processing* are frequently used interchangeably. Technically, most optical systems to date are processors and not computers. One difference is that computers must have some form of

memory, whereas processors may not. In the optics area, computing, when distinguished from processing, may mean a more general-purpose machine, a machine that is programmable, or merely a machine that operates on numbers as opposed to images.

> **Objective of digital optical computing:** To perform computations that are too inefficient, expensive, time consuming, or otherwise impractical on an electronic computer.

We can see from this objective that digital optical computing will play a complementary role to digital electronic computing. In the future there will likely be all-electronic computers, computers that are part electronic and part optical, and possibly computers that are entirely optical.

In the remainder of this report, Sec. 2 discusses characteristics, classifications, examples, and possible applications of digital optical computers. Section 3 describes the current state-of-the-art of the different components needed for digital optical computing, namely gates or nonlinearity, memory, and input/output, and some of the promising research in these areas. Section 4 compares digital optical computing to alternative technologies, and Section 5 discusses the use of optical fixed and reconfigurable interconnection networks in electronic and optical computers and communication switching.

2. ARCHITECTURES AND APPLICATIONS

In this section we discuss digital optical computing (DOC) architectures, both logical (functional) and optical (physical). Optical computers have certain inherent characteristics that should be exploited; some ways of exploiting these features from an architectural point of view are described in Sec. 2.1. Methods of classifying digital optical computers are given in Sec. 2.2 and 2.3. Examples of DOC architectures are given in Sec. 2.4, and some possible applications for digital optical computers are also given in Sec. 2.5.

First we pose the question: will a digital optical computer be a special-purpose or general-purpose machine? The answer to this is less important than one might think. Most likely it will be a *specialized* machine, and could be general-purpose or special-purpose. *Specialized* means optimized for certain classes of problems. General-purpose has a mathematical definition (can implement a Turing machine). In simple terms, a computer is general-purpose if and only if it can, given infinite memory and a very large amount of time, compute any computable function or execute any algorithm (to completion). The point here is that we always have a finite amount of memory and typically have a very limited amount of time. Useful optical computers (special purpose or general purpose) will be ones that can solve certain classes of problems substantially faster than electronic machines can. General purpose is preferable, because then one is assured that there will never be a step in an algorithm or program that the machine is not capable of performing. (Having to execute an unexpected instruction would just slow the machine down.)

2.1. Design Constraints of Optical and Electronic Systems

The design of processors and computers in any technology is constrained by the inherent characteristics of the technology. In this section we compare some of these constraints for electronic and optical systems and discuss how they affect system architectures.

Since the development of electronic LSI and VLSI, the cost of individual gates has constituted only a minor factor in the overall system cost function. The major concern has become internal and external communications [2,3]. The internal wiring network affects the amount of active chip area available for gates; in current systems it is common for more than 70% of the chip area to be devoted to interconnections [4]. Because of the resistance and parasitic capacitance associated with each on-chip wire, the response time of a gate and the propagation delay of a wire both become functions of the length of the wire; power consumption also becomes a function of wire length.

Timing and clock skew become problems because of the differing wire lengths [5,6]. (Although the wire resistance can be reduced by process technology, e.g., using thick metal layers, it appears that the wire lengths will still be a limiting factor in the system timing considerations [7].) Another restriction related to interconnections is the limited number of pin-outs which becomes more apparent as the number of gates per chip increases. One result of these restrictions is the need to minimize the number and length of interconnections.

For optical logic systems, the major design considerations are admittedly not so well defined as for VLSI, but it is clear that the cost function is much different. In particular, most of the communication costs affecting VLSI design are not associated with optical systems. An optical system can be made so that all interconnections have the same length to first order. (For example, this is the case in most of the DOC systems described in Sec. 2.4.) Thus synchronization problems due to clock skew can be eliminated, making large synchronized systems more feasible. Being able to synchronize the circuits eliminates the need for handshaking or other asynchronous techniques which introduce waiting time for individual circuit elements. The other design constraints associated with wire length, namely power consumption and device area utilization, can be avoided with optical systems, for example by using free space propagation in the third dimension for the interconnections.

Pin-outs are not a constraint in optical systems. Optical systems can accept a large number of parallel inputs and can generate a large number of parallel outputs. These are usually in the form of 2-D arrays of data or bits, e.g. bit planes of images. The careful partitioning of large systems is then unnecessary, and limitations on concurrent and pipelined processing due to large I/O requirements is relieved.

Of course, digital optical systems will have constraints of there own. It appears that these will be primarily concerned with the gates. The design constraints dictated by the gates may be in the total number used or in the average repetition rate at which they are switched; making a (binary) decision may be more costly with optics than with electronics. In general there may be a preference for regular or repeated interconnections. In addition, other limitations may of course surface as the technology progresses. As a rule of thumb, in electronics interconnections are expensive and gates (decisions) are inexpensive, and in DOC interconnections are inexpensive but gates may be

expensive. Thus we see the complimentary natures of the two technologies.

Because of these differences in the cost functions of electronic and optical systems, certain application areas are specifically well-suited to one or the other. VLSI systems are being particularly considered for applications which involve very regular structures and simple data flow that can be handled with only local communications. An example is systolic array architectures [6,8], which are well-suited to many vector-matrix and matrix-matrix operations. On the other hand, algorithms which inherently require global communications cannot be conveniently handled by VLSI, but could, in principle, be implemented with an optical system. Examples of such communication-limited operations include some fast Fourier transform (FFT) algorithms which required global communications due to their butterfly structure [9], and some image processing operations such as histogramming and regional property computation.

2.2. Classification of Digital Optical Computers

Many parameters can be used to classify DOCs. These include the architecture (physical or functional); gates, interconnections, and memory; devices and and materials. The term *physical architecture* refers to the physical components used and how they are organized and connected. A diagram of a physical architecture typically involves spatial light modulators, lenses, holograms, etc. *Functional (or logical) architecture* refers to a functional level description, which by itself is essentially technology-independent. A diagram of a functional architecture typically involves boxes labeled processing element (PE), memory, interconnection network, etc, or may involve lower level descriptions such as showing the gates and how they are interconnected.

Classification by functional architecture is probably the most general. Since functional architectures are essentially technology independent, classifications developed by the digital electronic computing community apply. Some of these are discussed briefly in the next section. Classification by physical architecture has not been pursued yet, primarily because there have been so few digital optical architectures demonstrated or even discussed to date. Without the restriction of "digital", there are clearly some methods of classifying: (1) analog, digital, number theoretic; (2) integrated optics, acousto-optics, systems using 2-D spatial light modulators; (3) active, passive; (4) linear space-invariant, linear space-variant, nonlinear space-invariant, nonlinear space-variant. In Sec. 2.4 examples of physical architectures of DOCs will be given.

Gates will be discussed in Sec. 3.1.1 and can be classified by:
(1) *signal level encoding technique.* The 0 and 1 logic levels can be encoded as intensity levels, intensity patterns, spot position, grating orientations, polarization orientations, etc. All of these have been demonstrated.
(2) *materials used.* In the case of active gates this may involve photoconductor materials, nonlinear optical materials such as GaAs, $BaTiO_3$, or electro-optic materials such as BSO ($Bi_{12}SiO_{20}$) or liquid crystal.
(3) *device(s) used,* such as photorefractive crystal; spatial light modulator such as the liquid crystal light valve, microchannel spatial light modulator, or Pockels readout optical modulator [10]; bistable device array [11]; or passive devices such as spatial filters or holograms.

Interconnections will be discussed in Sec. 5; briefly, they can be classified according to: (1) whether they are passive or active, (2) fixed or reconfigurable; (3) whether they use free space or guided wave optics; and (4) what level they are used at in a computer (gate, board, processor, etc.).

Memory can also be classified, according to its structure and organization - it can be location addressed, content addressed, associative, or "time addressed" (e.g. delay line). A location addressed memory is a conventional memory as used in von Neumann computers. It is an address (or location) in, data out memory. A content addressed memory is just the opposite - data in, address out. Finally, an associative memory is data in, data out. It effectively associates a given value of the input data with output data. An obvious application for associative memory is in databases. It should be pointed out that often in the electronics literature content-addressable and associative memories are treated synonymously, and other definitions of the distinction also exist. Other important memory issues include retention time, read/write or read only, and parallelism of access.

Two characteristics must be provided in any digital computing system: a nonlinearity and gain. These are provided by materials and devices. In the optics case, an array of threshold elements, preferably inverting, is usually used, whether or not gates are being used as part of the system (some architectures may not even have gates). Any computer needs to have a non-interaction of signals (e.g. for interconnections) and a mechanism for signals to interact (e.g. for gates). In the optics case the signals can be linearly combined as part of the (linear) interconnections, and then a nonlinear element completes the interaction between the signals.

2.3. Classification of Functional Architectures

Here we will include classification schemes invented by the digital electronics community, but will discuss only what is pertinent to the optics case. A good reference for most of this is [12]. Other references are given where needed.

Flynn's taxonomy provides a simple method of classifying parallel computers. It considers the number of instruction streams (either single or multiple) and the number of data streams (either single or multiple) in the operation of a computer. A conventional computer has but a *s*ingle *i*nstruction stream and a *s*ingle *d*ata stream, or SISD, since it performs operations sequentially on one data element at a time. SIMD computers (single instruction but multiple data streams) perform the same operation on many data elements synchronously. There is effectively one master giving the same order to many slaves simultaneously. An example is an array processor. Finally, there are MIMD (multiple instruction, multiple data) machines, in which there is more than one processing element (PE) performing computations in parallel, but different processors might be doing different things. Communication between PEs is done by passing messages through an interconnection network to other PEs, or to and from memory that is shared with other PEs. SISD computers are not useful for optical implementation because they do not take advantage of the inherent parallelism of optical systems. Both SIMD and MIMD computers can.

There are (functional) architectures which do not fit into Flynn's taxonomy, that are interesting because they may also be well-suited to the capabilities of optical systems. Some examples are given here. One example is a pipelined machine which functions in the same manner as an assembly line, with each station performing a different operation as the data passes by. Another example is a data flow machine. Data flow machines do not follow a set program sequence. Instead, each operation executes when its operands are available. For example, for $p = x + y$, the addition operation will be performed when x and y are available (instead of when the previous statement in a program finishes). This will produce p, which will then be available for other operations that need it. "Programs" can be written by making a graph of the data dependencies - e.g., an arrow is made from the $p = x + y$ node to the nodes of operations that use p. The computer then mimics the graph. Data flow computers may be a good candidate for optical computers because they need a large number of interconnections. (It should be pointed out, though, that it is not yet known, even in the computing architecture community, how efficient data flow machines will be for performing real tasks, or how easy or difficult it will be to program them in the general case.)

Another class of computers is based to varying degrees of accuracy on how the brain works as we understand it. Many models of neural nets have been proposed, and they can be used for computation. For example, see [13]. One method of implementing a simple neural net model has been described by Hopfield [14] and others (Hopfield was actually not the first to describe it). It consists of a matrix-vector multiply, followed by thresholding of each vector element, and the resultant vector is fed back to the input. The vector elements are binary valued after thresholding. It has received a great deal of attention, primarily because it is relatively simple and is workable. These models are of interest in optics for the following reasons: (1) They typically use a distributed representation for memory - i.e., one memory element is not located at one particular neuron, but instead corresponds to a state of a collection of neurons, which can provide for fault tolerance ("graceful" degradation of memory); (2) The human brain has a very large number of neurons (estimates range from 10^9 to 10^{11}), each of which has a very slow response time by electronics standards (approximately. 10 ms); current parallel optical switching arrays (spatial light modulators) also have slow switching times, but can implement a fairly large number of elements (presently 10^5 to 10^6 switches per device); (3) The number of connections between neurons is very high - estimated at 10^3 to 10^4 connections to each neuron in the human brain; optical systems should be more viable than electronic systems for implementing these interconnections.

Finally, computers for AI and inference machines could be considered as a separate category [15]. While they may be implemented via some of the above architectures (for example the neural net models), other architectures, specifically intended for these applications, may be desirable. Some of these architectures are massively parallel, and thus may fit optical systems well.

Other methods of classifying optical computers by functional architecture include classifying the instruction sets, and classifying by the grain size. A large grain system has a small number of large (powerful) PEs, whereas a small grain system has a large number of small (simple) PEs.

2.4. Examples of Optical Computing Systems and Architectures

In this section systems will be described that have been demonstrated, have been described, or might be worth considering. Four basic systems will be described which have been reported in the literature. A generic architecture will also be given to provide unity and a perspective on past, present, and future optical computing systems.

One system has been described and experimentally demonstrated by S. H. Lee, et al. [16], and essentially consists of a gate array and memory arrays, with beamsplitters, mirrors, etc. to connect them. The system functions under the control of a microprocessor and is a SIMD machine. The gate arrays can perform logical operations such as AND, NOT, etc., the particular operation being chosen by the controlling microprocessor. The system can perform operations such as STORE IMAGE A, A OR B (pixel-by-pixel), SHIFT IMAGE A ONE PIXEL TO RIGHT, LEFT, UP, or DOWN. It can execute programs that are stored in the electronics. It has, however, two drawbacks: it can only shift images one pixel at a time in any direction, and it utilizes mechanically rotating mirrors as part of the interconnections. They have also described somewhat more general version that utilizes two gate arrays [17], but still uses rotating mirrors.

A hybrid optical/electronic implementation of a model of neural nets (described in the previous section) has been demonstrated by D. Psaltis et al. [18]. It has a 1-D array of inputs and outputs and an optical matrix-vector multiply in between. The thresholding and feedback were done electronically, although in principle it could also be done optically using a threshold array and an image rotator or fiber bundle. A different optical architecture has been described by Fisher and Giles [19] that implements a version of a neural net, but is more adaptive than the implementation of Psaltis et al. It uses spatial light modulators that do more than just threshold.

A Huang et al. have described and are presently working on an experimental demonstration of an optical computer that works on the basis of symbolic substitution. The basic concept and early designs of the optical system are given in [20]. (The optical system he is working on now is different from that described in the reference, but its function is still that of symbolic substitution.) The basic stages are first symbolic recognition, then an inversion (e.g., NOR operation), then symbolic replacement; the outputs can then be fed back to the inputs. The symbols are binary patterns of light intensity. The computer has a set of substitution rules to implement a given algorithm (e.g. replace symbol A with symbol C, and symbol B with symbol A, etc.). They have proved that a computer based on symbolic substitution operations can be a general-purpose machine. They have given simple examples for rules that can be used to perform digital addition. Two arrays of digital numbers can be added element-wise in essentially m steps, where m is the number of bits in each digital number to be added (thus the number of steps is independent of N, the array size). Other operations, such as multiplication of a scalar and a matrix is also dependent only on the precision of the operands. Applications they are considering are array processors and switching networks. Other possible applications include simulation of physical processes, and artificial intelligence where computation is done by applying a set of rules instead of numerical computations [21].

The last optical architecture we will describe is the free-interconnection sequential optical logic processor of Jenkins et al [22]. It also consists of a 2-D active array, in this case of NOR gates, and uses the third dimension for the interconnections. Here the gates are interconnected by an optical holographic system (Fig. 1). Since these

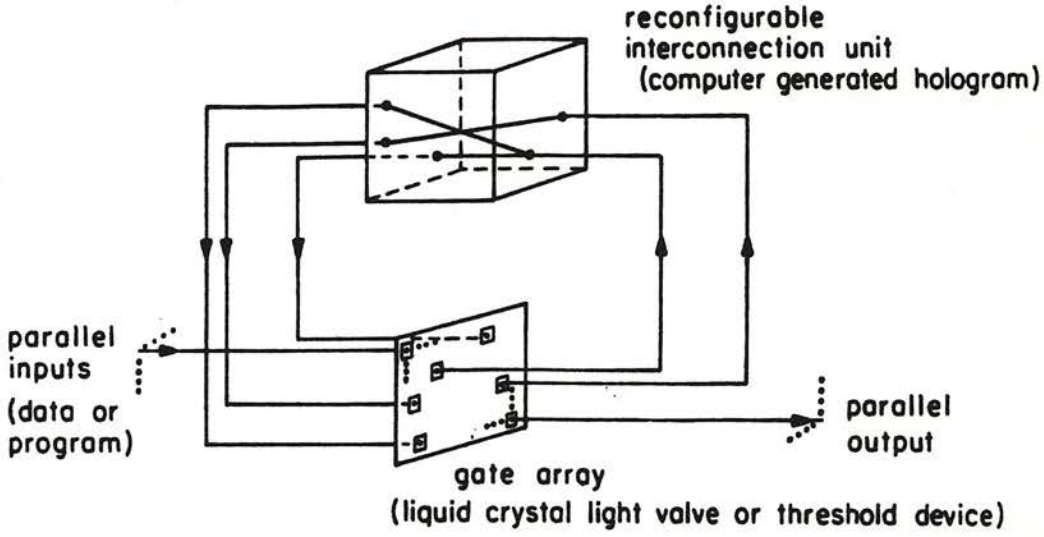

Fig. 1. Block Diagram of the parallel optical sequential logic system.

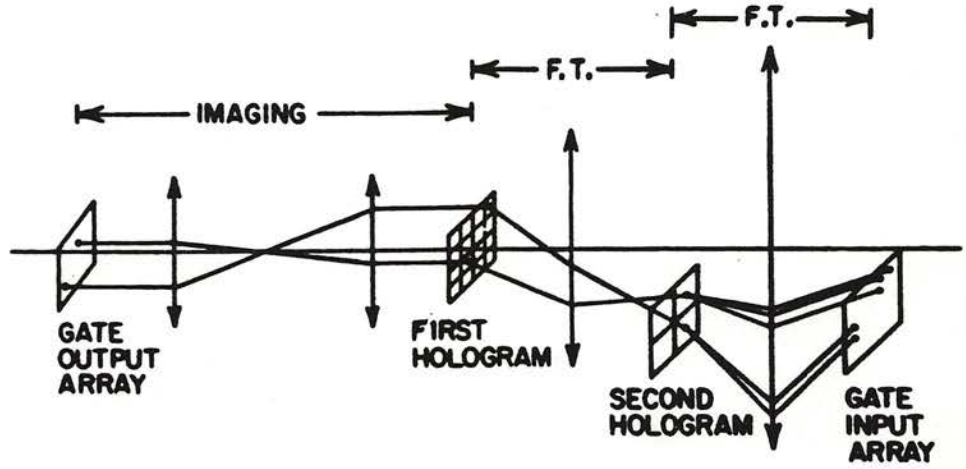

Fig. 2. Hybrid or basis-set interconnection system for the logic system of Fig. 1. The input is a gate output array and is imaged onto the first hologram. There is one hologram or facet for each point in the gate output array, which deflects the light through one of M subholograms in the second hologram. These subholograms act as space-invariant filters, each of which stores one interconnection pattern.

interconnections include a provision for feedback, then clocks and memory can be constructed in addition to combinatorial logic; these are the minimum hardware requirements to be able to implement, in principle, arbitrary digital processing operations. This system allows large numbers of interconnections between gates. Its inherent 3-D structure provides for a high degree of interconnection flexibility. The holographic interconnection system connects the output of each gate to inputs of other gates, effectively wiring up a circuit. For ease of manufacture, the holograms can be generated by (electronic) computer and written out using a computer plotting device. The concept of this system has been experimentally demonstrated with a 16-gate circuit; the circuit consisted of a master-slave flip-flop and driving clock [23].

Three different optical interconnection systems for interconnecting the gates have been described [22]. All of them use holograms in conjunction with free-space propagation. Their characteristics differ and this manifests itself in the kinds of circuits and processors that can be implemented most efficiently with each system. These systems are briefly described in the following two paragraphs, and in the third paragraph possible applications and functional architectures for the system are described. First we will discuss one of the interconnection systems, which is a hybrid space-variant/space-invariant system. This system has the most general applicability.

The hybrid interconnection system represents a basis-set approach to interconnections. The optical system consists of two holograms and two or more lenses (Fig. 2). The idea is to define a finite number, M, of distinct interconnection patterns, and then assemble the circuit using only these M patterns. (A circuit with only one interconnection pattern is shown in Fig. 3 - all gate outputs have exactly the same interconnection pattern.) We generally expect that $1 << M << N$, where N is the number of gates. Each different interconnection pattern is essentially stored in only one place on the hologram, and many gate outputs can use this same interconnection in a non-interfering manner. The number of gates and interconnection patterns that are implemented determine the complexity of the holograms. The hologram complexity or space-bandwidth product that can be achieved is limited by the capabilities of recording devices (e.g., computer plotting devices). Calculations indicate that with current plotting devices, if there are $M = 50$ interconnection patterns, then $N \sim 10^7$ gates can be interconnected [22]. Increasing M will decrease N, such that MN is constant. (Future gate arrays are expected to have $\sim 10^6$, possibly up to 10^7, gates.) Thus with this approach the designer has some minor limitations on the interconnections which can be used, but he has a potentially large number of gates at his disposal.

Two other interconnection systems have been described, and they are essentially special cases of the above system. A space-invariant system allows for only one interconnection pattern ($M = 1$), but can interconnect $\sim 10^8$ gates. Ways of making a useful circuit with only one interconnection pattern are described in [22]. The other extreme is a space-variant system which allows for completely arbitrary interconnections ($M = N$); it can be used to interconnect only up to $\sim 10^4$ gates with current hologram writing technology. Each of these systems uses just one hologram.

Of course, with a large enough M, any circuit can be implemented with the hybrid interconnection system. However, the potential of this system can be exploited more fully by implementing circuits with a high degree of regularity or symmetry. An example is a processor array. Typically interconnections between PEs have a considerable

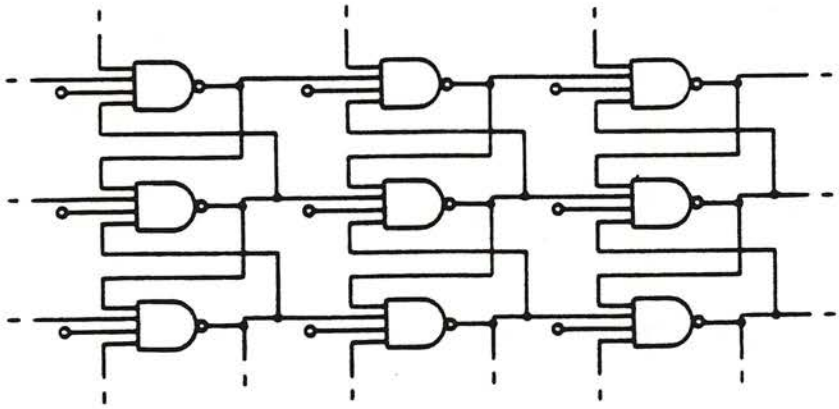

Fig. 3. Example of only one interconnection pattern being used for all gate outputs (M=1).

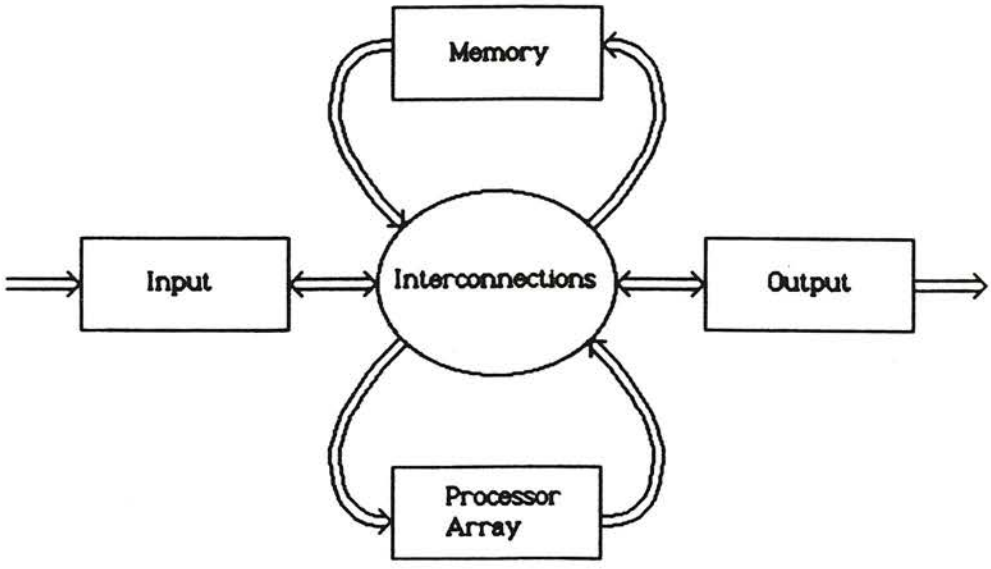

Fig. 4. Generic optical computing architecture.

amount of regularity, which can reduce the size of M. Examples include mesh-connected arrays, pyramids, and hypercubes. The interconnections within each PE may be completely arbitrary. The fact that many of the PEs would typically be identical provides a major reduction in M, since the within-PE interconnection patterns are only stored once. We also point out that whether the interconnections are global or local has essentially no effect on M or N.

Possible architectures for future optical computers include any of those in the previous paragraph. Other functional architectures discussed in the electronics community are also possibilities. The prime considerations are use of parallelism, a careful use of the number of gates, an exploitation of the interconnection capabilities of optics, as discussed in Sec. 2.1. Systems for symbolic and AI computing machines may be well suited to optics, because of the parallelism that can be exploited in searching and other needed operations. In addition, fewer thresholds or nonlinearities may be required than in typical numerical processing.

A generic architecture that encompasses almost all optical computing systems (both digital and analog) to date is given in Fig. 4. It may be useful in providing a perspective on past and existing systems and a framework for new research ideas. All of the above systems map into this generic architecture. This architecture can also be cascaded to make larger systems.

2.5. Algorithms and Applications for Digital Optical Computers

Some of the design constraints and characteristics of optical computers and electronic computers were given in Sec. 2.1. The use of an optical computer is appropriate for problems that utilize some of the advantages that optical systems have over electronic systems. Briefly, this is when (1) global interconnections are needed, a large "density" of interconnections is needed, the interconnection topology does not map into 2-D layers efficiently, or a high bandwidth is needed in the interconnecting lines; (2) parallel input/output (I/O) is needed or partitioning (into chips) is impossible or inefficient; (3) the input and output are already photons; (4) large amounts of parallelism are needed or beneficial; (5) immunity to electromagnetic interference is desired.

Some examples of problems that digital optical computers could be useful for follow:

(1) *Image processing and image understanding.* A highly parallel architecture can be used for image processing, namely a cellular logic machine that puts a PE behind each pixel of an image. Some of the needed operations require global communication, and these include histogramming, regional property computations, transforms. In addition, communication between PEs and memory, and input and output of data to the PEs, put high demands on the communication and interconnection requirements. This communication can be a major factor in the computation time when these architectures are implemented with large electronic cellular logic machines [24]. In image understanding computational requirements can be extreme, and again a large amount of parallelism is involved.

(2) The *fast Fourier transform* algorithm mentioned earlier requires global communication among its data elements, and could in principle be easily implemented with a highly parallel optical machine.

(3) *Communications switching.* Here the input and output are already photons. In addition, the speeds required will be too fast for electronic switching systems. Here the possibility of very fast optical switches (sub-picosecond) gives them a distinct advantage over electronics. (This advantage cannot be exploited everywhere because the optical power needed may be rather high. A 10 fJ switching energy, if switched every 0.1 ps, implies a power of 0.1 W per gate.)
(4) *Simulation of large physical systems.* This can involve a high degree of parallelism because of the physical processes involved. Examples include weather prediction, aerodynamics, fluid mechanics.
(5) *Monte Carlo simulations*, or simulation by running repeated trials of a process and accumulating statistics. Conventional computers cannot generate the needed truly random numbers, whereas by using speckle an optical system in principle can.
(6) *Artificial intelligence, expert systems, and symbolic computing.* These involve a large amount of computation using traditional methods on conventional computers. There may be large amounts of parallelism involved. Many of the required basic operations such as sorting and searching can in principle be done in highly parallel ways with optical systems. A simple optical correlation is in effect a parallel search over a 2-D field of data in one step.

This list is not all-inclusive, but is meant to enumerate some examples.

3. COMPONENTS: GATES, I/O, MEMORY; MATERIALS, DEVICES

In this section components needed for DOC systems will be discussed. The current state of these components and current research pertinent to DOC will be assessed.

3.1. Gates

Classes of optical gates were given in Sec. 2.2. Space does not permit a discussion of all classes here. The most promising technology for gates has come from the area of optical bistability. Typically a nonlinear optical material is placed in a resonant (Fabry-Perot) cavity. The result is a nonlinear optical switch that can optionally provide a bistable response. For gates in an optical computer, a bistable response is not needed (or wanted). An inverting response is preferred - a logically complete set of operations is needed. This requires either an inverting response, e.g. NOR, or AND and NOT, etc., or an inversion only at the input of the processor if both the signals and their complements are kept at every stage of the processor. Fortunately, these bistable devices can exhibit an inverting response.

3.1.1. Devices

In the last few years rapid progress has been made in optical bistability for logic gates. Switching energies between 1 and 10 pJ have been experimentally demonstrated for devices that have response times of 1 to 10 ns [25], and that have a NOR response (other responses have also been demonstrated.) These are nonlinear materials in a Fabry-Perot etalon, and actually respond to power density, not total power. These devices typically switch much faster in one direction than the other, i.e. switch-on vs. switch-off (they are driven in one direction but relax in the other). The numbers quoted above refer to the slower of the two switching times.

Another promising device is the self electro-optic effect device (SEED) [26]. It is based on the effect of increasing absorption with increasing incident intensity. It utilizes electronics (voltage source and resistor) in addition to the optics. It has also been experimentally demonstrated, and has the lowest projected power of any usable device so far. The experimental device has 600 μm diameter active area for one gate, and has an optical switching energy of $\sim$ 1 nJ; this is an energy density of 4 fJ/μm^2. As the device is scaled down, its speed increases and its power requirements decrease. Scaled down to the physical limit (λ^2/n^2 gate area), it has a projected power requirement of $\sim$ 1 fJ (electrical plus optical power). The importance of the power requirement will be discussed below. Switch-off times of 400 nS were demonstrated (limited only by power), whereas switch-on were up to 5 times higher. Switching times of 1-10 nS look feasible.

Very recently, at the Optical Society of America meeting in October 1985, some pertinent results have been announced. Partially successful results were reported of 2 x 4 arrays of the Fabry-Perot etalon devices mentioned above for implementing NOR gates [27] with spot (gate) diameters of 6-8 μm and spacings of 30 μm. The device was not completely uniformly thick, however, and the same gate operation was not achieved over all pixels (gates) simultaneously. In addition, a different device was reported [28] that has demonstrated optical switching energies of 0.5-3.0 fJ and response times on the order of ns. This is a hybrid device and had total (electrical + optical) switching energy of approximately 20 pJ. It is an active device which is an InGaAsP laser amplifier in a resonant cavity. Arrays of large numbers of these may be more difficult to make due to power dissipation but in principle can be done. It also demonstrates that optical switching can be performed at these very low incident power levels, and is compatible with laser diode wavelengths. One other result was a 200 ps recovery time of optical gates in a Fabry-Perot etalon array [29]. This is the fastest recovery to date of an optical gate of this kind and is inherently suited to gate arrays. It was also performed in a GaAs/GaAlAs structure in a Fabry-Perot etalon.

Power is an important consideration, not only because of power dissipation of the elements, which puts constraints on how closely gates can be packed together, but because of the source power required. Statistically, 1000 photons per switching event are required for reliable switching [30]. (This assumes the gate bases its decision on all 1000 photons, i.e. they all interact with the material. As some detectors have quantum efficiencies $\sim$ 1.0, this may be achievable. Also, it is conceivable that there may be ways of reducing this 1000 photon limit, but as yet there has not been any work in this area. This might be done by using other methods of encoding 1's and 0's, instead of intensity.) Devices that utilize a signal level encoded as intensity, as do the above bistable devices, are limited by this number. 1000 photons at $\lambda = 0.5\mu$m have an energy of 0.4 fJ. For an array of 10^6 gates switching every 1 nS, and allowing a factor of 5 for fan-out and other intensity losses, 2 W of optical power are required. This is approximately what today's highest power 1-D laser diode arrays can provide. The SEED device above, with a gate area of 10μm (approximately the minimum that could be used in a gate array in an optical system) would require 10^4 to 10^5 photons per switching event. This is surprisingly close considering current spatial light modulators require $\sim 10^9$ or more photons. Improvements in switching powers have been fairly rapid, and may continue to improve even more. An example is the switching laser amplifier device mentioned above with fJ switching energies.

3.1.2. Materials

The most promising materials for bistable gates are III-V materials, particularly GaAs or GaAs/GaAlAs multiple quantum well (MQW) or superlattice structures (crystals that are essentially built by alternating thin layers of two materials to alter the microscopic and macroscopic properties), InSb, and InAs (the II-VI materials ZnSe and ZnS are also worthy of mention but are probably less promising), although there are other materials that may prove viable as well. The methods of molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD) are important and extremely helpful in materials development. They essentially enable new (previously unexisting) materials to be manufactured. The GaAs/GaAlAs MQW can provide substantially larger optical interaction coefficients than GaAs. The GaAs or GaAs/GaAlAs materials also have the advantage that they can easily be used at laser diode wavelengths (0.8-0.9μm is common).

3.2. Memory

An important point in the implementation of memory for an optical computer is to avoid copying the von Neumann architecture of conventional electronic computers. The purpose of the addressing mechanism is to reduce the number of communication lines to the memory, and in optics the communication lines are cheap. In this section different methods of implementing optical memory for use in an optical computer will be discussed.

Memory can be made out of gates and interconnections, e.g. by wiring up flip-flops. These can easily be interspersed with processor gates. Bistable elements could also be used, an array of these providing a 2-D optical memory (this is essentially the same as the flip-flops, except the feedback is internal to the gates instead of external, and it would most likely be a physically separate device from the gate array). This would provide a maximum memory size of 10^6, maybe 10^7, bits per array. It could be easily accessed in parallel from the CPU or gate array.

Larger memories may be possible by using holograms. A fixed holographic memory (read-only memory) could be a planar or volume hologram, the volume hologram providing more storage. The volume hologram can store pages of data, but the number of pages cannot become huge because of limitations on dynamic range of the material, crosstalk, angular sectivity, etc. Holographic memory could be made to be content addressable. It could also be used as part of an associative memory. A dynamic hologram could be used to implement a read/write memory. Four wave mixing in photorefractive materials such as $BaTiO_3$ can be used to write holograms in real time. This method is not practical yet because of the high power required and/or the slow time response of the material.

Holographic content-addressable read-only memory has been described and experimentally demonstrated [31,32]. It can also be used as a truth table look-up memory to implement gate operations.

Architectures for implementing optical associative memory have been described [18,19], and demonstrated [18]. As described previously, an associative memory has data as input and the associated data as output (no addresses). They essentially provide a match of the input data to one of the stored data words. These also provide "best-

match" capability, i.e. when the input data does not exactly match any of the stored words, the memory finds the closest match (usually minimum Hamming distance) to one of the stored words. Such memories have applications, for example, in pattern recognition and in AI operations such as searching.

A memory could consist simply of free-space propagation. Arrays of data, in passing from one device to the next, are effectively being stored in a free-space memory, if the propagation time is longer than the gate response time or machine clock cycle. Thus this memory could be called "time addressed" (although I have not seen this term used in the literature). In this case the computer has to be designed so it can use the array when it arrives. Light travels 30 cm in 1 nS, so switching times $<<$ 1 nS are needed for small systems. Since switching times much less than 0.1 ns may be impractical in optical parallel computers for reasons of power consumption (cf. Sec. 3.1.1), this kind of memory may not be practical.

Finally, optical disks are mentioned because are at an advanced stage of development. However, they were developed for electronic computers and have a von Neumann addressing scheme. Possibilities for parallel readout exist - by illuminating part or all of the disk in parallel. The data would then have to be reformatted by an optical interconnection or coordinate transforming system. Whether this is practical is not yet known. Multiple heads might be usable for writing but would most likely only provide a small number of parallel write channels. A 1-D continuous beam to write all tracks in parallel would provide more write channels, but again no one has addressed this question; whether the power requirements would ever be practical, and what other problems there might be is unknown.

3.3. Input/Output

Input and output devices must also be considered. What is needed is application dependent. Initially, electronic-to-optical and optical-to-electronic transducers will be needed since the optical computer will probably not be a stand-alone system. Ultimately, however, they may be less important as interfacing to electronic computers might not be as vital.

The fastest transducers are laser diodes for electronic-to-optical conversion; and detectors (almost by definition) are the primary means of optical-to-electronic conversion. (Another possibility for optical-to-electronic conversion is direct input to an electronic memory where the optical signals create charge in the corresponding memory cells; this is essentially a detector array with memory.) Laser diodes can operate at 10 GHz and higher; detectors can also operate close to this speed.

Arrays are slower, particularly in the case of detectors. Here the problem is primarily in the electronics - current detector arrays do not have parallel readout. The signals are multiplexed onto a line electronically, then read out serially. Thus the electronics creates a bottleneck which slows down the detector array. In many cases the amount of data is reduced as computation takes place within the optical computer, so such a bottleneck can be tolerated. An example of such a computation is pattern recognition or image understanding, where an entire image is input but only a few words are output.

Laser diode (LD) arrays currently only exist as 1-D arrays. This is primarily because they are edge-emitting devices. Research is being pursued on developing

surface-emitting LDs, so that they can be integrated in the middle of a chip. The problems are in developing appropriate LD geometries and processing techniques. Power dissipation also presents a problem for densely packed LD arrays. LDs are pertinent not only for (signal) input transducers, but also for inputting optical power into the system (i.e., optical power supply).

A variety of spatial light modulators (SLMs) exist for input to an optical computer. These devices modulate light rather than creating it (a simple uniform beam is then used as the input to the SLM). Table 1 summarizes properties of the most pertinent ones, categorized by digital electronic to digital optical (DE → DO), with 1-D and 2-D optical outputs, and analog optical to digital optical (AO → DO) - i.e., an optical A/D converter. Some of these devices have an inherently analog output (which can be used for digital signals); others have inherently digital outputs - thus the bottom line in the table. An optical A/D has been demonstrated in principle [33] as a 2-D image parallel system. The frame time with current SLMs is 10 - 100 mS; it will improve with faster SLMs. Integrated optic A/D converters have also been demonstrated [34] and can operate at Gb/S rates but only operate on one input channel at a time; an array of 10^6 elements would thus take 1 mS.

4. CHARACTERISTICS AND COMPARISON TO ALTERNATIVE TECHNOLOGIES

In this section digital optical computing will be compared to electronic, integrated optic, and acousto-optic computing. Many of the general architectural features and characteristics of DOC have been discussed in Sec. 2.1 and 2.5. Here more specific questions are addressed. Comparisons can be made on the basis of current technology, projected future technology, or on fundamental characteristics and limits. In the case of present technology, DOC appears inferior to all of these alternatives, primarily because it is a newer, less mature, and less developed technology. Estimates of projected future state-of-the-art will be given wherever possible; these will be based on the current state of knowledge in DOC and research directions that are currently believed to be the most promising. Fundamental limits will also be given wherever they are known; these are more important than the other considerations but can in some cases turn out to be misleading. Any derivation of fundamental limitations relies on assumptions on the physical mechanisms used; when new physical mechanisms are discovered or different ones used for the process, the derived fundamental limits no longer apply.

4.1. Throughput, switching speed, and power

Power considerations alone dictate many of the characteristics of optical computing systems. Any nonlinear computing involves a detection or decision process. The statistics of this process forces a requirement on the number of photons per decision in order to obtain reliable, error-free switching. This, when compared to expected practical optical source powers, affect response time, array size, and therefore throughput. The analog and digital cases must be considered separately. The relation (switching energy) / (switching time) = (switching power) will be used for both.

For the digital (optical) computing case, statistically 1000 photons are needed per decision or switching event, which translates to 0.25 fJ for a near-infrared wavelength of

Table 1. Specifications of input devices. Abbreviations: digital electronic (DE), digital optical (DO), analog optical (AO).

	DE→DO (1-D)	DE→DO (2-D)	AO→DO
SIZE	30-5,000	256 x 256	500 x 500
FRAME TIME	10NS - 100 μS	1μS - 10MS	10 - 100MS
INPUT FORMAT	SERIAL OR PARTIALLY PARALLEL	SERIAL OR 1-D PARALLEL	2-D PARALLEL
OUTPUT LEVELS	ANALOG OR BINARY	ANALOG OR BINARY	BIT PLANES

0.8 μm. A gate or threshold array of 10^6 gates or pixels is assumed, because it appears attainable from both the optical system and the device point of view (current analog SLMs have close to 10^6 resolvable elements). If we further assume an optical power source of 10 W per gate array is attainable (e.g. using laser diode arrays), allow for a factor of 4 light loss for fan-out and system efficiency, and assume the detector (or nonlinearity) has a quantum efficiency of 1.0 (attainable with some present detectors), the maximum clock rate attainable is on the order of 10 GHz (0.01-0.1 nS switching times). Current electronic systems have clock rates up to 100 MHz. While optical switching has been demonstrated at sub-picosecond rates, from this analysis it is apparent that for parallel processing and computing with large gate arrays, such response times will not be used. Since this assumes that most gates are switched at or near the clock rate, there is a possibility of faster switching times if on the average the gates switch (or make a decision) a smaller percentage of the time. It should also be noted that other methods of encoding the 1's and 0's (such as polarization [35]) may reduce this 1000 photon requirement, permitting faster operation.

In the analog case, the number of photons required per switching event depends on the number of levels for each signal; for 1000 levels approximately 10^6 photons are required per decision [36] (a decision effectively assigns one of the 1000 levels to the signal). For the same assumptions as above, the maximum rate of successive decisions is 10 MHz.

Recently, in electronics a 28 pS delay Si MOSFET switch and a 12 pS delay GaAs/GaAlAs switch have been demonstrated [37]. In optics, $\sim$ 0.1 pS switching has been demonstrated and optical pulses $\sim$ 30 fS have been demonstrated [38]. One should not, however, draw conclusions on computing ability from switching times alone. In the electronics case, switching times that are too fast are impractical because they cannot be interconnected (crosstalk, capacitive loading, and interconnection power dissipation increase with faster signals); in the optics case, switching times that are too fast may be impractical because there is not enough optical source power available to switch them at that rate (cf. previous paragraph).

For the DOC case, 10 GHz x 10^6 gates yields 10^{16} gate operations per second per "chip"; a more conservative estimate would use 100 MHz yielding 10^{14} gates operations per second per chip. For the electronics case, 100 MHz and 10^6 gates on a chip may also be attainable. Electronic computers are commercially available with 100 MHz clock rates; VLSI MOS circuits are at present limited to about 30 MHz; and Si MOS and GaAs MESFET technologies have been demonstrated with GHz clock rates, although the level of integration is much lower than with VLSI [39]. 10^5 gates have already been achieved on a commercial electronic chip; 10^7 gates per chip is predicted by some. Distributing a fast, synchronizing clock over such a chip is nontrivial, and may not be feasible. The result may be relatively slow clocks (10 MHz for large gate counts) or resorting to asynchronous systems (locally synchronized but globally asynchronous), or a design trade-off between gate count and clock rate.

Thus a future parallel computing system in electronics may have 10^7 gates per chip, a clock rate of 100 MHz which might not be able to be distributed over the entire chip, and in optics may have 10^6 gates with a clock rate of 1 GHz that can be distributed over the entire chip. Each gives a maximum of 10^{15} gate operations per second per chip. But when the interconnection characteristics are compared, the free-interconnection and

parallel I/O capabilities of optical systems give them a distinct advantage. The limitations on the interconnections in electronics would most likely limit the throughput of algorithms to a substantially lower rate.

In parallel processing systems the speed-up is defined as:

$$\text{speed–up} = \frac{\text{execution time on 1 processor}}{\text{execution time on N processors}}.$$

It is generally less than N (in fact is conjectured to be proportional to log N in the general case - see Fig. 5), because of: (1) communication limitations, (2) contention (e.g. two or more PEs accessing a memory simultaneously), and (3) algorithm limitations (inherent sequentiality). Optical systems have the potential of substantially reducing (1) and (2).

As an example of the effect of communication limitations, Kushner and Rosenfeld [40] have enumerated classes of communication tasks in parallel processors for image processing. For image processing tasks, local operations (e.g. pixel-by-pixel multiplication of two images) can be done in O(1) steps (i.e. the execution time is independent of the image size). But more global tasks, (as needed in, for example, histogramming and transforms) take O(N) or $O(N^2)$ steps for simple locally-connected arrays. For large arrays these can thus dominate overall computation time of algorithms. By increasing the connectivity of the processor array (as in an optical system), many of these global computation tasks can be performed in $O(\log_2 N)$ time.

In acousto-optics, a very large number of systems for performing matrix operations have been described. The limiting elements in terms of speed are the AO device input and the output detector (array) and any formatting or conversion of the output data. The output detectors, geometry, and any conversion of the data format varies from architecture to architecture. Common to all AO architectures is (are) the AO device(s); from the associated speed we can derive a limit on the speed of all AO architectures. AO devices can have carrier frequencies up to approximately 4 GHz. For analog signal levels we cannot expect to modulate the signal at more than 1 GHz (this depends on the number of levels, accuracy, and degree of isolation or crosstalk; minimal requirements on these may permit up to 1 GHz). In the best case a system can then perform a multiplication and addition at GHz rates. A strobed system would perform on the average one multiplication and addition each nS; for a multichannel AO cell (up to $\sim$ 100 channels are available), processing rates of 100 GOPS (Giga analog operations per second) may be attainable. With a non-strobed system, for example some systolic systems, rates of 1GHz times [the number of data elements in the time window of the processor] may be attainable. With 200 such elements in a 100 channel AO cell, the upper limit is 2×10^4 GOPS. These numbers represent generic upper limits; what a given system can actually achieve is determined by the architecture and the operation being performed. An architecture that approaches this limit using multichannel cells has been described that may be able to perform up to 1000 GOPS [41]. For comparison, commercially available electronic array processors perform from 1 MFLOP (Mega floating point operation per second) at a price of a few thousand dollars to 100 MFLOPs at a price of a quarter million dollars.

Another important limitation is due to the detection process and number of photons requirement as described above. However, this is not of concern in some of these AO processors because they integrate over time or space before each detection (thus

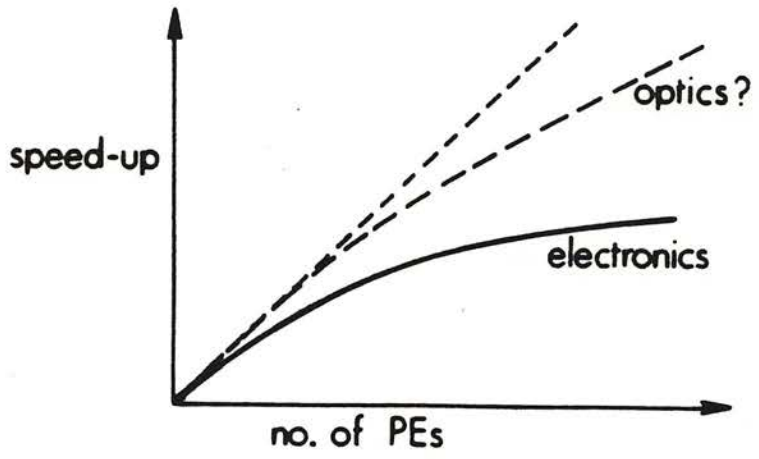

Fig. 5. Speed-up vs. number of processors for parallel computers, showing the conjectured curve for electronic machines, the ideal case of linear speed-up, and where digital optical computing may fit in.

increasing the number of photons per decision). The electronics that read out the detected information is a bottleneck in some architectures - detector arrays with parallel read-out do not presently exist to our knowledge, and shift-and-add or multiplexing output detectors are limited by the speed of the output line(s) (to 1-100 MHz), and typically cannot operate at above a few GOPS [41]. Processors that involve a large amount of data reduction circumvent this problem by only outputting a relatively small amount of data.

The accuracy of these analog AO processors is limited to approximately 8 bits. Other systems have been described with higher accuracy by trading off number of data elements or throughput. Many of these systems (including digital multiplication by analog convolution (DMAC)) use binary representation of the numbers at the input, and the system outputs a mixed binary representation - i.e., each digit of an output number represents a power of 2 as in the binary case, only each digit can range in value from 1 to n. The electronics performs an analog to digital conversion on each digit and then converts to a proper binary representation. Unfortunately, this electronic conversion has proven to limit the speeds of these processors to levels that are nearly attainable with all-electronic systems. Note that in the case of AO matrix processors there is a direct speed competition with electronics since these operations can be performed with only local interconnections (e.g. systolic arrays). Ref. [42] classifies different high accuracy AO matrix processors and gives the hardware requirements on the electronics and the associated speed for each.

Integrated optics can provide signal processing operations at high throughput rates but has very different characteristics from DOC. Integrated optic systems process individual signal lines or at most a 1-D array, whereas DOC can also process 2-D arrays. Integrated optic systems are inherently a 2-D geometry. If problems of waveguide routing and waveguides crossing over each other get solved, it would still suffer from many of the limitations of electronic systems because of the 2-D constraints. In addition, the physical limit to the minimum feature size is equal to or larger than that in electronics (0.8 μm vs. 0.3-0.5 μm). Very recently the possibility of sending signals in an integrated optic chip directly out of the plane into the third dimension has been described [43]. This offers new hope for the use of integrated optics in optical computing systems. Possibilities could include input or output devices, but have not been considered yet primarily because this is a new result first announced publically in October 1985. Some of the same technologies for gates can be used in integrated optics as for DOC, and similar limits apply.

4.2. Reliability, radiation effects

Reliability will be discussed on two levels - physical failure of the hardware and computational errors. Computational errors will be discussed first. In a DOC computational errors may be due to fluctuations of the laser source power output, statistics of the detection process, or system effects such as crosstalk from the interconnections or stray light. The 1000 photon per gate requirement discussed above assures reliable detection; whether it is sufficient to also account for the laser power fluctuations depends on the method used to stabilized its output power, and has not been specifically addressed. Some researchers assume that 1000 photons is sufficient for suppression of both effects [30]. In the semiconductor electronics case, many kT must be expended per

switching event for reliable switching. At room temperature kT = 0.004 aJ whereas one 0.8 μm wavelength photon has energy 0.25 aJ (1 aJ = 10^{-18} J). Integrated optics gates would also require 1000 photons. Fundamentally, then, reliable switching can probably be attained in optics and in electronics, but the electronics case has a power advantage. Interconnections must also be considered. Here electronic systems are less reliable than bulk optical systems, because of stray capacitances and crosstalk. AO systems are not easily divided into gates and interconnections; however the reliability of AO devices has in principle been demonstrated by commercial uses of some AO devices.

The vast majority of physical hardware failures in the electronics case is due to the interconnections [37,44], including connectors, wires, plugs, soldered connections, and on-board and on-chip interconnections. Electromigration is an example of failure of on-chip interconnections that becomes more severe as dimensions are scaled down, and is discussed in Sec. 5.1. Another reliability issue is yield in the manufacturing process; this is an extremely important consideration in larger scale and particularly wafer scale integration. Circuits must be designed for very high degrees of fault tolerance; the probability of an entire wafer being manufactured with no hardware faults is extremely low. An additional consideration is the lack of isolation from gate to gate and circuit to circuit in electronics. A short on a chip or wafer can draw too much current from neighboring gates or circuits, causing them to fail and the process can be catastrophic. The isolation in DOC systems from one gate to the next avoids this problem. Errors in a gate or interconnection of course may appear when large optical circuits are manufactured. Non-functioning gates may be more common in optics because of the requirements on the thickness of the gate array [37], so fault-tolerant designs are again needed but for a different reason.

The author has not found any unclassified information on nuclear radiation survivability of optical computing systems or devices. Optical systems should be less susceptible to electromagnetic interference (EMI) and electromagnetic pulse (EMP) than electronic systems. The interconnections should be relatively immune to EMI and EMP. The effects of EMI, EMP, gamma rays, and neutron radiation on the nonlinear optical properties materials used for optical gates and nonlinearities (cf. materials listed in Sec. 3.1.2) should be investigated; it is possible that some of this has been reported as the author has not performed an exhaustive literature search on this subject.

4.3. Packaging, size, weight, power consumption

4.3.1. Packaging

Semiconductor electronics represents a well-developed, mature technology. Its packaging is known, highly manufacturable, sturdy, fairly reliable and fairly inexpensive. However, in order to achieve these properties compromises have been made. Limited pin-outs of chips, and limited amounts of chip-to-chip and board-to-board wiring are necessary to prevent the cost from becoming prohibitive; this puts constraints on the design and implementation of processors (cf. Sec. 2.1). Integrated optics is one way of packaging optical systems, and it may share many of the advantages and disadvantages of electronic packages, only with much higher bandwidth per interconnecting line. AO systems may be bulk or integrated onto a chip. DOC systems should be bulk; it is the 3-D interconnections that gives them many of their advantages. Their packaging has

not yet been addressed by the research community. However, contrary to electronics, it is unlikely that dramatically reducing the number of communication lines will have much effect on the cost, so DOC systems will likely retain much of their communications advantage. Possibilities include the sequential logic system of Fig. 1, only operating with a reflection device and a reflection hologram to make the system more compact. Stability needs to be addressed in these 3-D optical systems. The mounts need to keep the components aligned well enough to prevent errors in the interconnections; this could be done using rigid mounts or by cementing the components (e.g. hologram and gate array) to different sides of a block of transparent material. It should be noted that interferrometric optical systems are commercially available and have much more stringent stability requirements than a DOC would.

4.3.2. Size

Minimum feature size: Semiconductor electronics is predicted to progress to a minimum feature size of approximately 0.5 μm (at smaller sizes the *physics* changes - whether smaller sizes can be made (and how they would be made) is unknown). Integrated optics has somewhat larger feature sizes. Waveguides of dimension less than a wavelength have extremely large losses; confining waves to the waveguide dimensions may require sizes on the order of 5 wavelengths, thus approximately 4 or 5 μm. Integrated optics switches for coupling waveguides are quite long at present (on the order of 1 cm), but may improve as new materials and techniques are discovered. AO systems typically divide an AO cell into many resolvable spots. For a time bandwidth of 2000 and a 2 cm long cell, the feature size is 10 μm if all points on the AO cell address one point, and is approximately 0.5 mm if N points on the AO cell each address one (or more) of N points in the output plane (these numbers are not fundamental limits; what is attainable depends on the time bandwidth product of future AO cells). For DOC, the size of one gate on a 2-D array is fundamentally limited to 1 wavelength. To avoid crosstalk, a few wavelengths separation will be needed.

Overall system size: In commercial electronics, a VLSI version of a VAX is packaged in a box that measures roughly 1 x 2 x 2 ft, not including peripherals. A VLSI chip can be made with up to 10^5 gates on it, in a size of a few cm. on each side. Taking minimum feature size into account, we may see 10^7 gates on a 3 cm x 3 cm active area for a general processing chip. Integrated optic chips are most competitive when they perform analog special-purpose processing, such as spectrum analyzers. Then they can be as small or smaller in size than electronics. As an example, the synthetic aperture radar processor of [43] will be significantly smaller and lighter in weight than the electronic counterpart, and will be much faster at the same time. AO systems can also be competitive in size for special-purpose analog processing. DOC systems may incorporate a 10^6 gate array with interconnections in a 2 cm x 2cm x 1cm volume (or even 2 cm x 2 cm x 0.2 cm). Overall system size would depend on power dissipation and cooling requirements and other parameters.

4.3.3. Weight

In current electronic computers, the contributing factors to weight are: chips, interconnections and wires, mounting hardware, and power supply. The chips take a relatively small percentage of the total weight. The other factors depend on level of

integration, and decrease in weight as the level of integration is increased and switching energies are reduced. Optical systems have two advantages in reducing weight. One is that the interconnections weigh significantly less; both fibers and free-space interconnections have very high bandwidth and weigh little when compared to the metal interconnections in electronics. Mounting hardware is needed in both cases and will probably be roughly comparable; rigidity is needed for some of the optics mounts but grounding and shielding is not needed. The power supply will probably also be roughly comparable. Laser diodes (LDs) weight little but must be powered by an electronic source; LDs have efficiencies on the order of 50%. Switching energies will probably be roughly comparable - 1 - 10 fJ. (At present electronics is 0.1 - 1 pJ and will decrease with feature size; optics is 1 - 10 pJ for gate operations but very recently optical switching energies αp fJ have been demonstrated) Finally, in the DOC case the weight of chips will probably also be comparable; in some cases, e.g. if a block of transparent material is used, the optics chips may weigh more. The other advantage of optical systems in the weight occurs when they are used as special purpose analog processors, as in AO and integrated optics. In this case a relatively small number of "operations" is needed to perform tasks that would take many more (binary) operations (and thus many more gates and interconnections) in the digital electronics case. An example is the synthetic aperture radar processor currently being developed [43]. It should also be pointed out that estimating the relative weights of future electronic systems and particularly future DOC systems is speculative.

4.3.4. Power consumption

Switching energies have been discussed in the preceding paragraphs because they dictate many of the other parameters that were discussed. Briefly, for digital computing the fundamental limit in electronic gates is approximately a factor of 50 to 100 lower than with optical gates. (The electronics literature claims that statistically, a few to several hundred kT are needed for reliable switching, but practically many more are needed to obtain nonlinearities. In the optics case, 1000 photons are claimed needed for reliable switching. 1 photon has αp 50 times more E than 1 kT at room temperature.) While optical switching energies close to the statistical limit have been recently demonstrated, to the authors knowledge this is not true in electronics.

Now, interconnections can also use power. For the case of DOC, interconnections in principle do not consume significant amounts of power, except for fan-out. Fan-out does simply because each gate requires a certain amount of power at its input; fanning out to n gates requires a total of n times as much power as one gate. In electronics, the interconnections themselves also consume power because they are essentially RC circuits. In practice, only 1/4 to 1/3 of the power dissipated in an electronic chip is due to the gates; most of the rest is due to the interconnections. In the case of analog processors including integrated optics and AO, more photons per decision are required, but fewer decisions are needed. The net power requirement is application dependent. They are attractive when high accuracy is not needed, the problem or algorithm maps well into optics, and the output is relatively insensitive to noise at the device plane.

4.4. Fabrication, cost, availability

The more mature technology, electronics, is the winner here. AO devices are commercially available, although systems using them are generally not. Both integrated optic chips and acousto-optic systems have been built in laboratories. DOC systems are at a more infant stage. The ultimate cost of any of these systems will depend on volume of production. Even the least mature at present, DOC, may be very manufacturable when it is better developed; much of the technology needed to make the device arrays and holograms has been developed for semiconductor electronic materials and devices. Of course, fairly subtle points can affect manufacturability and it is far too early to be able to forsee these for DOC.

5. OPTICAL INTERCONNECTIONS

In this section optical interconnections will be described for use in digital optical computers, digital electronic computers, and communications switching. Both fixed and reconfigurable interconnections will be discussed.

5.1. Comparison of optical and electronic network characteristics

There are many limitations and disadvantages to electronic interconnections in digital electronic computers. Many of these were mentioned in Sec. 2.1. At the gate level, the resistance and parasitic capacitance of each on-chip wire create many problems. The wire acts as an RC circuit and has a time constant. As the feature size is reduced, the gate time response decreases linearly but the wire time response does not change. Soon the wire time response will be the limiting factor. The RC constant and thus the response time of the wire depend on the length of the wire. The implications of this in the design of processors was discussed in Sec. 2.1. In addition, the power consumption depends on the length of the wire, so larger transistors must be used to drive longer wires. Another problem is due to the current density, which increases as device dimensions are scaled down. If all dimensions are reduced in size by a factor α, cross sectional area of the wire is reduced by α^2, so resistance per unit length increases by α^2, but current only decreases by α; current density then increases by α. Electromigration is the physical displacement of atoms in a wire due to the bombardment of electrons caused by the current in the wire; this becomes a problem when current densities are too high. Chip designers have to perform careful modelling of VLSI circuits to make sure current densities stay within operating limits. Also, as mentioned in Sec. 2.1, the device area must be shared between gates and interconnections. Interconnection characteristics have a major impact on the functional architectures that can be implemented.

Optical interconnections do not share these problems. They are non-interfering, parallel, and can be highly dense. The use of optics for communication purposes has been increasing at an enormous pace; its use at the gate level, chip level, and in local networking has received considerable attention. Fiber optic communication for transfer of data has reached a high state of development. Because of the high bandwidth of some optical systems, it is theoretically possible to multiplex data from thousands of processing elements over a single optical channel without causing a speed bottleneck. For parallel computing, electronic gates, circuits, or processors combined with optical interconnections may make a valuable combination. In this section both fixed and

reconfigurable interconnections will be discussed.

Table 2 shows different scales of interconnections, where optics fits in now and how it may fit in in the future. In general, fixed optical interconnections will be used for the lowest level of interconnections (gate to gate and chip to chip), whereas reconfigurable interconnections may be possible for higher levels (wafer and up). This includes local area networks and possibly telecommunications applications (although in telecommunications, large switches are essentially done using computers, since control and overhead represent the major part of the switching operation).

The general problem is one of connecting a 2-D array of points or nodes to another 2-D array of points or nodes with as as much generality as possible. In general, this requires a space-bandwidth product (SBWP) of N^4 for connecting arrays of N^2 points. There are ways of reducing this requirement without a major loss af interconnection generality. For example, the hybrid interconnection system discussed in Sec. 2.4 (Fig. 2) is essentially a multiplexing scheme to reduce this SBWP requirement. Some interconnection networks discussed in section connect a 1-D array to a 1-D array, which in general requires a SBWP of N^2 which can be attained with a 2-D mask or approximated by multiple 1-D stages.

5.2. Fixed interconnections

Fixed interconnections to connect a 2-D array of sources to a 2-D array of detectors can be made using one or more holograms. Three such systems that use holograms were described in Sec. 2.4 as part of the parallel optical sequential logic system. The holograms used can be generated by computer for flexibility and ease of manufacture. Computer generated holograms have relatively low optical efficiencies: the theoretical maximum is less than 10% for binary absorption holograms and less than 40% for binary phase holograms. In practice efficiencies seem to be much lower, although with increases in computing power it may be possible to significantly increase diffraction efficiencies in the future. Another possibility is to make a high-efficiency optical analog copy of a computer-generated hologram. This has been demonstrated; efficiencies of 50% are attainable. One other possibility is an automated system under computer control that exposes the hologram optically with movable point sources. This has also been demonstrated and efficiencies close to 100% have been obtained. All of these results have been at visible wavelengths.

It should be pointed out, however, that it is at present not easy to use optical holograms with laser diode wavelengths. Computer generated holograms could probably be made in appropriate materials but are inefficient. Optically recorded holograms are difficult because at present there are very few materials for recording holograms at such wavelengths, and these materials are very insensitive. Research and development is needed into materials for holograms at wavelengths of 0.8 - 0.9 μm, and perhaps even longer wavelengths; or into methods of making holograms at visible wavelengths that can be used at longer wavelengths.

Goodman et al. [45] have been working on using a fixed volume reflection hologram to interconnect VLSI chips (Fig. 6). The holograms provide focusing power to focus the light from a source onto one or more detectors. They have recently calculated that using a laser diode, hologram, and detector to connect between chips uses a comparable amount of power to the case of electronic interconnections. For on-chip

Table 2. Levels of interconnections.

INTERCONNECTION OF ELECTRONIC COMPONENTS

	SPATIAL SCALE	OPTICS - NOW	OPTICS - FUTURE
GATE	10^{-6} - 10^{-3}m		FIXED (HOLOGRAMS OR FIBERS)
CHIP	10^{-3} - 10^{-2}m		
WAFER	10^{-2} - 10^{-1}m		RECONFIGURABLE (FREE SPACE OR FIBER)
BOARD	10^{-1} - 10^{0}m		
COMPUTER	10^{0} - 10^{3}m	FIXED (FIBERS)	

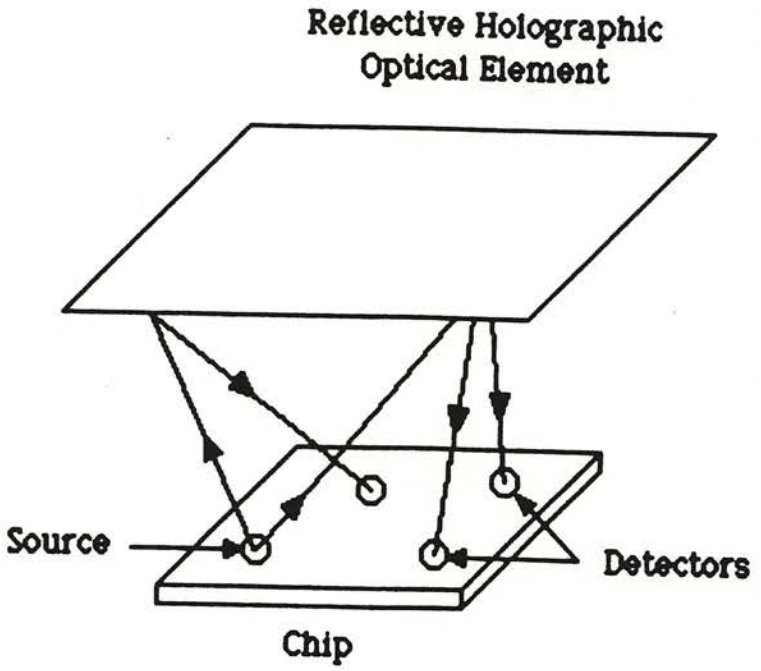

Fig. 6. Use of reflective volume holographic optical element for interconnecting within an electronic chip; current work is actually focusing on connections between neighboring chips [39,45].

interconnections, electronic interconnections at present use less power than optical interconnections. Efficiencies of the source and detector were included in these analyses. To date their work has used visible illumination with LEDs as sources. Other problems that need to be worked out are the integration of the sources and detectors on a chip.

The next fixed interconnection systems to be considered are really part of a reconfigurable interconnection network and represent the perfect shuffle permutation. The perfect shuffle is shown in Fig. 7, and has been implemented by A. Huang et al. [46] and by A. W. Lohmann et al. [47] (Fig. 8) using classical optical components - prisms, lenses, beamsplitters, etc. These systems really do N 1-D perfect shuffles in parallel, each on a 1-D array of inputs and outputs. These systems can in principle implement approximately 1000 perfect shuffles in parallel, each on 1000 points. A perfect shuffle can also be performed on a 2-D array using the hybrid holographic interconnection system of Sec. 2.4 [48], which can implement 1 perfect shuffle on approximately 20,000 points.

The use of fixed interconnections for medium and high level tasks in electronic computers is discussed in [49]. The use of fiber optics for tele-communications has already reached the commercial stage and is in use. Fiber optics for chip to chip communications is being pursued [50]. This is essentially the replacement of some of the electronic pins and connected wires with optical sources and fibers, and provides a much higher bandwidth interconnection than electronic pins and wires.

5.3. Reconfigurable interconnections

Reconfigurable interconnection networks are desirable for interconnecting processors (or PEs) to memories or to other processors. With future electronic and optical technology it is conceivable for parallel computers to have large numbers of processors (and memories). Connecting every PE to every other PE (or memory) with a fixed, dedicated line is not viable because of the number of lines and connections required. A reconfigurable interconnection network that effectively establishes interconnections only when they are needed is a solution. Networks range in complexity from a simple bus to a crossbar. As a compromise between these extremes, many different types of multistage networks with O(N log N) switches have been described [51].

We will discuss two types of networks: crossbars which are single stage networks, and omega networks which consist of repeated shuffle-exchange stages. First we will compare optical networks to electronic networks by enumerating some of the important characteristics and describing how they differ in electronic and optical implementations.

Passive vs. active. Electronic interconnection networks are almost exclusively active, meaning that the signals pass through gates or transistors, getting detected and regenerated as they travel through the network. In the optics case, both active and passive networks have been described. Passive networks involve generation of photons only at the input to the network, and the optical signals propagate through the network and may be detected at the output. In the near future passive optical interconnection networks look viable. Many of their characteristics are discussed below. One major advantage of them is that the bandwidth of each line is limited only by the bandwidth of the sources and detectors. Active optical networks utilize detection and regeneration (e.g. optical gates) throughout the network. This prevents noise buildup and optical power losses, but may increase the total amount of power needed and may slow the network

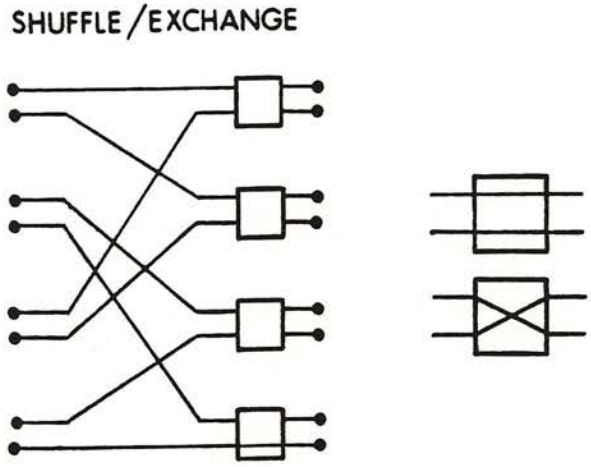

Fig. 7. One Shuffle/exchange stage. The wires perform the perfect shuffle operation and the boxes perform the exchange (i.e. the boxes are 2 x 2 crossbars). A rearrangeable interconnection network can be made by cascading $\sim 3N \log_2 N$ of these stages.

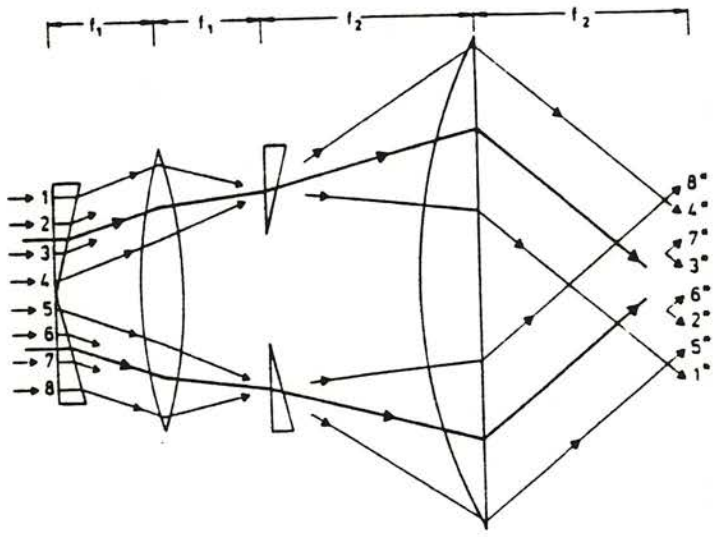

Fig. 8. An optical implementation of the perfect shuffle (from [47]).

down. It is a longer term prospect because of the present status of gate or switch arrays.

Bandwidth. A major advantage of optical networks is the bandwidth of each line in the network. For passive optical networks, the bandwidth is limited by sources and detectors, which can easily operate at 1 Gb/s rates or higher. (Sources (laser diodes) and detectors are available at 20 GHz and 40 ps, respectively. A 1 GHz channel can communicate at 1 Gb/s rate when the signal-to-noise ratio is 1.) In contrast, electronic systems typically operate at 10 Mb/s for each line. Increasing speeds of electronic lines beyond 1 Gb/s will be impractical because of stray capacitances and crosstalk. While active optical networks in principal are not limited in speed as are electronic networks, practically they are limited by optical switching times and power needed to switch.

Reconfiguration time. Reconfiguration of optical networks is limited to approximately 1 μs for moderate or large networks, with current or near future technology; in fact, most switch arrays have a frame time of 1 ms or greater. For an electronically controlled optical switching array, it is a matter of switching time *and* addressing schemes to control all of the switch states; for an array of N^2 switches, N^2 lines are needed to change all switch states in parallel. This is impractical for large arrays. Optical control could provide parallel control, or the switches could be addressed by N lines coming in one side and N lines coming in the other. In the latter case a signal on row i and a signal on column j could address switch ij. This would require N steps to reconfigure the entire array. It should be mentioned that devices (SLMs) with both optical control and electronic control as described above exist; frame times are approximately 100 μs to 100 ms. Practical electronic switches can typically switch in 50-100 ns for each switch. The total reconfiguration time of a network depends on addressing schemes for circuit switched networks, and on decoding and switching for packet switched networks. Fast reconfiguration of the network is needed in an environment where the switching permutations change dynamically. However, slow switching could be tolerated in applications where each connection is followed by the transfer of large blocks of data.

Thus we see applications of optical networks when the reconfiguration time is much less than the time duration of data packets to be transferred; and when there are sufficient processors and parallelism for many simultaneous data transfers. These conditions hold, for example, for block-oriented data and large file transfers that occur frequently in graphics and image processing applications. (By configuring the data and network differently, however, it may be possible to reduce the reconfiguration time in the optics case. This may involve trade-offs with bandwidth.) For smaller networks ($N < 50$) and for longer term applications these restrictions may not apply.

Number of lines. The size of N, or the number of input and output lines, is an important consideration. Numbers depend on the optical architecture and will be given with the their descriptions.

Data word size. In the case of VLSI, pin limitations pose a major constraint on the number of data channels that can be switched in parallel. With optics, there are no pin-out constraints, and in addition in some cases many parallel channels can pass through the same optical switches to provide large bit-parallel transfers.

Unidirectional or Bidirectional Data Transfer. An interconnection network used for inter-processor communication needs data to be transferred in only one direction.

However, a processor-to-memory interconnection network should incorporate bidirectional switching elements. This implies extra control lines in an electronic system. In some optical systems little extra hardware is needed.

Broadcast Capability. If the interconnection network is capable of achieving broadcasts in one step, the overall time-complexity of many parallel algorithms can be reduced. In electronic implementations, the provision of broadcasting involves higher control complexity, a larger number of pins, and/or slower operation. In some optical systems, the addition of broadcast capability involves only very minor increases in complexity.

Propagation Delay. The total delay in the propagation of a signal through the network depends on the number of stages and the delay through each stage. In partitioned large electronic crossbar networks the delay can be excessive and is dependent on the source and destination of the signal. In large active multi-stage networks the delay is also large. In passive optical networks the delay is a function of the physical length of the path and is generally much shorter than in large electronic networks.

Circuit or packet switched. The states of the switches and the routes through the network can be set by a separate control circuit (circuit switching). Or, the data can move through the network in packets with tags on the front of each packet labeling its destination; in this case each switch in the network looks at the tag when it arrives and sets its own state accordingly. Both can be done in electronics, although with large multistage networks pin-out limitations may prevent circuit switching. Passive optical systems are primarily circuit switched; packet switching is difficult. Active optical networks could be circuit or packet switched.

Cascadability. Networks can be used as building blocks in making larger networks. Small crossbars (up to 8x8) are by far the most common building blocks. Pin-out limitations in electronics inhibit cascadability. Optical systems can implement fairly large crossbars (100x100 to 500x500) and in principle can cascade them since there are no pin-out limitations (although this will probably require an active device to regenerate the signal between stages). Such cascadability has not yet been investigated by optical researchers.

Number of stages. A crossbar is the most general network, and has only one stage, but requires N^2 switches. The number of switches (and therefore control signals) can be reduced at some expense of generality by going to a multistage network with O(N log N) switches. This also increases the propagation delay of the network.

Rearrangeability is the ability to realize any permutation of inputs to outputs. This is needed for a general interconnections, and is provided by crossbars and by multi-stage networks with a sufficient number of stages, usually O(log N) stages. However, a rearrangeable network can still be blocking.

Blocking/nonblocking. A rearrangeable network that is blocking may have to re-route some existing connections in order to connect a pair of idle terminals. In a nonblocking network, any pair of idle terminals can be connected without having to re-route any existing connections. This is desirable and represents a major problem with rearrangeable multi-stage networks. Crossbars are completely nonblocking.

5.3.1. Crossbar networks

A schematic diagram of a crossbar is shown in Fig. 9. There are N input lines, N output lines, and N^2 switches. There is no contention and it is completely nonblocking.

Until recently, large crossbar networks were bulky and expensive to construct; the advent of VLSI permits hardware for thousands of switches to be integrated into a single chip. However, the number of pins on a VLSI chip cannot exceed a few hundred and this primarily restricts the size of the largest crossbar that can be integrated into a single VLSI chip. Larger crossbars can be realized only by partitioning them into smaller crossbars, each of which is then implemented using a single chip. Thus, a full crossbar of size $N \times N \times B$ bits can be implemented out of modules of size $n \times n \times b$ bits, and this requires $(N^2/n^2).(B/b)$ modules.

Some implementations of electronic crossbar switching networks for multiprocessors have been reported [52-54]. IBM has constructed a few electronic crossbar switch units as part of experimental special-purpose parallel processor systems for logic simulation. The Los Gatos Logic Simulation Machine(LSM) employs a 64×64 unidirectional crossbar switch to interconnect 64 processing elements among themselves [53]. This is a synchronous system in which the switching sequence is pre-programmed. The system has broadcast capability. The Yorktown Simulation Engine(YSE) [54], a sequel to the Logic Simulation Machine, uses a $256 \times 256 \times 3$-bit wide crossbar switch to interconnect 256 logic processing elements. This switch is constructed out of 256 individual 256×1 multiplexers and has no broadcast capability. The design of a large partitioned electronic crossbar system with special-purpose VLSI chips is discussed in [52]; a system with approximately 500 inputs and outputs is conceived. This is an asynchronous implementation in terms of both control and transfer of data and possesses broadcast capability.

The electronic implementations suffer from a number of limitations including problems associated with wire delays, pin limitations, distribution of clock and control signals, and wiring problems, which are absent in some possible optical implementations. Moreover, optical systems can potentially provide a much higher bandwidth per data channel. The electronic switching systems described in [53,54] consume an enormous amount of space and power. In addition, several delays become significant as the system size increases and a large amount of parallelism is lost because of the physical limitations of the technology.

The remainder of this section (5.2) will discuss optical crossbars [55]. Suppose that N data input lines (each 1 bit wide) are to be connected to N data output lines (each 1 bit wide). The input data at an instant of time can be represented by a vector of length N, as can the output data. The state of the N^2 switches can be described by an N x N matrix, with 1's to represent closed switches (connections) and 0's to represent open switches. A crossbar can then be implemented as a matrix-vector multiply. Crossbars with each line being M bits wide (in parallel) can be implemented by a matrix-matrix multiply with the input and output data represented as N x M matrices instead of N x 1.

Many optical systems have been described for implementing matrix-vector multiplication. An example is a parallel inner product processor whose description follows.

Parallel Matrix-Vector Inner Product Processor. Fig. 10 schematically shows a system capable of performing optical matrix-vector multiplications in what is known as an

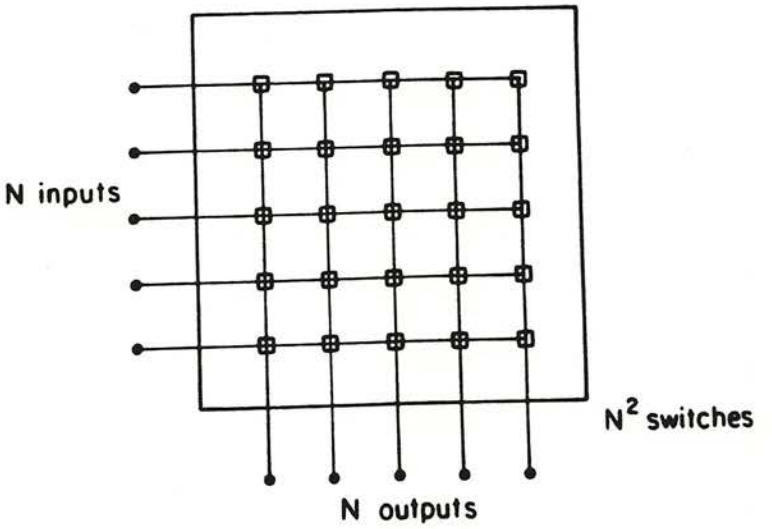

FIg. 9. Crossbar interconnection.

Fig. 10. Optical crossbar implemented with a parallel inner product matrix-vector multiplier. Light is generated by the input vector and travels from left to right. Imaging optics are omitted for clarity.

"inner product" format [56,57]. The N input lines drive an array of N light emitting diodes (LEDs) or laser diodes with a binary signal, so that a binary 1 is represented by light of a fixed intensity, and a binary 0 is represented by a lower (or zero) intensity. An optical system to the right of the input vector spreads the light from each input source into a vertical column that illuminates the crossbar mask shown. The crossbar mask consists of an $N \times N$ array of windows representing the entries in the $N \times N$ permutation matrix. An entry of 0 in the matrix corresponds to zero light transmission, or an opaque window in the mask. An entry of 1 corresponds to full light transmission or an open window. Using a photographic transparency with some pattern of opaque and open windows as the crossbar mask provides a fixed interconnection pattern. For reconfigurable interconnections the crossbar mask could be electrically- or optically- controllable shutter or switch array that is time-variable through the application of external control signals. Several different technologies are available for this purpose, including mechanical, electro-optic, and magneto-optic.

Following the crossbar mask, the next set of optics collects the light transmitted by each row of the mask, and sums the mask output onto a vertical array of N photodetectors corresponding to the N output lines. Thus the system performs a parallel matrix-vector multiplication. There are many possible ways to actually implement this schematic design, some of which use discrete optics or optical waveguides. A version of this optical system using fiber optics with guided-wave splitters and combiners has been experimentally demonstrated as a 4 x 4 crossbar [58]. Two different devices were each used in the system; one yielded switching times of 5 μs, the other 100 ms. The 4 x 4 device array could be addressed in parallel (there were 16 electrical lines coming out). They extrapolate from current experimental results to predict a 16 x 16 could be made, and they estimate that the theoretical upper limit on size would be approximately 500 x 500. Free space versions of Fig. 10 have also been implemented, only as matrix-vector multipliers. One potential drawback is that the system has an optical light efficiency of $1/N$ at most, when used for the usual crossbar operation. This occurs because $(N-1)/N$ of the light from each input source does not pass through the crossbar mask. With current or near future technology a 256 x 256 to 512 x 512 crossbar mask looks viable with a reconfiguration time of 100 μs to 1ms (approximately 1 μs for each switch with N cycles required for complete reconfiguration). Also needed are 1-D parallel source (laser diode) and detector arrays with parallel electronic connections. To the author's knowledge all 1-D detector arrays have serial readout which would be a major bottleneck in implementing such a system for interconnections in an electronic machine. In principle detector arrays with parallel readout could be made. Another possibility is to use optical fibers to route the signals to discrete detectors or to on-chip detectors where the signal is needed.

Acousto-Optic Matrix-Vector Processors. Another class of matrix-vector and matrix-matrix multipliers is based on the technology of acousto-optics. Acousto-optic devices [59] are essentially 1-D light modulators. An electrical signal drives a transducer on one end of the acousto-optic device. This signal propagates down the length of the device as an acoustic wave. The acoustic wave locally alters the optical index of refraction, which can be used to locally modulate a light beam passing through the device. The acoustic (electrical) signal is represented as an amplitude or frequency modulation, with the carrier frequency up to a few GHz.

Many different acousto-optic matrix-vector and matrix-matrix multipliers exist [60]. This includes systolic, engagement, and outer-product processors. Many use a shift-and-add detector array; others use a stationary detector array (preferably with parallel readout) or just a single detector. Compared to the systems that use 2-D optical switch arrays, acousto-optic crossbars are more light efficient, have lower bandwidth, but have shorter reconfiguration time. The limiting factor in terms of bandwidth for most systems is either the detector array or the AO device input, depending on the architecture. With sufficiently fast input and output devices, the AO cell will limit the bandwidth of each line to $\nu_0/2N$, where $\nu_0 \approx 1$ Gb/s. A number of these 1-D elements can be placed in parallel however, permitting each line to be M bits wide, thereby increasing the effective bandwidth by a factor of M ($M \approx 100$ with current devices). Since a new matrix is essentially read in for each new bit or word, the system can be reconfigured rapidly.

5.3.2. Multi-stage networks

Many multi-stage networks have been considered theoretically by the electronics parallel processing community. Nothing yet of any substantial size has been implemented in hardware. In this section we will consider the optical implementation of primarily shuffle/exchange networks. With these rearrangeability can be achieved with approximately $3\log_2 N$ shuffle/exchange stages or with $\log_2 N$ stages followed by $\log_2 N$ inverse shuffle/exchange stages (omega followed by inverse omega network). A diagram of a shuffle/exchange stage is shown in Fig. 7, where the boxes represent switches that perform the exchange operation (they are 2 x 2 crossbars).

Optical Sequential Logic System for Interconnection Networks. Since a computer can be used to implement an interconnection network, we can consider the use of an optical logic system. Here we will consider the implementation of active interconnection networks with the parallel optical sequential logic system of Fig. 1, and will assume that fast 2-D switching arrays as discussed in Sec. 3.1.1 will become available.

The number N of input and output nodes of a network implemented with this system is limited by the complexity of the hologram(s) and the number of gates on the device. If we assume an array of $10^3 \times 10^3$ gates on the device and currently available plotting devices for writing the hologram, we can estimate the maximum N that *one* device and *one* hologram interconnection unit can implement. Here we will assume centralized control with the control signals generated externally. A rearrangeable network consisting of a banyan followed by an inverse banyan, with all 2 log N stages implemented in hardware, can be implemented with $N \approx 12{,}000$ (limited by the number of gates). In this case the data can pass through the network in a pipelined fashion. Implementing one shuffle/exchange stage also permits $N \approx 12{,}000$ (limited by hologram complexity), but the data must pass through the same hardware stage repeatedly to obtain rearrangeability. A non-blocking network can be realized by implementing a Clos network ($N \approx 650$) or a crossbar ($N \approx 500$). It should be noted that these values of N are not really limits because more than one optical system can ultimately be used in conjunction to implement much larger networks.

A few characteristics of such an implementation of interconnection network should be mentioned. First, the optical system of Fig. 1 permits a large number of parallel input and output lines. This allows multiple optical systems to be connected to

implement large interconnection networks without having to worry about careful partitioning of the network. It also allows a separate control line to be input to each switch or exchange box in the case of centralized control. And this system differs from the optical matrix-vector inner product implementation discussed earlier: its data rate is lower and its reconfiguration rate is higher, and the signal level of the data is regenerated as it passes through the network. The potential advantages over electronics in this case arise from the number and length of gate interconnections within each optical interconnection unit, and the number of parallel lines that can run from one optical system to another.

Other shuffle/exchange operations. Implementations of the perfect shuffle were discussed in Sec. 5.2. The exchange stages can be implemented with switchable half-wave plates, as discussed by Lohmann et al [47]. At present, however, such devices require high voltages (100's of volts) to switch, making them impractical for high switching speeds. This situation may improve in the future. Note that this would be a passive interconnection system. Alternatively, the exchange could be implemented using optical gates as described above. In this case the system would be active. In either case with classical optical components used for the perfect shuffle, $N \approx 1000$ could be implemented with 1000 of these in parallel. Then the data could be sent through 30 successive stages to obtain rearrangeability, and 30 rearrangeable networks could be implemented. By adding $O(\log^2 N)$ stages to sort the output, a nonblocking network can be achieved, permitting the implementation of perhaps 1 to a few of these networks with such an optical system.

Multi-stage integrated optic network. A nonblocking multi-stage network has been experimentally demonstrated with limited success [61]. A 4 x 4 was built. It is a 2-D geometry, using electronically controlled optical directional couplers to perform exchange operations, built out of Titanium-diffused $LiNbO_3$. The problems that limit the maximum size were found to be insertion loss of the fibers and crosstalk of the directional couplers. They estimate that with current technology a 6 x 6 could be built. Switching times were fast, $\sim$ 10 ns.

Table 3 summarizes specifications of the reconfigurable optical interconnection networks discussed in this section. Numbers without question marks refer to present or near future estimates. Numbers with question marks are farther in the future and are therefore less reliable.

5.4. Conclusions

We have reviewed optical concepts and systems for the design of fixed and reconfigurable interconnection networks, including crossbar and multi-stage networks. Optical switching systems possess certain desirable characteristics, such as high bandwidth and parallelism, which make them suitable for the design of large crossbar networks, although electronic switching speeds are not achievable with the existing optical technology. Alternatively, an optical sequential logic system could be configured as a crossbar or multistage switching network to permit optically controlled and packet switched networks.

Table 3. Some estimated specifications of reconfigurable optical interconnection networks. (a) Matrix-vector inner product processor (free space). (b) Acousto-optic matrix-vector processor. (c) Acousto-optic crossbar (not a matrix-vector multiplier; this system was not described in this report; it has been presented recently but is as yet unpublished). (d) Use of the parallel optical sequential logic system of Fig. 1 for implementation of a Clos network. A Clos network is built out of smaller crossbars, and is in this case completely nonblocking. (e) Shuffle/exchange implementation for multi-stage networks. (f) Guided-wave version of the 3-D matrix-vector inner product processor. (g) 2-D integrated optic multi-stage network (nonblocking). Abbreviations: matrix-vector (M-V); spatial light modulator (SLM).

	Crossbar:	N=no. lines	Bandwidth	Recon-figuration Time	Broad-cast ?
Free space	(a) M-V with 2-D SLM	256 - ?	> 1 GHz*	100 μs-1ms	yes
	(b) M-V with AO	30 - 1,000	1 $GHz/_{2N}$	2N x 1ns.	yes
	(c) Other AO	1,000	>1 GHz*	1 μs	maybe
	Clos:				
	(d) D.O.C. (USC System)	up to 600	100 MHz?	10ns? †	yes
	(e) Multi-stage: (eg. shuffle/exchange)				
	D.O.C (USC System)	10^4	100 MHz?	10ns? †	
	Conventional Optics	N=1,000 ? 1,000 in parallel ?	>1GHz*	?	
Fiber					
	(f) Crossbar (3-D)	16 - 500 ?	1GHz*	5 μs - 100ms	yes
	(g) Multi-Stage (2-D)	4 - ?	1GHz*	10ns each switch	yes

* Passive line - Bandwidth limited only by detectors and sources

† Does not include calculation of routes

6. SUMMARY AND CONCLUSIONS

In summary, an overview of research in digital optical computing systems and optical interconnections has been given. Architectures, algorithms and applications, required components and requisite materials and devices have been discussed. The present state of the technology has been given, and comparisons have been made to other technologies where possible.

The conclusion is that digital optical computing is in a very early stage of development but is definitely worth pursuing because of potential benefits. Significant progress needs to be made for it to become viable, but recent progress and current research make it look promising. Digital optical computing compares favorably with projected alternative technologies in many application areas. Optical interconnections in electronic computers represent a compromise between DOC and digital electronic computing, and provide hope of solving many of the problems associated with electronic interconnections that are becoming more limiting as the technology progresses. Their role has already begun on a very high level and will increase in use to lower levels and probably to reconfigurable networks as well.

REFERENCES

1. A.A. Sawchuk and T.C. Strand, "Digital Optical Computing", *Proc. IEEE,* Vol. 72, p. 758, 1984

2. C.A. Mead and L.A. Conway, (eds.), *Introduction to VLSI Systems,* Addison-Wesley, Reading, Ma., Ch. 8, 1980.

3. L.S. Haynes, R.L. Lau, D.P. Siewiorek, and D.W. Mizell, "A Survey of Highly Parallel Computing", *Computer,* Vol. 15, no. 1, pp. 9-24, January 1982.

4. J.W. Tompkins and S. Hollock, "The Impact of Multi-Level Metalisation on Semi-Custom LSI", *Proc. 1982 Custom Integrated Circuits Conf.,* pp. 276-280, 1982.

5. C.L. Seitz, "Systems Timing", Chp. 7, in Ref. 2.

6. H.T. Kung, "Why Systolic Architectures", *Computer,* Vol. 15, no. 1, pp. 37-46, January 1982.

7. C.L. Seitz, "Ensemble Architectures for VLSI - A Survey and Taxonomy", *Proc. 1982 Conf. on Adv. Research in VLSI,* M.I.T., pp. 130-135, 1982.

8. H.T. Kung and C.E. Leiserson, "Algorithms for VLSI Processor Arrays", in Ref. 2.

9. S.Y. Kung, Private Communication.

10. A. D. Fisher, "A Review of Spatial Light Modulators," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper TuC1.

11. H.M. Gibbs, S.L. McCall, and T.N.C. Venkatesan, "Optical Bistable Devices: The Basic Components of All-Optical Systems?", *Optical Engineering,* Vol. 19, p. 463, 1980; D.A.B. Miller, "Bistable Optical Devices: Physics and Operating Characteristics", *Laser Focus,* Vol. 18, no. 4, p. 79, 1982.

12 K. Hwang and F. A. Briggs, *Computer Architecture and Parallel Processing,* (McGraw-Hill, New York, 1984).

13. T. Kohonen, *Self-Organization and Associative Memory,* (Springer-Verlag, New York, 1984).

14. J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA,* Vol. 79, pp. 2554-2558, 1982.

15. S. E. Fahlmann et al., "Massively Parallel Architectures for AI: NETL, Thistle, and Boltzmann Machines," *Proc. AAAI,* pp. 109-113, 1983.

16. S. H. Lee, "Nonlinear Optical Processing," in S. H. Lee, ed., *Optical Information Processing: Fundamentals* (Springer-Verlag, New York, 1981), Ch. 7.

17. H. Barr and S. H. Lee, "A Digital Optical Processing System," *Proc. Tenth International Optical Computing Conference,* IEEE Cat. No. 83CH1880-4, IEEE Computer Society, Silver Spring, Md., pp.171-177, 1983.

18. D. Psaltis and N. Farhat, "Optical information processing based on an associative-memory model of neural nets with thresholding and feedback," *Optics Letters,* Vol. 10, pp. 98-100, 1985.

19. A. D. Fisher, C. L. Giles, and John N. Lee, "An adaptive, associative, optical computing element," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper WB4.

20. K.-H. Brenner and A. Huang, "An optical processor based on symbolic substitution," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper WA4.

21. A. Huang, "Why use the parallelism of optics?" *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper WA2.

22. B.K. Jenkins et al., "Architectural Implications of a Digital Optical Processor", *Appl. Opt.,* Vol. 23, no. 19, pp. 3465-3474, 1984.

23. B.K. Jenkins, et al., Sequential Optical Logic Implementation", *Appl. Opt.,* Vol. 23, no. 19, pp. 3455-3464, 1984.

24. B. K. Jenkins and A. A. Sawchuk, "Optical cellular logic processors for image processing," *Proc. Computer Architecture for Pattern Analysis and Image Database Management,* Miami, November 1985, IEEE Computer Society, Siver Spring, Md. to appear, 1985.

25. J.L. Jewell, M.C. Rushford, and H.M. Gibbs, "Use of a Single Nonlinear Fabry-Perot Etalon as Optical Logic Gates", *Appl. Phys. Lett.,* Vol. 44, p. 172, 1984; J.L. Jewell, M.C. Rushford, H.M. Gibbs, and N. Peyghambarian, "Single-Etalon Optical Logic Gates", in *Technical Digest Conference on Lasers and Electrooptics,* Optical Society of America, Washington, D.C., paper THg2, 1984.

26. D. A. B. Miller, et al. "Novel hybrid optically bistable switch: the quantum well self-electro-optic effect device," *Appl. Phys. Lett.,* Vol. 45, no. 1, pp. 13-15, 1985.

27. J. L. Jewell and Y. H. Lee, "Parallel operation of GaAs bistable etalons: with and without crosstalk," *Technical Digest, 1985 Annual Meeting,* Optical Society of America, Washington, D.C., paper FU4, p. 111, 1985.

28. W. F. Sharfin and M. Dagenais, "Femtojoule optical switching in InGaAsP laser amplifiers," *Technical Digest, 1985 Annual Meeting,* Optical Society of America, Washington, D.C., paper FU3, p. 111, 1985.

29. Y. H. Lee, et al., "200-ps recovery of an optical gate in a GaAs etalon array," *Technical Digest, 1985 Annual Meeting,* Optical Society of America, Washington, D.C., paper FU5, p. 111, 1985.

30 P. W. Smith, "On the fundamental limits of digital optical switching and logic elements," *Bell Syst. Tech. J.,* Vol. 761, p. 1975, 1982.

31. C. C. Guest and T. K. Gaylord, "Truth-table look-up optical processing utilizing binary and residue arithmetic," *Applied Optics,* Vol. 19, pp. 1201-1207, 1980.

32. C. C. Guest, et al., "EXCLUSIVE OR processing (binary image subtraction) using thick Fourier holograms'" *Applied Optics,* vol. 23, pp.3444-3454, 1984.

33. A. Armand, et al., "Real-time parallel optical analog-to-digital conversion," *Optics Letters,* Vol. 5, pp. 129-131, 1980.

34. R. A. Becker, et al., "Wide-band electrooptic guided-wave analog-to-digital converters," *Proc. IEEE,* Vol. 72, pp. 802-819, 1985.

35. A. W. Lohmann, "Optical logic gates based on polarization," *Technical Digest, 1985 Annual Meeting,* Optical Society of America, Washington, D.C., paper TuY4, p. 111, 1985.

36. A. R. Tanguay, Jr., "Materials requirements for optical processing and computing devices," *Optical Engineering,* vol. 24, pp. 2-18, 1985.

37. R. W. Keyes, "Optical logic - in the light of computer technology," *Optica Acta,* vol. 32, pp. 525-536, 1985.

38. P. W. Smith, "Applications of all-optical switching and logic," *Phil. Trans. R. Soc. Lond. A,* Vol. 313, pp. 349-355, 1984.

39 J. W. Goodman, et al., "Optical interconnections for VLSI systems," *Proc. IEEE,* Vol. 72, pp. 850-866, 1985.

40 T.R. Kushner and A. Rosenfeld, "Interprocessor Communication Requirements for Parallel Image Processing," *Proc. Computer Architecture for Pattern Analysis and Image Database Management,* IEEE Cat No. 83CH1929-9, pp. 177-183, 1983.

41 K. H. Wagner and R. T. Weverka, "Space-integrating acousto-optics matrix-matrix multipliers," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper TuD5.

42 R. A. Athale and J. N. Lee, "Architectural and hardware issues in high accuracy optical matrix processors," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper TuD6.

43 T. J. Bicknell, et al., "Integrated-optical synthetic aperture radar processor," *Technical Digest, 1985 Annual Meeting,* Optical Society of America, Washington,

D.C., paper TuE6, p. 18, 1985.

44 R. W. Keyes, "Fundamental limits in digital information processing," *Proc. IEEE,* Vol. 69, pp. 267-278, 1985.

45 R. K. Kostuk, et al., "Optical imaging applied to microelectronic chip-to-chip interconnections," *Applied Optics,* Vol. 24, pp. 2851-2858, 1985.

46. A. Huang, "digital techniques in optical computing," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper TuA6.

47. A. Lohmann, et al., "Optical implementation of the perfect shuffle," *Technical Digest, Topical Meeting on Optical Computing,* Incline Village, Nevada, Optical Society of America, Washington, D.C., 1985, paper WA3.

48. B. K. Jenkins, "Recent developments in digital optical computing," *Technical Digest, 1985 Annual Meeting,* Optical Society of America, Washington, D.C., paper TuT4, p. 32, 1985.

49. A. Husain, "Optical interconnection of digital circuits and systems," *Proc. Soc. Photo-Opt. Instr. Eng.,* Vol. 466, 1984.

50 M. E. Kim, et al., "GaAs/GaAlAs integrated optoelectronics for optical interconnect applications," *Proc. Soc. Photo-Opt. Instr. Eng.,* Vol. 466, 1984.

51. H. J. Siegel, "Interconnection Networks for SIMD Machines," *Computer,* Vol. 12, No. 6, June 1979, pp 57-65.

52. G. Broomell, J. R. Heath, "An Integrated-Circuit Crossbar Switching System," *Proceedings of the 4th International Conference on Distributed Computing Systems,*" May 1984, pp 278-287.

53. T. Burggraff, et al., "The IBM Los Gatos Logic Simulation Machine Hardware," *Proceedings of the International Conference on Computer Design,* 1983, pp 584-587.

54 M. M. Denneau, "The Yorktown Simulation Engine," *Proceedings of the 19th Design Automation Conference, Las Vegas,* 1982, pp 55-59.

55. A. A. Sawchuk, et al., "Optical Interconnection Networks," *International Conference on Parallel Processing,* St. Charles, Illinois, IEEE Cat. No. 85CH2140-2, 1985, pp. 388-392.

56. J. W. Goodman, et al., "Fully Parallel High-Speed Incoherent Optical Method for Performing Discrete Fourier Transforms," *Opt. Lett.,* Vol. 2, 1978, pp 1-3.

57. J. W. Goodman, et al., "Some New Methods for Processing Electronics Image Data Using Incoherent Light," *Proc. of the International Commission for Optics - 11, Madrid, Spain,* 1978, pp 138-145.

58. A. Himeno and M. Kobayashi, "4 x 4 optical-gate matrix switch'" *J. of Lightwave Technology,* Vol. LT-3, pp.230-235, 1985.

59. N. J. Berg, J. N. Lee, *eds., Acousto-Optic Signal Processing,* Marcel Dekker, New York, 1983.

60. R. A. Athale, "Optical Matrix Algebraic Processors: A Survey," *Proc. Tenth Intl. Optical Computing Conf.,* IEEE Cat. No. 83CH1880-4, pp 24-31, 1983.

61. H. S. Hinton, "A nonblocking optical interconnection network using directional couplers," *Proc. IEEE Global Telecommunications Conference,* IEEE Cat. No. 84CH2064-4, pp. 885-889, 1984.