a reprint from Applied Optics

Reconfigurable intelligent optical backplane for parallel computing and communications

Ted H. Szymanski and H. Scott Hinton

A reconfigurable intelligent optical backplane architecture for parallel computing and communications is described. The backplane consists of a large number of reconfigurable optical channels organized in a ring with relatively simple point-to-point optical interconnections between neighboring smart-pixel arrays. The intelligent backplane can implement (1) dynamically reconfigurable connections between any printed circuit boards, (2) dynamic embeddings of classical interconnection networks such as buses, rings, multidimensional meshes, hypercubes, shuffles, and crossbars, (3) multipoint switching, (4) sorting, (5) parallel-prefix operations, (6) pattern-matching operations, (7) snoopy caches and intelligent memory systems, and (8) media-access control functions. The smart-pixel arrays can be enhanced to include more complex functions, such as queuing and routing, as the technologies mature. Descriptions of the architecture and the smart-pixel arrays and discussions of the system cost, availability, and performance are included.

Key words: Free space, backplane, smart pixels, reconfigurable, dynamic, programmable, intelligent.

1. Introduction

Over the past few decades, advances in technology have had a continual impact on systems architectures by reducing size and increasing performance by roughly a factor of 2 every year. If this trend continues, the supercomputers of tomorrow will fit within the volume currently occupied by a single shelf found in existing electronic cabinets. Based on the evolutionary trends of the past, this hardware compression implies that there will be more transistors per integrated circuit (IC), more IC's per printed circuit board (PCB), more PCB's per shelf, and finally more connections between the PCB's. In addition to this decrease in system volume, system performance will continue to increase. Thus today's connection-constrained supercomputers and telecommunications systems that occupy multiple cabinets of electronics and require bisection bandwidths in the tens of gigabits per second (Gb/s) will evolve into single-shelf systems supported by backplanes capable of terabits per second (Tb/s) bisection bandwidths (the bisection bandwidth can be defined as the bandwidth that crosses a bisector that splits the architecture into two halves of equal size).

Although there has been a great deal of progress in the development of electrical backplanes, they are ultimately constrained by the fundamental physical limitations of electronics.[1,2] Current metal interconnects are limited by the skin effect, which results in a greater attenuation in transmission lines at high frequencies, and parasitic inductance and capacitance, which reduce the usable bandwidth of the interconnections, typically to the hundreds of megabits per second (Mb/s). The power dissipation required per pin-out for electrical interconnects will limit the number of pin-outs that can be realized within one package. Finally, electronic backplanes will always be constrained by the two-dimensional (2D) planar nature of electronic traces on a PCB substrate; all the interconnections must be implemented by relatively large metal traces typically hundreds of micrometers wide in a 2D PCB substrate. Optical interconnects provide a massive degree of interconnection by exploiting three-dimensional (3D) free space; beams focused to spots tens of micrometers wide and spaced hundreds of micrometers apart can be routed through 3D free space at very high

T. H. Szymanski is with the Department of Electrical Engineering, McGill University, 3480 University Street, Montreal PQ H3A 2A7, Canada; H. S. Hinton is with the Department of Electrical and Computer Engineering, University of Colorado at Boulder, Boulder, Colorado 80301-0425.

Received 16 June 1995; revised manuscript received 7 November 1995.

0003-6935/96/081253-16$06.00/0

clock rates, with low energy per bit and with no electromagnetic interference.

One approach to overcome these interconnection limitations of electrical backplanes is to exploit the temporal and spatial bandwidth available with free-space optical technology. As shown in Fig. 1, a free-space optical backplane is composed of a large number of optical communication channels (OCC's) created by optically interconnecting smart-pixel arrays (SPA's). These SPA's are optoelectronic devices that consist of optical inputs and outputs (I/O's) and electronic processing circuitry. Signals to be transferred between PCB's are injected into the OCC's by means of the SPA's and then transferred to the destination PCB's, where the SPA's extract the optical signals and convert them back into electrical form. The potential connectivity of these optical backplanes includes the ability to provide over 10,000 high-performance connections per PCB while supporting bisection bandwidths in excess of 1 Tb/s. Perhaps more important than this raw connectivity is the ability for intelligence to be added to each of the OCC's through the electronic circuitry in the SPA's. It is this intelligence and its effective use that constitute what is referred to below as intelligent optical backplanes.

The unique ability to transport and process vast amounts of data per second may have an impact on future architectures by enabling new paradigms for computing and communications. Potential applications include terabit point-to-point and multipoint optical asynchronous transfer mode switching architectures, terabit shared memory and message-passing parallel computing architectures, terabit data-flow computing architectures, terabit intelligent memory systems, and terabit parallel database architectures. This paper begins with a review of free-space optical interconnection technologies, followed by a description of the HyperPlane intelligent optical backplane architecture and its associated SPA's. The following sections describe the embedding of a Cray Research supercomputer interconnect, the embedding of a 1024-channel broadcast-based multichanel switch, advanced applications for intelligent backplanes, and system cost, availability, and performance issues.

2. Review of Free-Space Optical Technology

This section provides a brief review of four smart-pixel technologies and two optical imaging technologies. A key point of the proposed intelligent backplane architecture is that it requires relatively simple point-to-point optical interconnections among neighboring SPA's without requiring any complex optical operations, such as one-dimensional (1D) or 2D perfect shuffles. The proposed architecture can use any of the following smart-pixel and imaging technologies, i.e., the architecture is not constrained to any particular technology. Readers familiar with these technologies may proceed directly to Section 3, the presentation of one version of the proposed backplane architecture.

A. Smart-Pixel Technologies

As the key component of an intelligent optical backplane, the SPA is an optoelectronic chip packaged and mounted on a PCB. SPA's communicate with the IC's on the PCB through electrical channels while communicating with the SPA's located on other PCBs through the 2D arrays of parallel free-space optical channels. The potential advantages of free-space optical interconnect technology include (1) high temporal bandwidth interconnects, (2) low skew between channels, (3) lower system power dissipation, (4) low cross talk between channels, (5) larger number of pin-outs per chip or board, (6) parallel structures that potentially reduce latency, and (7) the ability of a small electronic aggregate capacity to control a large optical aggregate capacity. Figure 2 illustrates the SPA aggregate capacity as a function of connectivity (pin-outs per chip) and per channel data rate (bits per second). The shaded upper right corner of the figure is the high-performance region that supports greater than a terabit aggregate capacity that can be accessed through smart pixels and free-space interconnection.

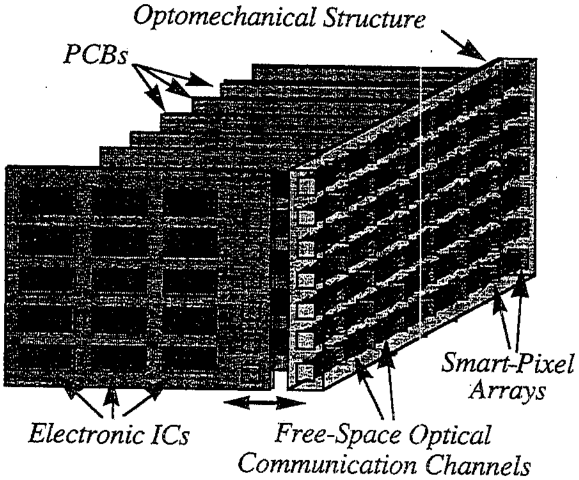


Fig. 1. Free-space optical backplane.

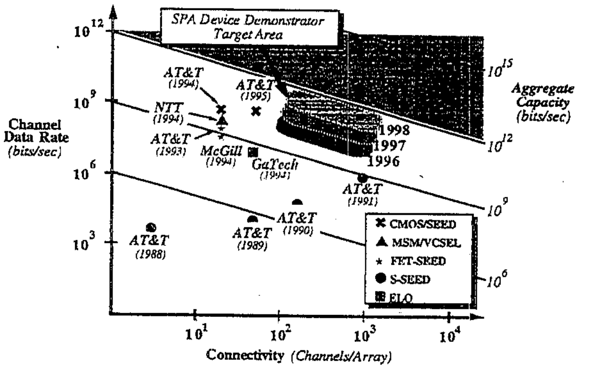


Fig. 2. Projected aggregate capacity per SPA. CMOS, complementary metal-oxide semiconductor; SEED, self-electro-optic-effect device; MSM, metal–semiconductor–metal; VCSEL, vertical-cavity surface-emitting laser; FET, field-effect transistor; S-SEED, symmetric SEED; ELO, epitaxial lift-off.

There are four SPA technologies that have been developed over the past 5 years; they include (1) the field-effect transistor (FET) self-electro-optic-effect-device (SEED) technology,[3] (2) the epitaxial lift-off (ELO) technology,[4] (3) the Hybrid-SEED technology,[5] and (4) the vertical-cavity surface-emitting laser (VCSEL) metal–semiconductor–metal (MSM) technology.[6] All these technologies use fairly standard semiconductor processing and are grown at the wafer scale. The SPA's are packaged in the same manner as VLSI IC chips and can be included on a PCB, a multichip module (MCM), or any other type of IC mounting method. Below is a brief description of each of these technologies.

FET-SEED smart pixels: FET-SEED smart pixels combine metal–semiconductor FET's (MESFET's) and SEED's by processing MESFET's on a GaAs substrate that contains the SEED multiple-quantum-well (MQW) structure. This process is truly monolithic, involving only a single substrate in creating both electronic and optic functions, and is an evolution of the symmetric-SEED smart pixels used earlier and referenced in Fig. 2. The SEED acts as a high-efficiency optical detector and an optical modulator. In use, optical power from an off-chip laser is modulated by SEED's in smart-pixel transmitters and then detected by SEED's in smart-pixel receivers. These devices excel as optical modulators and detectors because the substrate is grown specifically for SEED functionality. Optical detection–modulation and electronic processing rates are very fast (>1 GHz). One drawback is that FET-SEED circuitry is not yet optimized for dense, low-power designs.

ELO Smart Pixels: Leveraging on the current state of complementary metal-oxide semiconductor (CMOS) technology, ELO smart pixels have thin-film (devices removed from their growth substrate) optical devices placed into special wells left in a chip processed through a CMOS foundry. CMOS circuitry allows for the design of dense, low-power, high-speed smart-pixel systems. ELO processes electrical devices on a Si substrate through standard VLSI techniques and processes optical devices on another substrate well suited for optical detection, modulation, or emitting. The optical devices are lifted from their substrate and attached to the CMOS circuitry.

Hybrid-SEED Smart Pixels: The Hybrid-SEED smart pixel combines Si CMOS circuitry with GaAs SEED devices by flip-chip bonding. The CMOS circuitry contains a small (~15 μm) bonding pad at each location where an electrical contact is needed to a SEED device for optical detection or modulation. A GaAs substrate that contains an array of SEED's (with contacts positioned according to the CMOS contacts) is flip-chip bonded to the Si CMOS substrate. A final step removes the GaAs substrate, leaving behind an array of isolated SEED's each electrically attached to the CMOS circuitry. This technology is promising; it can currently integrate dense CMOS (400,000 transistors/cm^2) with dense optoelectronics (28,000 SEED's/cm^2) on a chip operating at rates greater than 250 Mb/s.

VCSEL/MSM Smart Pixels: The VCSEL/MSM smart pixel is the integration of VCSEL's and MSM detectors along with electronics. This smart-pixel technology uses optical sources instead of modulators as the output devices in each smart pixel. This removes the need for an off-chip laser and the associated optical components required for bringing its power onto the modulators. As the power required for driving VCSEL's is reduced, this will become an attractive smart-pixel technology.

At the current time both the ELO and the hybrid-SEED smart-pixel technologies are available for use in experimental optoelectronic systems through the U.S. Advanced Research Project Agency Consortium on Optoelectronic Technologies on smart pixels. The proposed backplane architecture has been designed into four hybrid-SEED SPA's and the devices are currently being fabricated through the U.S. Advanced Research Project Agency/AT&T/Consortium on Optoelectronic Technologies workshop. A demonstration of the architecture is planned for 1996.

B. Free-Space Optical Interconnects

Creating the optical connection channels between SPA's in a rugged package for a real-world environment is one of the issues central to realizing an optical backplane. The optical system in the backplane has one main function: to transfer optical energy from an array of smart-pixel modulators or emitters through free space and onto an array of smart-pixel detectors. This function can be considered a simple one-to-one imaging of an optical source plane onto an optical detector plane.

Figure 3 shows examples of two major approaches to the free-space optical interconnection of SPA's. The first method is based on the use of bulk optical imaging systems, which has been the preferred

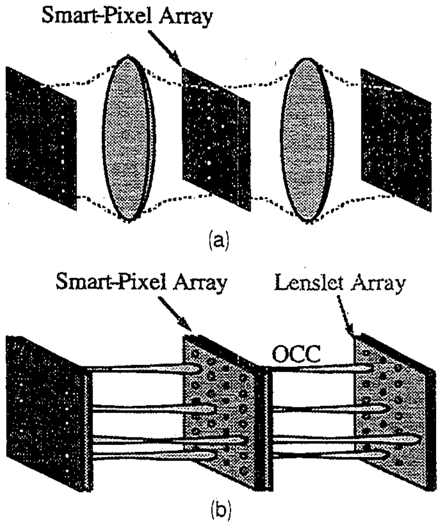


Fig. 3. Methods of point-to-point optical interconnection: (a) bulk optical imaging, (b) lenslet-based microchannels.

approach to this point in time. An example is the AT&T *System*$_4$ architecture, which imaged 32 × 32 optical channels between six SPA chips.[7] This method requires high-quality lenses for capturing and accurately transferring the entire SPA field of optical channels. Despite the tremendous amount of progress in developing these systems, they tend to be both bulky and expensive. A second approach uses small lenslets to create small optical microchannels between the SPA's. Each lens is responsible for imaging only a single smart pixel (or a small cluster of smart pixels[8]) through a dedicated micro-optic system. Currently both diffractive and refractive lenslet arrays are available. Although still in the early stages of development, the lenslet approach offers the hope of small and inexpensive systems.

In order to combine optics into an electronic computing system physically, this technology needs a compact and durable optical system that is compatible with current computer packaging. The push to decrease optical system size has dramatically changed the methods for system packaging. Previous free-space digital optics systems typically used nearly all 32 ft^2 (3 m^2) of an optical table. The large system size was due to the large size of the optical components and their mounting mechanisms. With the shrinking of optical components and the availability of high-power laser diodes, new optical system packaging methods that are compatible with the physical conventions of current electronic systems are evolving.

3. HyperPlane Architecture

The HyperPlane shown in Fig. 4 is a reconfigurable[12] multichannel optical backplane architecture that has the ability to embed classic networks dynamically.[9–11] Each PCB or MCM board accesses the Z optical backplane channels through I electronic injector channels and E electronic extractor channels, where $I \leq Z$ and $E \leq Z$ typically. The injectors provide the capability of injecting electronic signals into a selected subset of optical channels while the extractors are used to extract information from another subset of optical channels. With existing IC I/O technology, each VLSI die has at most tens of gigabits of electrical I/O bandwidth because of electronic constraints such as packaging, power, and clock-rate limitations.[1,2] However, each VLSI die may potentially have hundreds of gigabits of optical I/O bandwidth. The smart-pixel designs for the HyperPlane allow a VLSI smart-pixel die with tens of gigabits of electrical I/O bandwidth the capability of tapping an optical interconnect with hundreds of gigabits of optical bandwidth.

There are three fundamental methods for implementing an optical HyperPlane: time-division multiplexing, wavelength-division multiplexing, and space-division multiplexing. Combinations of these three basic approaches could also be used. In a time-division-multiplexed HyperPlane, a specific time segment TS$_i$ is equivalent to a single communication channel C_i. In a wavelength division-multiplexed HyperPlane, a specific wavelength λ_i is associated with each communication channel C_i. In a space-division multiplexed HyperPlane, a specific spatial location (x_i, y_i) is associated with each communication channel C_i. Space-division multiplexing exploits the large 3D spatial bandwidth made available through free space.

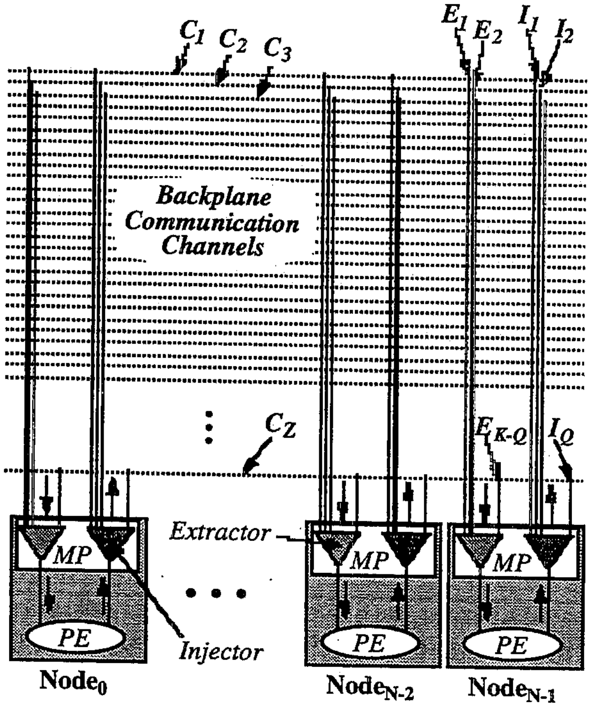
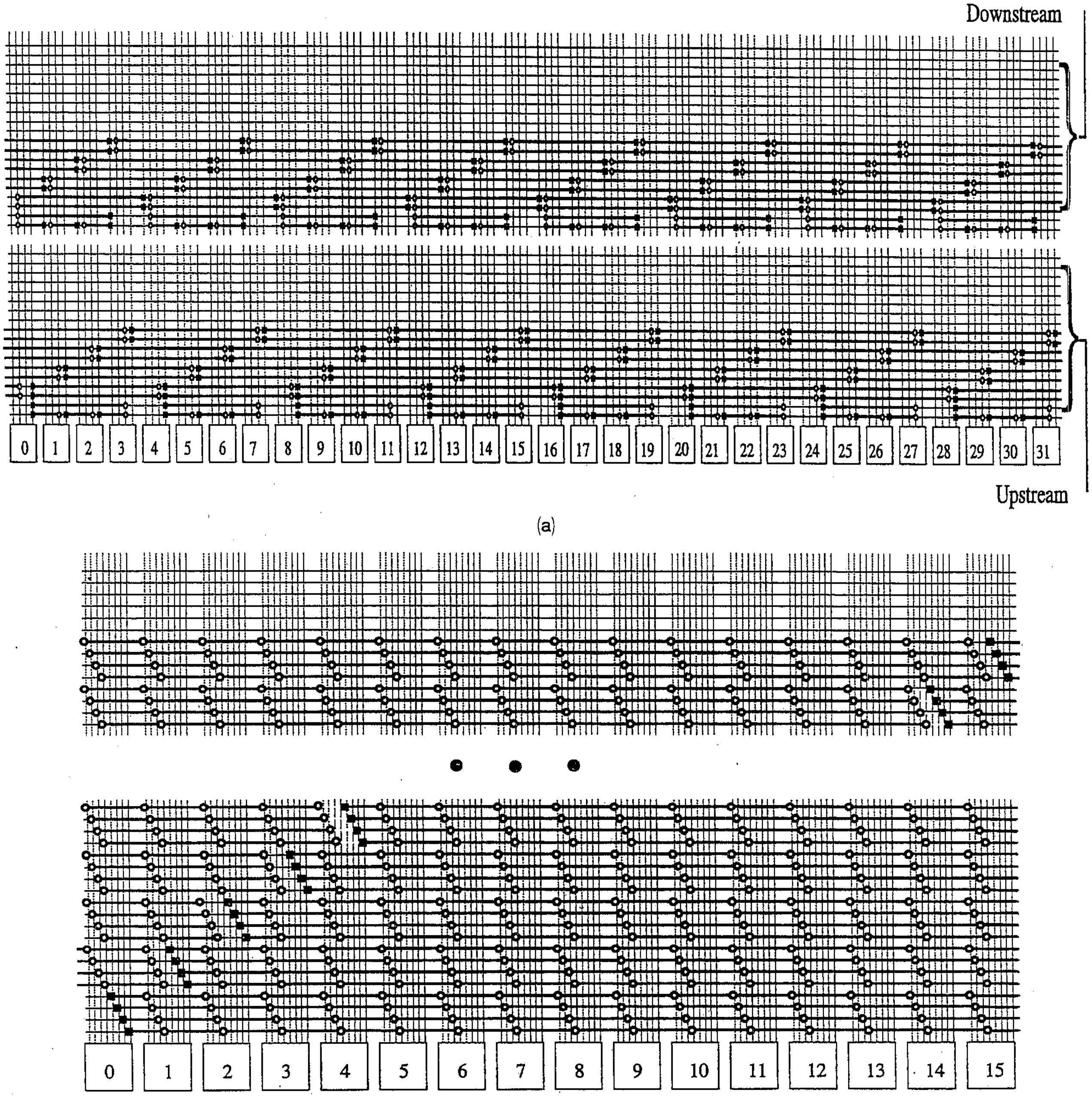


Fig. 4. HyperPlane connectivity model. MP's, message processors; PE's, processing elements.

In a circular HyperPlane, the optical channels are organized in a bidirectional ring. A key feature of the architecture is its ability to embed conventional interconnection networks used in computing and communications. The embedding of the 3D mesh interconnect found in the Cray Research T3D supercomputer[13] is shown in Fig. 5(a), and the embedding of a multiple bus switch is shown in Fig. 5(b). (These embeddings are described in Subsections 3.E. and 3.F.) These embedding templates use a box to denote each backplane PCB (or MCM). Each PCB has a number of vertical lines that represent the electrical injector and extractor channels to or from the PCB. The template has a large number of horizontal lines that, in this paper, represent logical optical channels that typically consist of 8 optical bits. In this paper the word channel refers to a bytewide logical optical channel. The horizontal lines in the 2D template denote optical channels without specifying their precise physical location in 3D free space, as it is difficult to convey 3D information pictorially. Photonics provides a large 3D spatial bandwidth, which the 2D embedding template does not fully reflect.



(a)

(b)

Fig. 5. (a) Embedding of a 3D 8 × 8 × 8 mesh for a Cray Research T3D supercomputer into a bidirectional circular HyperPlane. (Each bold line represents 24 bytewide channels, and the complete embedding uses 240 channels in each direction.) (b) Embedding of a multichannel broadcast switch into a unidirectional circular HyperPlane. (Each bold line represents four bytewide channels, and each PCB has access to a contention-free 128-bit-wide reconfigurable bus.)

The solid boxes, circles, and bold lines in Fig. 5 represent the connections in the circular HyperPlane. The solid boxes represent data-injection points, the circles represent data filtering and extraction points, and the bold lines represent established point-to-point or multipoint optical connections. The K vertical access channels emanating from a PCB represent the PCB degree, i.e., the maximum number of logical optical channels that can be accessed by a PCB. When a connection between PCB's is established, a bold horizontal line is drawn between the optical channel end points.

A. Smart-Pixel Technology Constraints

First the constraints of the AT&T Hybrid-SEED technology are reviewed, followed by a description of the HyperPlane SPA's that fit within these technology constraints. In AT&T's Hybrid-SEED technology an array of GaAs MQW optical I/O modulators are flip-chip bonded onto a Si substrate that contains the logic.[5] A 32 × 32 array of smart pixels with 1024 optical I/O bits per cm^2 is currently considered a realistic target area given current optical power limitations.

VLSI Layout: With existing bonding techniques, electronic I/O bonding pads are placed around the perimeter of a VLSI die. Each bonding pad requires an area of ≈160 μm × 80 μm, yielding ≈500 I/O pads per cm^2. Assuming that ≈100 pins are allocated to power and ground, ≈400 I/O pins are available for data and control. When a 32 × 32 array of pixels is placed into the remaining VLSI area, each pixel has an area of ≈300 μm × 300 μm. After 50 μm × 200 μm is allocated for interfacing to the optical I/O, each pixel has ≈280 μm × 280 μm for logic.

Transistor Density: With a conservative transistor density of 4000 transistors per square millimeter (achievable with CMOS), each pixel can hold ≈300 transistors. Assuming three transistors per two-input logic gate, each pixel can hold ≈100 logic gates. Densities of 20,000 transistors per square millimeter are achievable with existing CMOS technologies, yielding a density of ≈1500 transistors per pixel or ≈500 two-input logic gates per pixel. When switch logic (i.e., pass transistor logic) is used, the gate density can be increased by ≈50%–100%. In the remainder of this paper we assume a complexity of 50–100 gates per pixel, which is readily achievable with the 0.8-μm CMOS process available through the metal-oxide-silicon implantation system (MOSIS).

Clock Rates: Optical clock rates of 500–1000 Mb/s have been demonstrated with the Hybrid-SEED technology and are considered realistic. External electronic clock rates are limited by the slew rates of the electronic bonding pad drivers and are typically ≤500 Mb/s for CMOS processes.[1,5]

Summary: With the Hybrid-SEED and 0.8-μm CMOS technologies, a feasible SPA on a square centimeter die would offer a 32 × 32 array of pixels with 1024 optical I/O's, 500 electronic I/O's, an electronic clock rate of <500 Mb/s, and an optical clock rate of 500–1000 Mb/s. Assuming that the electronic and the optical clock rates of the SPA are both 250 Mb/s, the aggregate optical and electrical I/O bandwidths of each SPA are 256 and 48 Gb/s, respectively, and the ratio of optical to electrical bandwidth for each SPA is 5.3 to 1.

B. Basic Smart-Pixel-Array Design

In this section the basic design of a space division multiplexed SPA that supports multiple reconfigurable channels is described. The organization of a basic smart pixel is shown in Fig. 6. Each pixel has four basic states, the transparent, transmitting, receiving, and transmitting-and-receiving states, as shown in Fig. 7. The smart pixels are organized into a 2D array called a slice, as shown in Fig. 8.

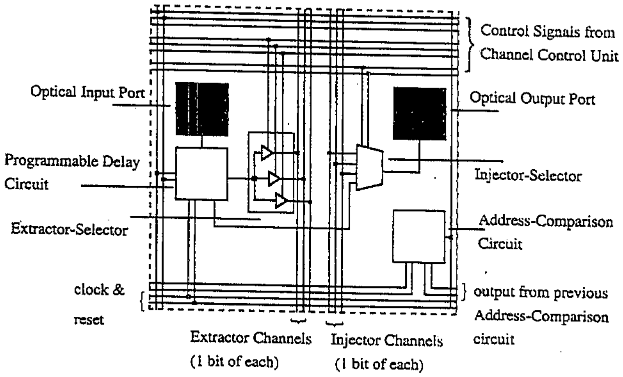


Fig. 6. Basic HyperPlane smart pixel.

Each slice is a self-contained module with optical I/O and electronic I/O: the slice implements the interface between *C* optical channels (each *w* = 8 bits wide in this example), *I* electronic injector channels, and *E* electronic extractor channels. The VLSI die may contain *S* independent slices, where *S* ≥ 1. The SPA's may have different ratios of electrical to optical I/O bandwidth by adjustment of the parameters *S*, *C*, *I*, *E*, and *w*. Figure 9 illustrates SPA's with various ratios of optical to electrical bandwidths. Different designs may be chosen to match the electronic I/O bandwidths of various IC packages.

Each slice typically includes an injector-selector (i.e., expander) switching circuit for switching *I* injector channels onto a subset of *C* ≥ *I* optical channels and an extractor-selector (i.e., concentrator) switching circuit for switching a subset of *E* channels selected from the *C* optical channels onto *E* extractor channels. (Expanders and concentrators are classic components of computing and communication systems.)

The backplane can be operated in two general modes, a reconfigurable mode and an intelligent mode. In the reconfigurable mode, the backplane can be reconfigured to embed any type of graph subject to the constraints on the number of electrical

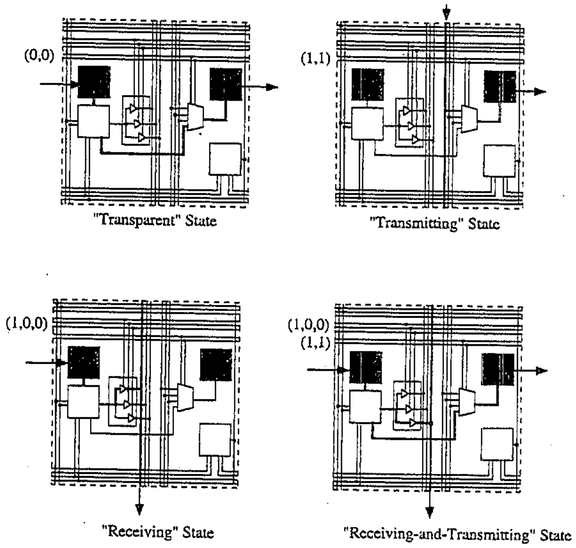


Fig. 7. Four basic states of a HyperPlane smart pixel.

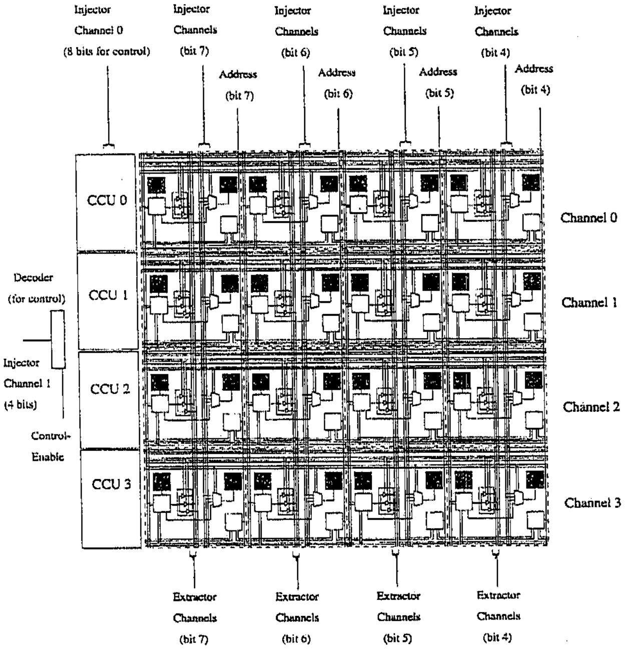
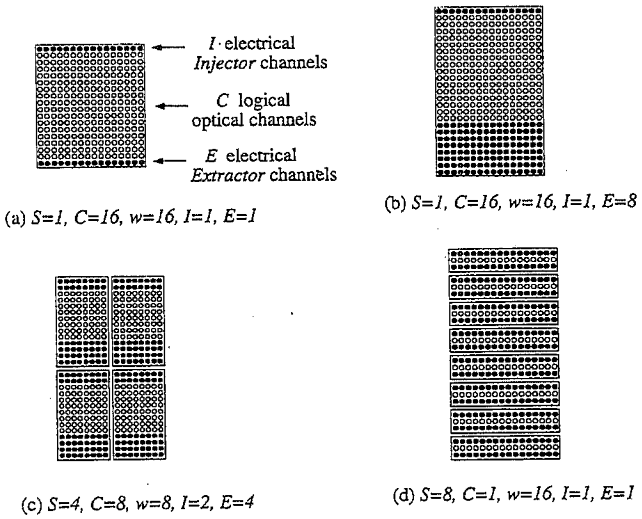


Fig. 8. 2D slice of pixels with four optical channels, three injector channels, and three extractor channels (only the four most significant bits of each channel are shown). CCU's, channel control units.

and optical channels. In the reconfigurable mode the SPA's do not perform packet processing and filtering functions. In the intelligent mode, the SPA's process packets of data as they travel down the backplane and make decisions on which packets to extract according to various extraction (or filtering) criteria.

Each 32 × 32 SPA has 1024 pixels that can be partitioned into 128 logical optical channels, each 8 bits wide (or 64 optical channels, each 16 bits wide, etc.). However, because of electronic packaging constraints, each SPA may have at most typically 24 electronic injector and extractor channels, each 8 bits wide (which would utilize approximately 400 electronic I/O pins). Along with power, ground, and control signals, this design utilizes the ~500 possible electronic I/O's discussed in Subsection 3.A.



(a) $S=1$, $C=16$, $w=16$, $I=1$, $E=1$

(b) $S=1$, $C=16$, $w=16$, $I=1$, $E=8$

(c) $S=4$, $C=8$, $w=8$, $I=2$, $E=4$

(d) $S=8$, $C=1$, $w=16$, $I=1$, $E=1$

Fig. 9. Smart-pixel organizations with varying ratios of electrical-to-optical output bandwidth: (a) 1/16, (b) 1/2, (c) 1/2, (d) 1/1.

Because of technological constraints it may be desirable to implement multiple smaller slices on a die rather than a single larger slice. The complexity of the concentrators, expanders, and arbitration circuits are proportional to the number of channels in a slice. For example, a single slice that spans 128 optical channels would have relatively complex and slow concentrators, expanders, and arbitration circuits. Smaller slices can utilize simpler and faster circuits. To achieve a reasonable pixel complexity and speed, the 32 × 32 array can be partitioned into eight slices with 16 optical channels and three injector and three extractor channels per slice ($S = 8$, $C = 16$, $I = E = 3$). Each slice therefore requires a 16-to-3 concentrator for extracting channels and a 3-to-16 expander for injecting channels. A 16-to-3 concentrator can be made in a regular layout suitable for CMOS VLSI if a smaller 4-to-3 concentrator subcircuit is implemented within each pixel for extraction. Similarly, a 3-to-16 expander can be made in a regular layout if a 4-to-1 multiplexer is implemented within each pixel, as shown in Figs. 6 and 7. An arbitration circuit that processes requests from $C = 16$ incoming optical channels and arbitrates access to three electronic extractor channels can be implemented as a finite state machine with the appropriate I/O's and state transitions. (Rather than increase the number of slices, we may also increase the width of each channel.)

Each pixel in Fig. 6 consists of an optical input port, a programmable latch circuit, an address-comparator circuit, a concentrator circuit, an expander circuit, and an optical output port. Each channel also has its own channel control unit (CCU), which stores the control signals that determine the state of the channel (see Fig. 7). In Figs. 6 and 7 the horizontal lines passing through each pixel are control signals leading to or from the CCU's. The address-comparator circuit performs the packet processing for determining which packets to extract. The CCU's also contain any arbitration circuits that generate control signals for the expanders or concentrators when the SPA's are operated in the intelligent mode, as shown in Fig. 10. A CCU typically consists of a bytewide control latch to store the state of the channel.

C. Detailed Smart-Pixel Operation

The operation of a representative SPA that performs a relatively simple basic address comparison and detection scheme is described. More detailed pixel processing circuits are described in Section 4. The pixels in a row are always comparing the incoming optical data with a unique PCB address stored in an address latch, using the distributed address-comparator circuits in a row, as shown in Figs. 6 and 11. Unique PCB addresses are downloaded from

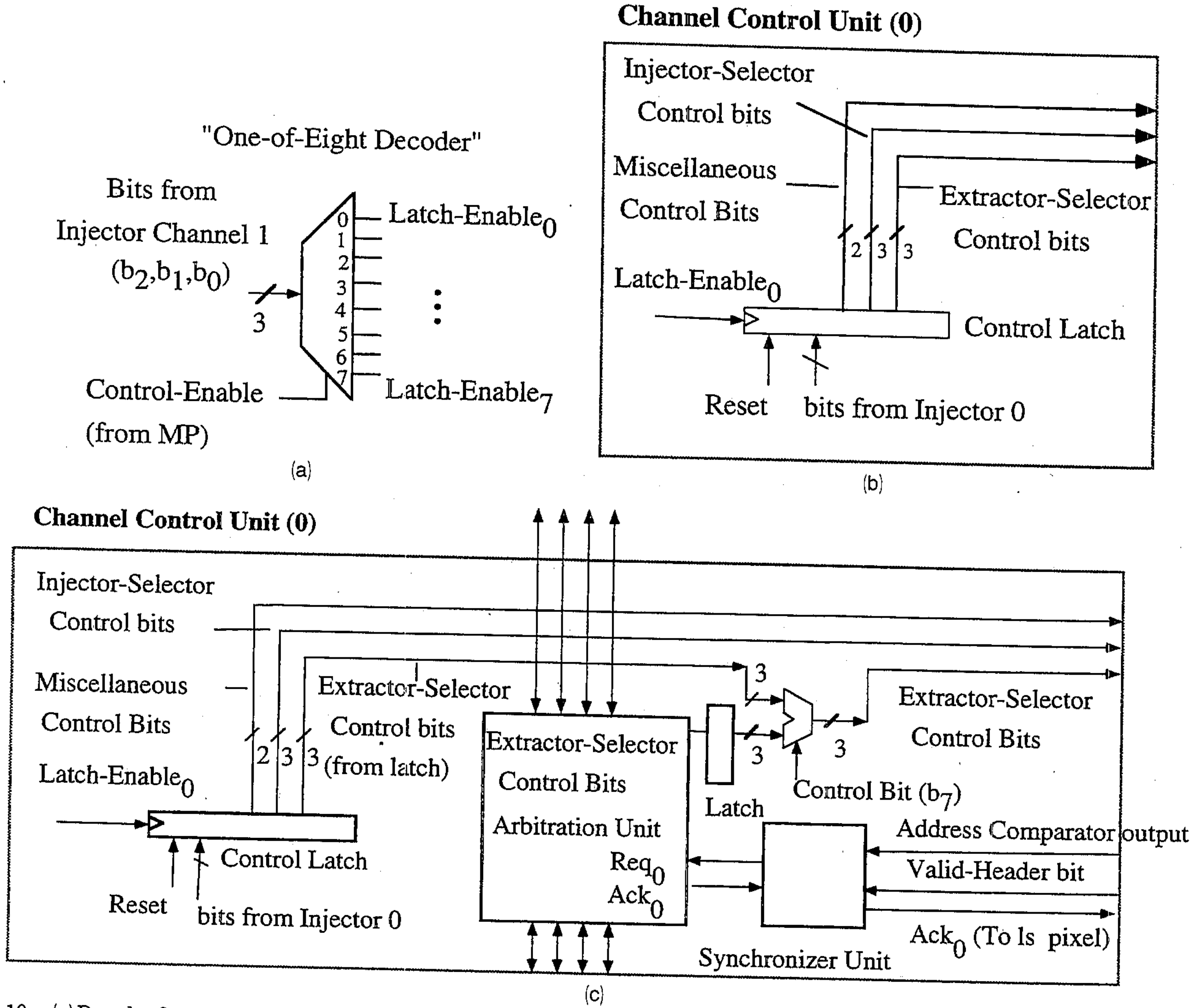


Fig. 10. (a) Decoder for writing control bits to control latches, (b) CCU for reconfigurable backplane, (c) CCU for reconfigurable intelligent backplane.

the message processor (MP) into each SPA so that a single VLSI die can be used. In Fig. 8, the unique PCB address is shown entering the top of the SPA. (To conserve I/O pins, the address bits may be loaded into the address latch bit serially.) When a row recognizes its address in the packet header it sets a Receive-Request bit for that channel, which is fed to the CCU. The arbitration circuitry in the CCU's examines all the Receive-Requests in a slice and generates the appropriate control signals (for the extractor selectors), which causes the selected channels to enter the receiving state. As shown in Fig. 11, to speed up the address detection process, all address bits in a packet can be compared with the unique PCB address in parallel.

The structure of a CCU for the intelligent backplane is shown in Fig. 10(c). [A reconfigurable backplane without packet processing capability needs only a control latch in its CCU, as shown in Figure 10(b).] The control latch stores the state of the channel: 1 bit for the state of the programmable latch circuit, either unbuffered or buffered, 3 bits for the state of the extractor selectors, 2 bits for the state of the injector selectors, and a few miscellaneous bits. Hence each CCU requires an 8-bit control latch. The receiving state of a channel is determined by three control signals coming from the CCU control latch, which determine the state of the extractor selectors in all the pixels in a channel. When receiving data, one out of three tristate drivers in each extractor selector is enabled, causing the arriving optical data to be written onto one of the three vertical extractor channels. The transmission state of a channel is controlled by two control signals coming from the CCU control latch, which determine the state of the injector selector in all the pixels in a channel. When transmitting data, each injector selector selects one datum from the three vertical injector lines or the programmable buffer circuit and forwards it to the optical output port, thereby injecting the data onto the backplane.

To configure the SPA in the reconfigurable mode, the appropriate control bits must be downloaded into the CCU's. The entire SPA also needs 8 bits to store

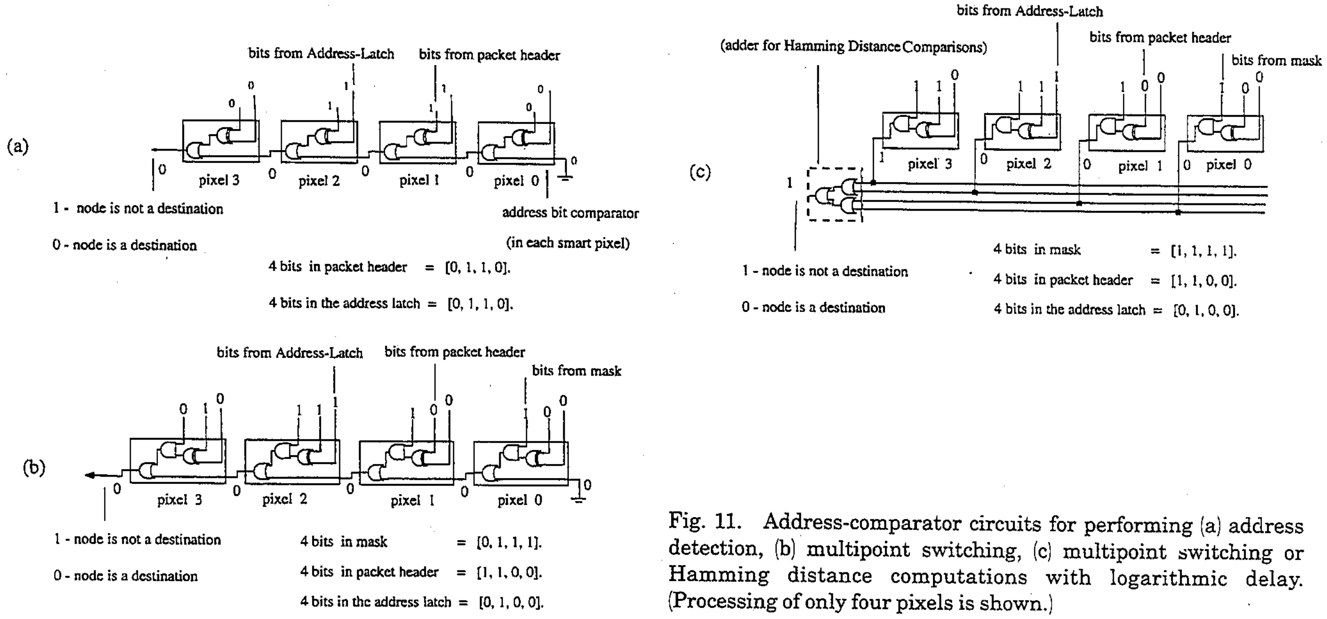

(a)
bits from Address-Latch
bits from packet header
pixel 3
pixel 2
pixel 1
pixel 0
1 - node is not a destination
0 - node is a destination
address bit comparator
(in each smart pixel)
4 bits in packet header = [0, 1, 1, 0].
4 bits in the address latch = [0, 1, 1, 0].
(b)
bits from Address-Latch
bits from packet header
bits from mask
pixel 3
pixel 2
pixel 1
pixel 0
1 - node is not a destination
0 - node is a destination
4 bits in mask = [0, 1, 1, 1].
4 bits in packet header = [1, 1, 0, 0].
4 bits in the address latch = [0, 1, 0, 0].
(c)
bits from Address-Latch
(adder for Hamming Distance Comparisons)
bits from packet header
bits from mask
pixel 3
pixel 2
pixel 1
pixel 0
1 - node is not a destination
0 - node is a destination
4 bits in mask = [1, 1, 1, 1].
4 bits in packet header = [1, 1, 0, 0].
4 bits in the address latch = [0, 1, 0, 0].

Fig. 11. Address-comparator circuits for performing (a) address detection, (b) multipoint switching, (c) multipoint switching or Hamming distance computations with logarithmic delay. (Processing of only four pixels is shown.)

its unique PCB address. The total number of control bits on the 32 × 32 SPA is 1032 bits. Assuming that one electronic injector channel can be multiplexed to provide a byte of control per clock cycle, then approximately 16 clock cycles are required for completely reconfiguring a slice, and all slices can be reconfigured in parallel. At a clock rate of 250 Mb/s, reconfiguration requires 64 ns. In the usual mode of operation, the array is programmed at initialization or each time an embedding is changed.

By appropriate setting of the control bits, the reconfigurable backplane can be used to embed classical interconnection networks such as buses, rings, 1D meshes, 2D meshes, hypercubes, shuffles, and arbitrary graph topologies in general. In the telecommunications environment, it can also be used to partition the switching system into multiple individually controlled private networks.

D. Optical Imaging Schemes

The proposed optical backplane architecture requires only a point-to-point optical interconnect between SPA's, without requiring any complex imaging operations such as 1D or 2D perfect shuffles or inverse shuffles. There are numerous approaches to implementing the point-to-point optical imaging that depend on the smart-pixel technology.[14] For example, the VCSEL-based technology described in Section 2 will use a completely different optical design compared with the modulator-based technology such as the Hybrid-SEED's described in Section 2. A key point is that our backplane architecture can function with any type of smart-pixel technology outlined in Subsection 2.A. and with any type of point-to-point optical imaging system outlined in Subsection 2.B. A review of various approaches for realizing point-to-point optical interconnections is given in Chap. 5 of Ref. 14.

E. Embedding of a 1024-Node Cray T3D Supercomputer

The Cray T3D Supercomputer is one of the most powerful supercomputers in existence.[13] The largest machine consists of 1024 processor nodes (DEC-Alpha's) arranged in a 3D 8 × 8 × 16 mesh that occupies many electronic cabinets occupying a large room. The machine has a commercial value of approximately $31,000,000 U.S. Each interprocessor channel in the Cray is 3 bytes wide in each direction (2 for data and 1 for control) and is clocked at 150 Mb/s for a channel bandwidth of 3.6 Gb/s. The bisection bandwidth of the machine is 128 × 16 × 150 Mb/s = 307 Gb/s. An 8 × 8 × 8 mesh that corresponds to the 512 processor Cray T3D interconnect can be embedded into the circular HyperPlane, as shown in Fig. 5(a). Let each PCB support 16 processors (which is achievable with advanced PCB and MCM technology; some hypercube machines support 64 processors per PCB). The 8 × 8 × 16 3D mesh can be contracted to yield a 4 × 8 2D mesh (i.e., the 16 processors on a PCB are collapsed into one node of the mesh), in which the edges in the 2D mesh now represent eight interprocessor channels from the Cray machine. The embedding of a 4 × 8 2D mesh into a circular HyperPlane with 32 PCB's with optical rings in each direction is shown in Fig. 5(a). The embedding of a 1024 processor Cray T3D will require 64 PCB's (assuming 16 processors per PCB) and will consist of the embedding from Fig. 5(a) extended to 64 PCB's. In Fig. 5(a) each bold line represents eight Cray channels (i.e., each bold line represents 24 bytewide channels), and these embeddings use 10 bold lines or equivalently 240 bytewide channels in each direction. Assuming two SPA's per ring in each direction, the backplane supports 256 bytewide channels in each direction. Hence the backplane has sufficient optical bandwidth to embed the 1024 processor Cray T3D supercomputer interconnect by using only four SPA's per PCB. Reliability and cost estimates for this system are discussed in Section 6.

F. Embedding of a 1024-Channel Multichannel Switch

Broadcast bus-based switches are often used in parallel computing machines.[15,16] Suppose that each PCB supports eight computing elements that generate traffic at the high-end bit rate of roughly 3.2 Gb/s each. With eight computing elements per PCB, each PCB is generating and receiving data at a rate of roughly 25.6 Gb/s. (The decision to place eight processors per PCB is arbitrary. With advances in MCM technology the trend is toward 8, 16, or 32 processors per PCB.)

The backplane can be configured by the allocation of a reserved broadcast bus for each PCB, as shown in Fig. 5(b) (which illustrates only 16 PCB's; the 32 PCB system would extend the embedding pattern over another 32 PCB's). Let there be four SPA's per PCB, providing 512 optical channels in the backplane. Because there are 32 PCB's and 512 optical channels, then each PCB can reserve 16 optical channels for a contention-free 128-bit-wide bus. This bus can be reconfigured so that each processor reserves two optical channels that represent 4 Gb/s of I/O bandwidth (at a clock rate of 250 Mb/s). In Fig. 5(b), the small solid boxes represent transmitters over four bytewide optical channels and each solid line represents four bytewide optical channels. A processor will never face contention when broadcasting over its own reserved channels. However, a packet transmission over the backplane can face contention at the receiving PCB if too many other packets simultaneously attempt to communicate with the same PCB. Various criteria for extracting packets are discussed in Section 4, and the performance of this multichannel switch is considered in Section 5.

4. Advanced Applications for Intelligent Optical Backplanes

A key feature of the intelligent optical backplane is the ability of smart pixels to simultaneously transport and process data packets as they travel down the backplane and make decisions on which data to extract. The ability to process terabits of data per second is not generally possible with other free-space architectures, such as the optical network described in Ref. 17. The processing schemes for point-to-point and multipoint optical switching are first described. The functionality of an intelligent optical backplane can be enhanced to cover the needs of most advanced communication and computing systems by providing more advanced processing capabilities on a SPA. The extensions are described in the subsections below.

In an intelligent backplane, data must be formatted into packets so that addressing information can be prepended to the packet headers. In an asynchronous system the packet transmissions can occur at any time without synchronization to a central clock. In a synchronous system, time is divided into slots of fixed duration, and a fixed-size packet can be transferred from one PCB to any other in a single time slot. In this paper we consider an asynchronous HyperPlane. However, a synchronous HyperPlane is easily designed.

Multipoint Optical Switching: To enable multipoint optical switching, the packet header may consists of two fields, the mask and the destination fields, in which a 0 in a mask bit implies a logical DON'T CARE for that bit position. This functionality enables multipoint switching to a wide range of selected subsets. The processing requires an EXOR, OR, and AND gate per pair of pixels (assuming that the mask and the destination appear on separate pixels), as shown by the logic equations below and in the logic circuit in Fig. 11(a):

$$L - \text{bit mask} = m_{L-1} \cdots \cdots m_0,$$

$$L - \text{bit address} = I_{L-1} \cdots \cdots I_0,$$

$$L - \text{bit destination} = d_{L-1} \cdots \cdots d_0,$$

$$\text{Extract} = \text{NOT}[m_{L-1} \bullet (I_{L-1} \oplus d_{L-1}) + \cdots + m_0 \bullet (I_0 \oplus d_0)],$$

where • indicates AND, ⊕ indicates EXOR, and + indicates OR.

Sorting: SPA's that detect inclusion within a range, in which the ranges are integers or floating-point numbers, can be used in an intelligent backplane that performs distributed sorting efficiently. (It has been estimated that a significant fraction of the world's computing power is spent sorting.) The packet header may consist of one field that denotes an integer or floating point number called the key. Each SPA is supplied with two bounds from the MP. The criteria for extraction may be inclusion or exclusion of the key within the lower and the upper bounds or various others, as shown below. To support numbers greater than w bits wide, the comparisons may occur over multiple clock cycles (or multiple channels). This may require additional latches on the chip to store the larger numbers. The CCU may also require a small finite-state machine that keeps track of the current status of the comparison. SPA's that detect the minimum or the maximum from all keys are also useful. This form of header processing requires ~12 binary gates per pixel, as shown:

$$L - \text{bit lower bound} = M_{L-1} \cdots \cdots M_0,$$

$$L - \text{bit upper bound} = N_{L-1} \cdots \cdots N_0,$$

$$L - \text{bit key} = k_{L-1} \cdots \cdots k_0,$$

$$\text{Extract} = (\text{key} \geq M) \bullet (\text{key} \leq N),$$

$$\text{Extract} = \text{NOT}[(\text{key} \geq M) \bullet (\text{key} \leq N)],$$

$$\text{Extract} = (\text{key} \geq M),$$

$$\text{Extract} = (\text{key} \leq N).$$

Resource Arbitration: A parallel-prefix operation is well known in parallel computing,[15] i.e., synchronization and resource assignment. Let each node i have a key k_i. After the parallel prefix, each node i contains $k_0 + \cdots + k_i$ (the addition can be replaced with any associative operator including logical AND or OR). To implement the parallel prefix, each SPA in PCB i may operate on the keys broadcast by PCB's $0 \cdots i - 1$, compute the running sum (i.e., the rank), and report the rank to its MP. Alternatively each array may operate on its own key and an incoming rank and report its rank to its MP and simultaneously forward it to the next PCB. Parallel-prefix computations occur frequently and are often hard wired into parallel computing machines to make execution faster. Synchronization is one example of a parallel prefix. Typically, when a processor finishes executing a subtask it sets a finished bit. The completion of the global task is determined by a logical AND of all these bits. Because the logical AND is an associative operator, this synchronization scheme is one instance of a parallel prefix. The Cray T3D supercomputer uses this scheme.[13] A parallel-prefix based on the addition of a running sum with a key will require a full adder (12 logic gates) in each pixel.

The parallel-prefix operation can also be used to provide arbitration for resources. Each PCB with a request for a resource generates a request bit as its key. Over one optical channel, the backplane may perform a parallel prefix that assigns each PCB a unique rank. If there are J resources to be allocated, then each PCB with rank $\leq J$ is assigned the resource identified by its rank; the PCB's with higher ranks lose out in the contention process and must wait for another arbitration cycle. It is possible to allocate access to the optical channels in the backplane itself if the parallel-prefix operation is used over a reserved optical channel, i.e., the backplane can be treated as multiple bus system with x buses (each $4096/x$ bits wide), and access to these buses can be arbitrated by an optical control channel that is performing a parallel prefix.

Pattern Matching: Functional memory systems such as the content-addressable memories (CAM's) allow the preprocessing of data before they are extracted from a dense VLSI memory.[18] SPA's that perform pattern matching over terabits of data may enable new models for distributed data caches, CAM's, data-flow architectures, and parallel database systems. The VLSI CAM memory provides storage and retrieval with limited I/O bandwidth and with dense processing capabilities (perhaps many thousands of comparisons within a single CAM IC). SPA's generally provide a very large I/O and processing bandwidth with potentially many comparisons occurring within the IC. Hence the SPA's may find applications as intelligent gateways that perform transportation, processing, and selection of search keys at terabit aggregate rates, leading to further processing on the processing boards.

Let each SPA store i patterns and each packet header contain one or more search keys. The optical channels perform comparisons with the search keys and the patterns in parallel and matching keys are extracted. To allow the processing of long search keys, the comparison may span multiple clock cycles. In this case the CCU may contain a small finite-state machine that keeps track of the status of the comparison. The previous functionality can be enhanced when a bit mask is associated with each search key, in which the comparators examine only the bits specified by a nonzero mask bit. The logic equations are shown below, and a typical circuit is shown in Fig. 11(b).

$$\begin{aligned}
L - \text{bit pattern}_i &= p_{L-1,i} \cdots \cdots p_{0,i}, \\
L - \text{bit mask} &= m_{L-1} \cdots \cdots m_0, \\
L - \text{bit key} &= d_{L-1} \cdots \cdots d_0, \\
\text{Extract} &= \max[[m_{L-1} \bullet (p_{L-1,i} \Theta d_{L-1})] \\
&\quad + \cdots + [m_0 \bullet (p_{0,i} \Theta d_0)]], \\
\text{Extract} &= f([[m_{L-1} \bullet (p_{L-1,i} \Theta d_{L-1})] \\
&\quad + \cdots + [m_0 \bullet (p_{0,i} \Theta d_0)]]).
\end{aligned}$$

The pattern-matching concept can be extended by compution of the Hamming distances between the search keys and patterns (according to the bits specified in a mask field) and extraction of the data if a threshold is exceeded, as shown above (where Θ now denotes logical EXOR and f is a Boolean function). Keys that match in b or more bits meet the threshold criterion and are extracted for further off-chip processing. The circuit diagram for this processing is similar to the circuit in Fig. 11(b), except the OR tree is replaced by an adder circuit that returns the Hamming distance. One may envision a terabit CAM distributed over multiple PCB's in which the strict match criterion of conventional CAM's is replaced by an exact or a near match based on Hamming distance. The optical backplane may find applications in parallel database systems and fuzzy logic inference systems.

Snoopy Caches: A number of basic functions for parallel processing can be implemented within the intelligent backplane. Parallel processing based on shared-memory schemes often relies on snoopy caches.[16] Because the intelligent backplane can support multiple broadcast channels, snoopy caching can be directly supported. When a processor changes a cached shared variable, it broadcasts the change to all other processors, who invalidate their cache entries or update their caches.

Media Access: The SPA's can also implement lower-level media-access (MAC) protocols commonly used in computing and communication systems. For example, the arrays can implement multiple token ring, packet ring, and multiple bus MAC protocols.[19] The SPA's can also implement deflection routing schemes common in computer networks and the table lookup routing schemes common in communications networks.[19]

Channel Control: Communication channels often have hardware support for basic functions such as buffering, flow control, congestion control, error control, and acknowledgment.[13] Many parallel computer networks also provide hardware support for these basic functions and additional basic functions used primarily in computing.[13,15,16] Some machines, like Thinking Machine's CM5 supercomputer, even have separate control networks; see Ref. 15, which describes the basic functions of the CM5 Control Network. These basic functions may vary among machines and applications. Control signals for these basic functions can be transported over the optical backplane by statically embedding channels for this purpose. Additional hardware support for basic functions can also be supplied within the SPA's themselves. For example, the intelligent backplane provides hardware acknowledgment of successful packet transmissions, which can also be used to provide a backpressure flow control mechanism (i.e., a sender is suspended until a positive acknowledge is received). If the transistor densities are sufficiently large in the future then entire functions may be relegated to the SPA's.

5. Performance Model

The performance of the optical backplane is analyzed in terms of basic technological parameters in this section. The following parameters with their typical values are identified:

- Z is the number of optical bytewide channels in the backplane (512).
- B is the bit-clock rate (250 Mb/s). A bit clock is defined as one period (4 ns).
- L is the length of a packet in bytes (typically 32 bytes).
- E is the effective bandwidth of a logical channel in bits per second.
- T is the transmission delay of a packet over a channel in bit clocks and is equal to L/E.
- V is the skip factor, which is the number of PCB's that can be traversed in one bit clock (related to the delay per PCB in an unbuffered optical byte channel).
- P is the propagation delay down the backplane.
- α is the probability that a data source has a packet to transmit at any time.

Consider a multichannel photonic switch with I independent broadcast channels or buses reserved for each of $N = 32$ PCB's. The backplane must support IN independent broadcast channels in the optical ring. A broadcast channel is defined as a number of optical channels that are operated in parallel and switched together as an indivisible entity. Let there be $Z = 512$ optical channels in the backplane, so that each broadcast channel can be allocated $Z/(IN) = 16/I$ optical channels. If $I = 1$, then all 16 optical channels can be operated in parallel as one broadcast channel or broadcast bus that is 128 bits wide. At a clock rate of 250 Mb/s, timing analysis indicates that optical transmissions should traverse $V = 4$ PCB's in one bit-clock period (4 ns), i.e., optical signals pass through SPA's with a delay of 1 ns per PCB. In the worst case, a transmission from PCB 0 must reach PCB 31, thereby passing through $N - 1$ PCB's, so that the end-to-end propagation delay is $(N - 1)/V$ bit clocks. (The optical channel can be operated completely unbuffered so that optical signals are essentially traveling waves with their own clocks; the source uncompelled protocol.[20] To minimize skew and maximize pipelining, we may program PCB's 3, 7, . . . , 27 to be buffered and the other PCB's to be unbuffered. The buffering will resynchronize the parallel signals in a byte channel.) Assuming that $I = 8$ and $L = 32$, the transmission delay of a packet is 16 bit clocks and the propagation delay is 8 bit clocks, yielding a minimum packet delay of ~100 ns:

$$T = \lceil L/(Zw/IN)\rceil = \lceil 32/\lceil 512/(8 \times 32)\rceil\rceil = 16,$$
$$P = \lceil (N - 1)/V\rceil = \lceil 31/4\rceil = 8.$$

In the absence of blocking and at full load, the multichannel optical switch delivers one packet over each broadcast channel in each packet transmission interval. Given a random traffic model, the multichannel optical switch has a blocking probability PB given in Subsection 5.A.; hence the aggregate bandwidth of the multichannel switch embedded in the HyperPlane is given by

$$\mathrm{BW} \approx \alpha Z(1 - \mathrm{PB})(wB).$$

A. Blocking Model

Blocking occurs when an excessive number of packets simultaneously attempt to connect to the same PCB. Each slice in a SPA has E extractor channels to remove packets from the backplane. When more than E packets arrive at the same time at one slice, E packets will be selected and routed out, and the remaining packets will not be received. The expected blocking probability, given a random uniform traffic model, will be quantified for the intelligent optical backplane.

Each asynchronous traffic source can be modeled as a Poisson process that is generating packets at a fixed rate. When a new packet is generated it is transmitted over the backplane and in the absence of blocking the packet is thereby removed from the source. The fraction of time a traffic source is transmitting a packet (in the absence of blocking) is given by α. Hence $\alpha = 1$ corresponds to a fully loaded traffic source. Assuming that all sources are identical and independent, the probability that j sources are transmitting simultaneously to one destination is a binomially distributed random variable. With the Poisson approximation to the binomial, it can be shown that the probability that a packet is blocked at the receiving slice is given by

$$\mathrm{PB} = \frac{N}{\alpha IC}\left[\sum_{j=E+1}^{\infty} (j - E)\,\frac{\exp(-\alpha IC/N)(\alpha IC/N)^j}{j!}\right]. \quad (1)$$

B. Throughput and Delay Models

Queuing can be used to handle contention that may occur over the backplane. A packet transmission over the backplane can be blocked if too many packets attempt to communicate with the same PCB. The sender can be notified when a packet is blocked through an acknowledge bit that is also transmitted back to the sender over the optical ring. Hence each processor can utilize a small queue to handle the case when its packet transmission blocks so that the packet can be retransmitted after the blocking clears. In general, queues may be placed at the input side, the output side, or both sides of a multichannel switch.

The approximate queueing analysis has been performed based on the well-known $M/M/Y/\infty$ continuous-time queuing models and yields first-order approximations on the throughput and delay of the backplane.[19] It is reasonably accurate for the broad design-space exploration presented in this section. Packets are generated on each PCB by the eight computing elements and forwarded into eight dedicated input queues at the aggregate rate of λ packets/s. Packets are removed from the input queues and transmitted over the free-space backplane in a first-in first-out order at the rate of $\mu_{\max}$ packets/s, where $\mu_{\max}$ is the peak carried bandwidth per PCB expressed in packets per second. Alternatively, all eight input queues can be merged into a single shared input queue with eight servers. The multiserver input queue can remove and transmit eight packets simultaneously over the backplane in the absence of any blocking. The multiserver input queue can be modeled as a continuous-time $M/M/Y/\infty$ queuing system, with memoryless interarrival and service time distribution, with $Y \geq 1$ servers, and with infinite queuing capacity.[19]

Assuming infinite queueing capacity, the expected delay $\overline{D}$ is given by Little's law, $\overline{D} = \overline{C}/\overline{Z}$, where $\overline{C}$ is the expected number of customers in the queuing system and where $\overline{Z}$ is the throughput. Defining the utilization $\rho = \lambda/Y\mu_{\max}$, the expected number of customers is given by

$$\overline{C} = \sum_{i=0}^{Y} \frac{Y^i \rho^i}{i!}\, iP_0 + \sum_{i=Y+1}^{\infty} \frac{Y^Y \rho^i}{Y!}\, iP_0,$$

where

$$\overline{Z} = \begin{cases} \lambda & \text{if } \lambda \leq Y\mu \\ Y\mu & \text{if } \lambda > Y\mu \end{cases},$$

and where P_i is the equilibrium probability that the queue has i packets.[19] (To compare the packet delays fairly between different embedded switches, the offered loads to each embedded switch must be held equal.)

C. Numerical Results and Discussion

Figure 12 illustrates the blocking probability versus offered load for the multichannel backplane, given a random uniform traffic model, as a function of the number of extractor channels per slice. Figure 12(a) applies to a backplane in which each PCB has a single injector and a varying number of extractors ($I = 1, E = 1 \cdots 8$). As expected, as the number of extractors increases, the blocking drops dramatically. In Fig. 12(b) each PCB has four injectors and a varying number of extractors ($I = 4, E = 1 \cdots 8$). For the same number of extractors, the blocking probability increases slightly compared with Fig. 12(a), primarily because there are four times as many injectors competing for access to the same number of extractors. Although the blocking probability is slightly higher, this switch has a higher utilization of its extractor channels and therefore carries more traffic. In Fig. 12(c) each PCB has S slices, each with 16 optical channels, one injector, and E extractors ($S = 8, C = 16, I = 1, E = 1 \cdots 8$). Compared with Figs. 12(a) and 12(b), the blocking in Fig. 12(c) is reduced considerably because of the larger number of extractor channels made available from the S separate slices. Figure 12 illustrates that the blocking probability of a multichannel switch can be designed to be very low so that, on average, a packet will be successfully received on its first transmission attempt.

Figure 13(a) illustrates that an aggregate bandwidth of 900 Gb/s is achievable in a 32 PCB backplane, given a random uniform traffic model. Figure 13(b) indicates that when each PCB supports a bandwidth of ~25.6 Gb/s, a packet delay in the range of 100–200 ns can be expected for offered loads with ~800–900 Gb/s aggregate bandwidth. Hence the analytic model indicates that the optical backplane will have acceptable performance and that packet blocking will not be significant.

6. System Cost and Availability Estimates

A. System Cost Estimates

It is expected that smart-pixel optical technology will initially appear in high-end supercomputing and communications architectures, where the bandwidth limitations are currently among the most challenging aspects of system design[15] (favoring an optical solution) and in which the initial cost of the optical technology can be absorbed. After a sufficient industrial infrastructure is developed, the cost of optical technology is expected to drop significantly as its use increases.

In Subsection 3.E. it was shown that the 1024-node Cray T3D supercomputer could be packaged into a 64-PCB backplane with four SPA's and 16 processors per PCB. An estimate of the cost of optics is presented. Assuming that the cost of each SPA and its associated optomechanics can be limited

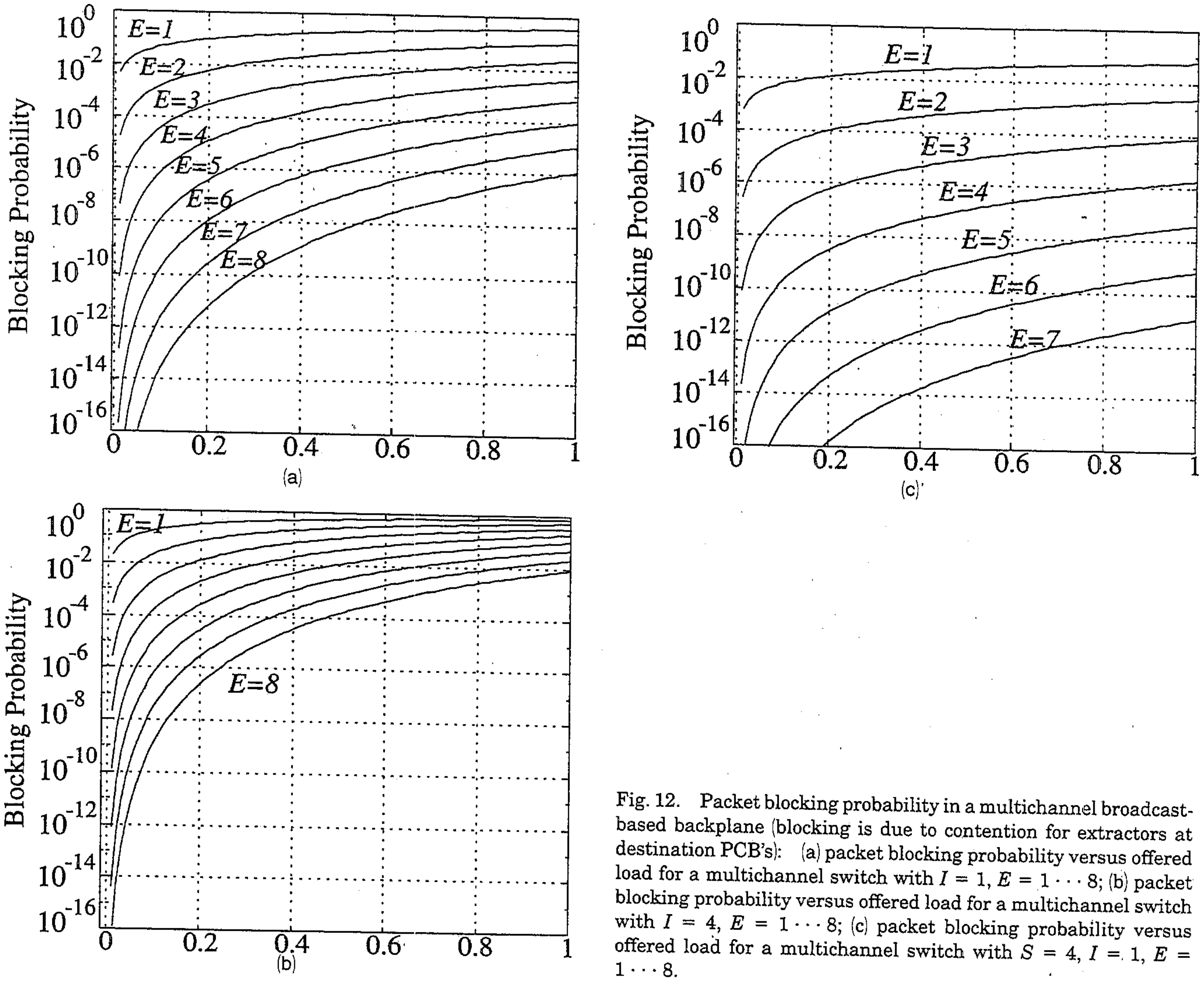


Fig. 12. Packet blocking probability in a multichannel broadcast-based backplane (blocking is due to contention for extractors at destination PCB's): (a) packet blocking probability versus offered load for a multichannel switch with $I = 1$, $E = 1 \cdots 8$; (b) packet blocking probability versus offered load for a multichannel switch with $I = 4$, $E = 1 \cdots 8$; (c) packet blocking probability versus offered load for a multichannel switch with $S = 4$, $I = 1$, $E = 1 \cdots 8$.

to ≤$5000, the cost of the optics per PCB will be limited to ≤$20,000. At current prices, the commercial value of the 16 processor PCB will be approximately $500,000, i.e., a total T3D system cost of $31,000,000 amortized over 64 PCB's with their associated optics. (This figure is consistent with other high-end PCBs; for example, the switching PCB in the IBM supercomputer has a commercial value of approximately $100,000, and the IBM PCB is simpler than the proposed Cray PCB.) Hence the cost of the optical components will represent approximately 4% of the commercial value of each PCB. In summary, the cost for optics can be more easily absorbed in the near term in high-end supercomputing and communication applications, in which the commercial value of the PCB's varies from approximately $100,000 to $500,000. However, the cost of the optics will drop as its use increases.

B. System Availability Estimates

The most common failure mode in Si IC's is individual electronic pin-out failure. This type of failure can be caused by excessive heating, temperature cycling, excessive current, oxidation, etc. The probability of some other type of catastrophic failure that entirely disables an otherwise functioning IC is much less likely and is usually ignored in reliability analysis. In a SPA the individual optical I/O failures may also be modeled (although sufficient empirical data have not yet been gathered).

In Subsection 3.E., each PCB in the proposed Cray system utilized 240 channels out of 256 available channels in each direction. Hence each PCB has access to 16 spare channels in each direction. Because the backplane can be dynamically reconfigured to bypass faulty channels, it follows that each PCB can tolerate the failure of up to 16 channels in each direction and the system will still function at 100% availability. The failure of 17 or more channels in one direction in any one PCB will start to have an impact on the system performance. In practice, we expect the probability of 17 or more random pin-out failures over two CMOS IC's on one particular PCB between maintenance intervals to be very small (otherwise existing nonreconfigurable electronic systems would be unable to function). To

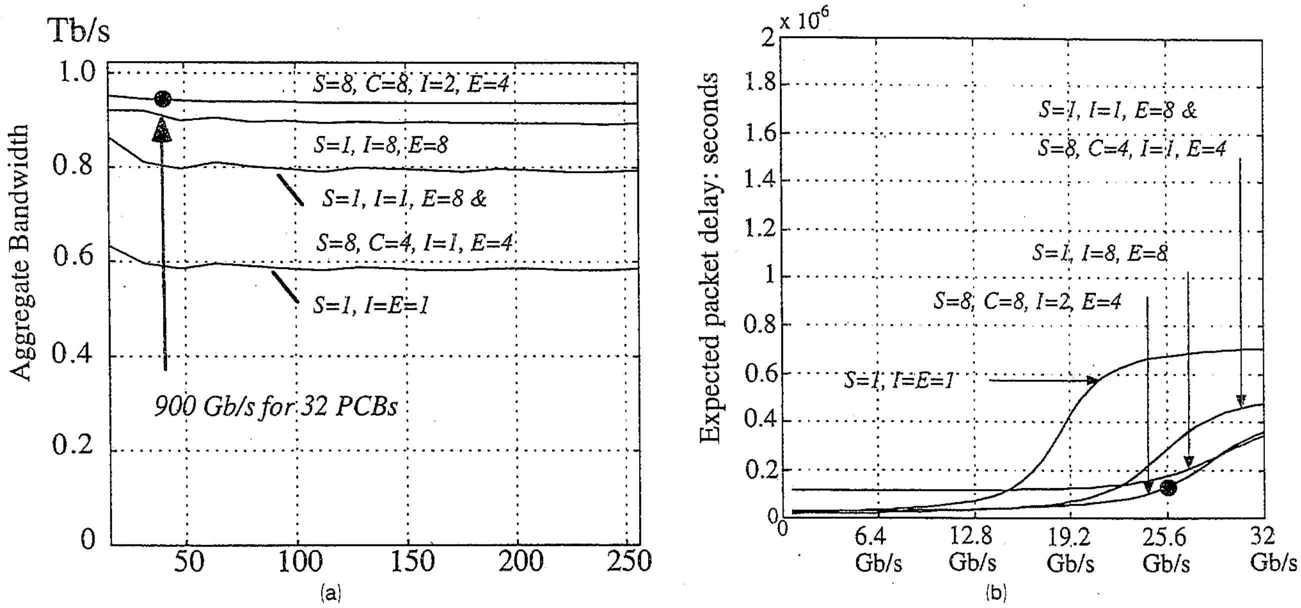


Fig. 13. (a) Aggregate bandwidth of 900 Gb/s is achievable, given a random uniform traffic model. (b) Packet delays of 100–200 ns are achievable in a system that support 256 processors and 32 PCB's, with each PCB generating backplane traffic at 25.6 Gb/s. The filled circles indicate the operating points of the 32-board optical backplane.

improve reliability further, the PCB's can be designed to have reconfigurable channel widths by the use of reconfigurable message processors. Such a system could be dynamically reconfigured to use narrower channels in the presence of many channel failures. For example, each channel in the Cray could be reconfigured to be 2 bytes wide rather than 3. In this case, the 3D mesh embedding in Subsection 3.E. will require 160 channels out of 256 available channels. It follows that each PCB could tolerate up to 96 channel failures out of 256 (i.e., an exceptionally high 37.5% failure ratio), and the system would still function at 100% of its reduced load. In summary, the dynamic reconfigurability of the backplane can ensure high system performance and availability.

7. Conclusions

The architecture of a reconfigurable intelligent optical backplane for computing and communications applications has been described. The paper included a description of the basic HyperPlane architecture and associated SPA's and how they could be used to implement (1) dynamically reconfigurable connections between any PCB's, (2) dynamic embeddings of classical interconnection networks, (3) multipoint switching, (4) sorting, (5) parallel prefix, (6) pattern matching, (7) snoopy caches and intelligent memory systems, and (8) MAC and channel control. A performance model of the intelligent backplane indicates that it has the ability to support bisection bandwidths in excess of a terabit per second.

The contributions of Manoj Verghese, Palash Desai, Michael Kim, Stephane Gagnon, Thomas Obenaus, Sherif Sherif, Boonchuay Supmonchai, and David Plant of McGill University are gratefully acknowledged.

References

1. R. A. Nordin, A. F. Levi, R. N. Nottenburg, J. O'Gorman, T. Tanbun-Ek, and R. A. Logan, "A systems perspective on digital interconnection technology," J. Lightwave Technol. **10,** 811–827 (1992).
2. R. R. Tummala and E. J. Rymaszewski, eds., *MicroElectronics Packaging Handbook,* (Reinhold, New York, 1989).
3. L. A. D'Asaro, L. M. Chirovsky, E. J. Laskowski, S. S. Pei, T. K. Woodward, A. L. Lentine, R. E. Leibenguth, M. W. Fucht, J. M. Freund, G. G. Guth, and L. E. Smith, "Batch fabrication and operation of GaAs–Al$_x$Ga$_1$–xAs field-effect transistor-self-electrooptic effect device (FET-SEED) smart pixel arrays," IEEE J. Quantum Electron. **29,** 670–677 (1993).
4. C. Camperi-Ginestet, B. Buchanan, S. Wilkinson, N. M. Jokerst, and M. A. Brooke, "Integration of InP-based thin film emitters and detectors onto a single silicon circuit," in *Optical Computing,* Vol. 10 of 1995 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1995), pp. 145–147.
5. K. Goossen, J. A. Walker, L. A. D'Asaro, S. P. Hui, B. Tseng, R. Leibenguth, D. Kossives, D. D. Bacon, D. Dahringer, L. M. F. Chirovsky, A. L. Lentine, and D. A. B. Miller, "GaAs MQW modulators integrated with silicon CMOS," IEEE Photon. Technol. Lett. **7,** 360–362 (1995).
6. S. Matsuo, T. Nakahara, Y. Kohama, Y. Ohisa, S. Fukushima, T. Kurokawa, "Photonic switch monolithically integrating an MSM PD, MESFETs, and a vertical-cavity surface-emitting laser," in *LEOS '94* (IEEE, New York, 1994), postdeadline paper PD2.1.
7. H. S. Hinton, "Free space digital optical systems," Proc. IEEE, **82,** 1632–1649 (1994).
8. D. R. Rolston, B. Robertson, H. S. Hinton, and D. V. Plant, "Analysis of a microchannel interconnect based on the clustering of smart-pixel-device windows," Appl. Opt. **35,** 1220–1233 (1996).

9. T. H. Szymanski and H. S. Hinton, "Architecture of a terabit free-space photonic backplane," in *Proceedings of the International Conference on Optical Computing* (IEEE, New York, 1995), pp. 141–144.

10. T. H. Szymanski, "Intelligent optical backplanes," in *Optical Computing,* Vol. 10 of 1995 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1995), pp. 11–13.

11. T. H. Szymanski and H. S. Hinton, "Design of a terabit free-space photonic backplane for parallel computing," in *Proceedings of the Second International Conference on Massively Parallel Processing using Optical Interconnects* (IEEE Computer Society, Washington, D.C., 1995), pp. 16–27.

12. H. S. Hinton and T. H. Szymanski, "Intelligent optical backplanes," in *Proceedings of the International Conference on Massively Parallel Processing using Optical Interconnects* (IEEE Computer Society, Washington, D.C., 1995), pp. 133–143.

13. *Cray T3D System Architectural Overview,* (Cray Research Inc., Chippewa Falls, Wisc., 1993).

14. H. S. Hinton, *An Introduction to Photonic Switching Fabrics* (Plenum, New York, 1993), Chap. 5, p. 245.

15. K. Hwang, *Advanced Computer Architecture: Parallelism, Scalability, Programmability,* (McGraw-Hill, New York, 1993), Chap. 8.5, p. 465.

16. J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach* (Morgan Kauffman, San Mateo, Calif., 1995), Chap. 9, p. 720.

17. I. Redmond and E. Schenfeld, "A distributed reconfigurable free-space optical interconnection network for massively parallel processing architectures," in *Proceedings of the International Conference on Optical Computing* (IEEE, New York, 1995), pp. 215–218.

18. K. Tamaru, "The trend of functional memory development," Inst. Electron. Inf. Commun. Eng. (Jpn.) Trans. Electron. **E76 C,** 1545–1554 (1993).

19. D. Bertsekas and R. Gallager, *Data Networks,* 2nd ed. (Prentice-Hall, Englewood Cliffs, N.J., 1992), Chap. 2, p. 37.

20. J. Dia Giacomo, *Digital Bus Handbook* (McGraw-Hill, New York, 1990), Chap. 18, p. 18.1.