Optical Centralized Shared Bus Architecture for High-Performance Multiprocessing Systems

Xuliang Han, Ray T. Chen

Microelectronics Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin

ABSTRACT

With the increasing demand for solving more complex problems, high-performance multiprocessing systems are attracting more and more research efforts. One of the challenges is to effectively support the communications among the processes running in parallel on the multiprocessors. Due to the physical limitations of electrical interconnects, interconnection networks impose a potential bottleneck limiting the overall performance. On the other hand, optics has many advantages as an interconnect technology. In this paper, benefits of optics are evaluated along with a comparison of two mainstream system topologies, shared bus and switched media. This analysis leads to an innovative interconnect architecture, optical centralized shared bus. The crucial design aspects of this architecture, including system organization, working principle, and conversion between free-space propagation and substrate-guided mode propagation by using volume holographic gratings, are delineated. To ensure the feasibility of using this architecture as high-performance interconnection networks in real multiprocessing systems, a PCI implementation of the centralized shared memory multiprocessor system is proposed. In this prototype, the required connectivity is accomplished by using the optical centralized shared bus architecture. Some preliminary results are presented.

Keywords: Optical Interconnect, Shared Bus, Switched Media, Multiprocessing, Centralized Shared Memory Multiprocessor, Optoelectronic Interface

1. INTRODUCTION

With the increasing demand for solving more complex problems, high-performance multiprocessing systems are attracting more and more research efforts. One of the challenges is to effectively support the communications among the processes running in parallel on the multiprocessors. Thus, interconnect is becoming an even more dominant factor in modern computation systems. As shown from a typical electrical backplane bus in Fig. 1, the simplest way to connect multiple nodes is to have them share a common media, the metal traces on the backplane in this example, and daughter boards can be plugged into the designated backplane connectors to obtain access to this shared media. Another mainstream approach is to use a switched media. Both topologies are illustrated in Fig. 2 [1]. One straightforward way to increase the system performance is to utilize higher clock frequency. As the underlying implementation technology, however, electrical interconnects face numerous challenges, such as power consumption, signal integrity, and electromagnetic interference as the clock frequency increases. On the shared bus, there is physical discontinuity at each tap. This discontinuity produces reflection waves that degrade the signal integrity. Therefore, in order to utilize higher clock frequency, in the electrical domain we are now seeing system topology changing from the simple shared bus toward the complex switched media as shown in Fig. 2. The switched media allows multiple pairs of nodes to communicate simultaneously, providing huge aggregation bandwidth. The simple point-to-point configuration relieves the physical limitations of electrical interconnects, allowing high clock frequency to be utilized. These are the main rationales for this topology trend in the electrical domain. However, the switched media is an indirect network, and thus cannot effectively support broadcast. Besides the additional latency of going through the switch fabric, routing latency incurs whenever a link needed to be established, and this latency increases with the complexity of the switch fabric. In contrast, the shared bus is a direct network. All messages incur single-hop latency once the access to the network is

Photonics Packaging and Integration III, Randy A. Heyler, David J. Robbins, Ghassan E. Jabbour, Editors, Proceedings of SPIE Vol. 4997 (2003) © 2003 SPIE · 0277-786X/03/$15.00

granted. The broadcast nature allows effective utilization of snooping protocols to support cache coherence on a multiprocessor, reducing communication overhead. Thus, retaining the shared bus architecture at the board-to-board level is beneficial for the overall performance, especially in the latency-sensitive scenarios. The employment of optical interconnects will be one of the major alternatives for upgrading the interconnect performance [2]. Machine-to-machine interconnection has already been significantly improved by utilizing optical means. The major research thrusts in optical interconnects are at the backplane and board levels where the physical limitations of electrical interconnects are imposing a prominent bottleneck. Optical interconnects have been investigated in both shared bus and switched media scenarios. Besides the impact of topology on the system performance previously discussed, the effectiveness of using optical interconnects also depends on the maturity of non-linear optical technologies. Inside a switch fabric, there are two important function units, switches and buffers. So far, optical technologies cannot effectively implement switching and buffering. Thus, these two function units still rely on electronic technologies. This fact implies OE and EO conversion at the interface of the switch fabric in an optically interconnected network. This overhead introduces additional latency. Furthermore, the number of optoelectronic interface modules is at least doubled, which dramatically increases the cost. Therefore, the remaining part of this paper will focus on using optical interconnects to implement a shared bus based architecture.

At the board-to-board level, optical implementation techniques can be classified into three categories, optical waveguide interconnections [3], free-space interconnections [4], and substrate-guided interconnections [5], as illustrated in Fig. 3. Similar to the metal traces, optical waveguides provide interconnects in a planar configuration. Free-space optics provides both high bandwidth potential, and the possibility of flexible routing. However, only simple point-to-point interconnects can be implemented, thus, not suitable for the development of shared bus based architectures. Another disadvantage is that the free-space interconnections are subject to the disturbance in the surrounding environment. This problem can be avoided by confining optical signals within a waveguiding substrate, as shown in Fig. 3. At the source node, the optical signal is coupled into the substrate by a properly designed holographic grating. In order to confine the light within the substrate, the incident angle at the interface must be larger than the critical angle, i.e., the total internal reflection (TIR) condition must be satisfied. At the destination node, another properly designed holographic grating couples the light back into free-space propagation toward the receiver. Thus, an optical link from the source to the destination is established. With appropriate design of the types and positions of the holographic gratings, broadcast can be effectively implemented. Thus, it is possible to utilize this method to develop shared bus based architectures, as reported in [6] and [7]. These previously developed optical shared-bus architectures, however, are difficult to implement in practice due to the large variation among the optical signal fan-outs that dramatically increases the complexity of optoelectronic interface modules. The difficulty in equalizing the fan-outs is mainly due to the requirement to support the bi-directionality of signal flows on the backplane bus. In our previous report [8], a new method for rebroadcast signals in an optical backplane bus system was described. Uniform optical signal fan-outs become possible by using this method. Upon this method, we present in this paper an innovative architecture called the optical centralized shared-bus, which is, to the best our knowledge, the first fan-out equalized optical backplane bus. As shown in Fig. 4, the electrical bus provides interconnects for the non-critical paths, whereas the optical bus for the critical paths. To implement optical interconnects, volume holographic gratings are used for coupling optical signals into and out of the optical wave-guiding plate within which optical signals propagate as substrate-guided waves. The details of this innovative architecture, especially the feature of uniform optical signal fan-outs, are described in the Section 2. In Section 3, a PCI implementation of the centralized shared memory multiprocessor system is proposed to ensure the feasibility of this innovative architecture. Finally, a summary is given in Section 4.

2. OPTICAL CENTRALIZED SHARED BUS ARCHITECTURE

Fig. 4 illustrates the architectural concept of the optical centralized shared-bus. The electrical backplane layer provides interconnects for the non-critical paths. The daughter board that is inserted into the central backplane connector plays a pivotal role in this architecture and is referred to as distributor in this paper. The optoelectronic interface modules, including VCSELs and photodetectors, are integrated on the backside of the electrical backplane and aligned with the underlying optical backplane layer. Different from the other modules, the positions of the VCSEL and the photodetector in the central module are swapped. The optical backplane layer consists of an optical wave-guiding plate with the properly designed volume holographic gratings integrated on its top surface. Underlying the distributor is a

double-grating hologram, and the others are single-grating holograms. In this way, this architecture provides the bi-directionality of signal flows on the backplane bus.

In this architecture, there are two optical paths for each signal line. One is for the source daughter board to deliver the signal to the distributor, and the other one for the distributor to broadcast the signal to all the daughter boards on the backplane bus. A complete data transfer from a daughter board to (1) another daughter board (point-to-point), (2) more than one daughter board (multicast), and (3) all the daughter boards on the backplane bus (broadcast), generally involves two processes, which are single-hop delivery from the source daughter board to the distributor and then broadcast from the distributor. This explains the reason we name this architecture as optical centralized shared-bus. First, the VCSEL of the source daughter board projects the light surface-normally on its underlying holographic grating. This signal is coupled into the optical wave-guiding plate and propagates within the plate under the TIR condition. Then, this signal is surface-normally coupled out of the plate by the central double-grating hologram and detected by the receiver of the distributor. Second, the distributor regenerates the same optical signal and projects it surface-normally on its underlying double-grating hologram. This signal is diffracted into two beams and coupled into the optical wave-guiding plate, propagating along the two opposite directions within the plate under the TIR condition. During the propagation, a portion of the light is surface-normally coupled out of the plate by a daughter board's underlying holographic grating and detected by its receiver. This daughter board takes appropriate actions on the received data. If the distributor is the data source, the first process will not happen. If the distributor is the only data destination, the second process is not necessary. Furthermore, in conformity with a specific global topology, a hierarchical interconnection network can be constructed by using the optical centralized shared-bus as the building block and the distributor as the socket.

The most attractive feature of this architecture is to achieve uniform optical signal fan-outs. Assuming that the VCSELs of all the optoelectronic interface modules emit the same optical power, uniform fan-outs mean that (1) the power of signals delivered from any daughter board to the distributor is same, (2) the power of signals broadcast from the distributor to all the daughter boards on the backplane bus is same, and (3) the power of signals broadcast from the distributor equals that delivered from any daughter board to the distributor. Thus, this architecture specifies a uniform interface between the electrical and the optical backplane layers in contrast to other proposed architectures, e.g., [6] and [7]. The fan-outs are equalized by specifying the diffraction efficiency of the volume holographic gratings. Because of the symmetric configuration, it is obvious that the two multiplexed gratings inside the central hologram should have the same diffraction efficiency. The analysis in [8] shows that the fan-outs are equalized if the following iterative equation is satisfied,

$$\eta_{(i+1)} = \frac{\eta_i}{1-\eta_i}, \quad (1)$$

where η_i represents the diffraction efficiency of the i-th single-grating hologram counted from the central double-grating hologram. In the case of uniform fan-outs, the fan-out coefficient, which is defined as the ratio of the fan-out power to the effective VCSEL fan-in power, is obviously equal to the reciprocal of the total number of the daughter boards (not including the distributor) on the backplane bus. Thus, the fan-out capacity, i.e., the maximum number of daughter boards that one optical centralized shared-bus can accommodate, can be calculated if the bit error rate requirement, marginal power penalty, and the parameters of the optoelectronic interface modules, such as VCSEL emission power and photodetector sensitivity, are specified.

The volume holographic gratings specified by the optical centralized shared-bus architecture were recorded in dry photopolymer films (HRF-600X014-20, DuPont). The photopolymer-based volume hologram is an attractive candidate for making high efficiency gratings. The advantages of this material over other types of emulsion, such as dichromated gelatin and silver halides, include dry-processing capability, long shelf life, and good photo-speed [9]. The two-beam interference method was used to form the gratings in the dry photopolymer films. This material consists of monomers, polymeric binders, and photoinitiators. The monomers are polymerized when exposed to the light of specific wave-length, and the refractive index of the film is determined by the polymer concentration. While being exposed to an interference pattern, there are more monomers being polymerized in the bright regions than in the dark regions. This non-uniform illumination sets up monomer concentration gradients, driving the monomers to diffuse from the dark regions to the neighborhood bright regions. A final uniform illumination is required to polymerize the remaining monomers and stabilize the spatial distribution of the polymer concentration, which conforms to the original illumination

pattern. Thus, a grating structure is formed inside the film. To obtain the double-grating holograms, two sequential exposure steps are required to form the two multiplexed gratings inside the films. Our objectives are to obtain single-grating holograms with accurate diffraction efficiency that satisfies iterative equation (1), and high-efficiency equal-strength double-grating holograms. The quality of these volume holographic gratings directly affects the fan-out variation and capacity, therefore, are pivotal to the implementation of the optical centralized shared-bus architecture. The recording schedules were developed by using the method described in [10]. Following these recording schedules, we were able to control the accuracy of the diffraction efficiency within 2% and obtained 47%/47% double-grating holograms. Fig. 5 (a) demonstrates the uniform optical signal fan-outs with single bus line. In conformity with the specification of the optical centralized shared-bus architecture, a 47%/47% double-grating hologram is integrated in the middle of the optical wave-guiding plate. The other two are single-grating holograms with 50% diffraction efficiency. The diffraction angles within the plate are 45°, which satisfies the TIR condition. The two 22.5° bevels at both ends are coated with aluminum, providing nearly 100% reflection efficiency. An 850nm VCSEL source was used to obtain the result as shown in Fig. 5 (a). The fan-out variation was measured to be within 2.5%. A two-dimensional (2-D) bus line configuration [11] can be implemented by replacing the optical source with a 2-D VCSEL array as shown in Fig. 5 (b).

3. PCI IMPLEMENTATION OF THE CENTRALIZED SHARED MEMORY SYSTEM

Classified by the memory organization, there are two multiprocessor models as illustrated in Fig. 6. As previously discussed, compared with the distributed memory model the centralized shared memory model is more cost-effective in the latency-sensitive scenarios. In a centralized shared memory multiprocessor, several microprocessors share a single physical memory connected by a common bus. The shared data may be replicated in multiple caches. All cache controllers snoop on the bus in order to keep all caches coherent. Besides providing high bandwidth potential, the centralized shared bus fits this particular multiprocessor model from the architecture point of view. Thus, a demonstration of the centralized shared memory multiprocessor where the required connectivity is accomplished by using the optical centralized shared bus architecture will ensure the feasibility of this innovative approach in real multiprocessing systems.

Peripheral Component Interconnect (PCI) protocol can be adapted to emulate the centralized shared memory multiprocessor system. In our proposed demonstration system as shown in Fig. 7, a single board computer (SBC), a PCI memory card, and a Gigabit Network Interface Card (NIC) reside on a passive PCI backplane. The microprocessor communicates with the PCI subsystem via a special chipset known as host bridge, which handles all necessary tasks to transfer data from or to the microprocessor or the main memory subsystem, thus, completely separates the PCI subsystem from the main memory subsystem. The main memory of the SBC is referred to as local memory as in the real centralized shared memory multiprocessor system. The PCI memory card resides on the PCI bus as a bulletin board for data sharing. Its memory is mapped into I/O space, which creates a separate and distinct memory partition. The Gigabit NIC is directly connected to another Gigabit NIC of another workstation by using a RJ-45 crossover cable and thus functions an asynchronous data agent, emulating the other microprocessor on the PCI bus. Both the PCI memory card and the Gigabit NIC are capable to initiate transactions on the PCI bus. With bus master transfer mode, a card can transfer its data to the target without any CPU action. This is similar to the Direct Memory Access (DMA) transfer on an Industrial Standard Architecture (ISA) bus with the only difference that the bus master controller sits on the PCI card. These features make it possible to emulate the centralized shared memory multiprocessor system by using a modified PCI protocol. The implementation of such a system by using the optical centralized shared bus architecture is being developed in our Lab.

4. CONCLUSION

In this paper, benefits of optics are evaluated along with a comparison of two mainstream system topologies, shared bus and switched media. This analysis leads to an innovative interconnect architecture, optical centralized shared bus. This architecture retains the advantages of shared-bus topology while at the same time specifying a uniform interface between the electrical and the optical backplane layer in contrast to other proposed architectures. For the first time, a fan-out equalized optical backplane bus is developed. A PCI implementation of the centralized shared memory

multiprocessor system is proposed. In this prototype, the required connectivity will be accomplished by using the optical centralized shared bus architecture. This demonstration will ensure the feasibility of using this architecture as high-performance interconnection networks in real multiprocessing systems.

ACKNOWLEDGEMENTS

The authors thank BMDO, DARPA, ONR, AFOSR, and the ATP program of the State of Texas for supporting this study.

REFERENCES

1. D. Bouvier, "An embedded system component network architecture," www.RapidIO.org
2. M. R. Feldman, S. C. Esener, C. C. Guest, and S. H. Lee, "Comparison between optical and electrical interconnects based on power and speed characteristics," *Applied Optics*, vol. 27, pp. 1742-1751, May 1988.
3. Y. Liu, L. Lin, C. Choi, B. Bihari, R. T. Chen, "Optoelectronic integration of polymer waveguide array and metal-semiconductor-metal photodetector through micromirror couplers," *IEEE Photonics Technology Letters*, Vol. 13, pp. 355-357, April 2001.
4. T. Sakano, T. Matsumoto, K. Noguchi, T. Sawabe, "Design and performance of a multiprocessor system employing board-to-board free-space optical interconnections, COSINE-1," *Applied Optics*, vol. 30, pp.2334-2343, June 1991.
5. K. Brenner, F. Sauer, "Diffractive-reflective optical interconnects," *Applied Optics*, vol. 27, pp. 4251-4254, October 1988.
6. J. Yeh, R. K. Kostuk, and K, Tu, "Hybrid free-space optical bus system for board-to-board interconnections," *Applied Optics*, vol. 35, pp. 6354-6364, November 1996.
7. S. Natarajan, C. Zhao, and R. T. Chen, "Bi-directional optical backplane bus for general purpose multi-processor board-to-board optoelectronic interconnects," *IEEE Journal of Lightwave Technology*, vol. 13, pp. 1031-1040, June 1995.
8. G. Kim, X. Han, and R. T. Chen, "A method for rebroadcasting signals in an optical backplane bus system," *IEEE Journal of Lightwave Technology*, vol. 19, pp. 959-965, July 2001.
9. W. J. Gambogi, A. M. Weber, and T. J. Trout, "Advances and applications of DuPont holographic photopolymers," in *Proceedings of SPIE*, vol. 2043, pp. 2-13, 1994.
10. X. Han, G. Kim, and R. T. Chen, "Accurate diffraction efficiency control for multiplexed volume holographic gratings," *Optical Engineering*, vol. 41, pp. 2799-2802, November 2002.
11. G. Kim, X. Han, and R. T. Chen, "Crosstalk and interconnection distance considerations for board-to-board optical interconnects using 2-D VCSEL and microlens array," *IEEE Photonics Technology Letters*, vol. 12, pp. 743-745, June 2000.

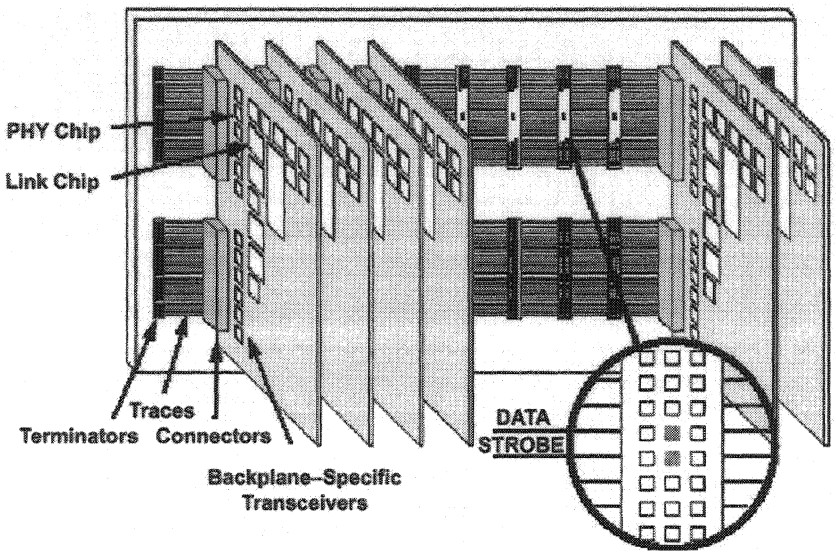


Fig. 1 Typical electrical backplane bus

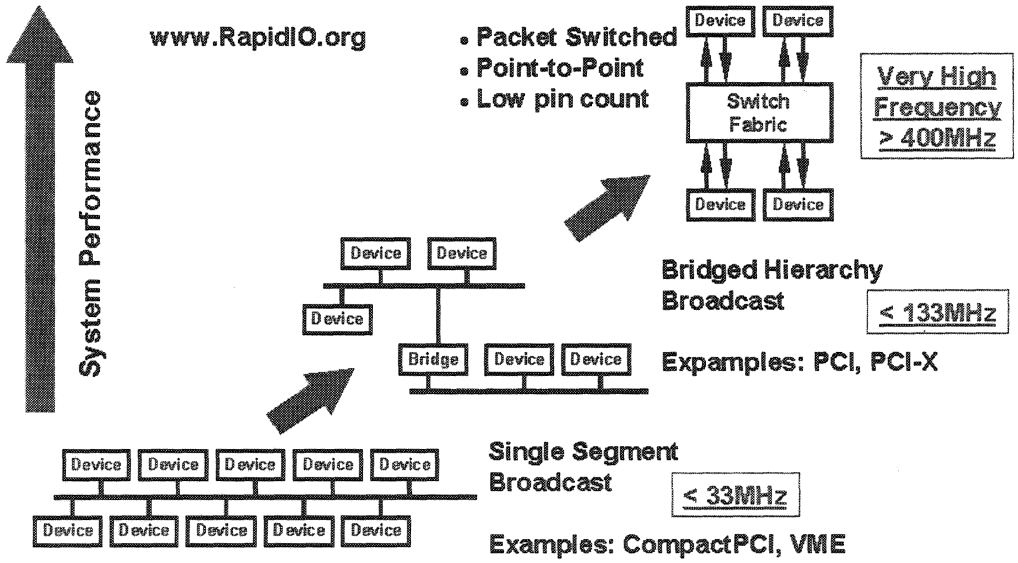


Fig. 2 System topology trend due to the limitations of electrical interconnects

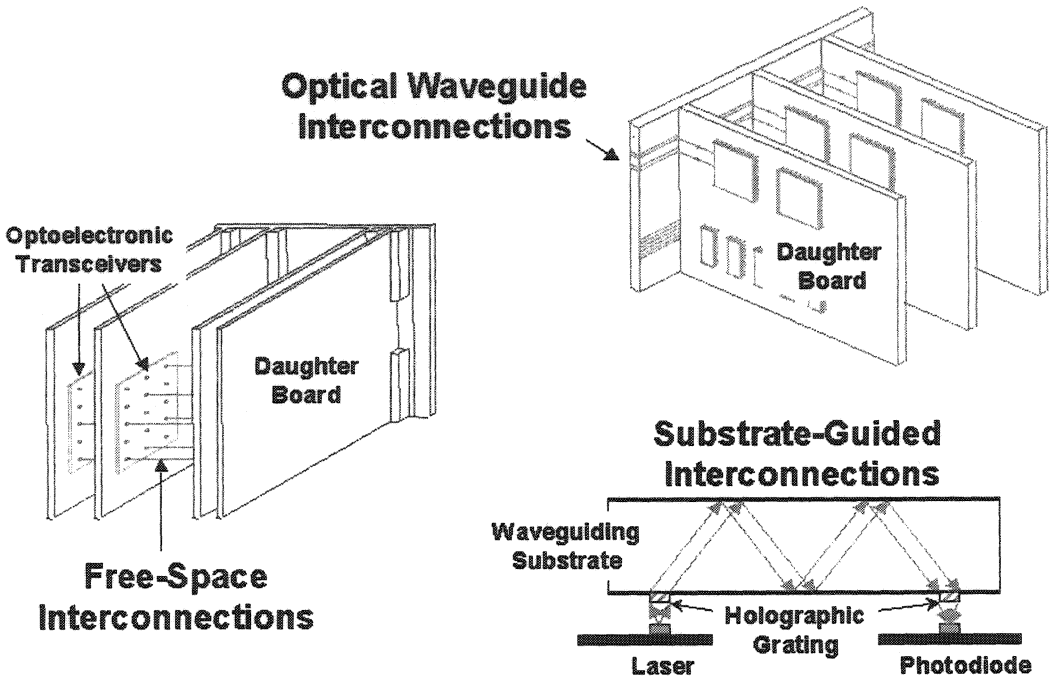


Fig. 3 Three different optical implementation methods: optical waveguide interconnections, free-space interconnections, and substrate-guided interconnections

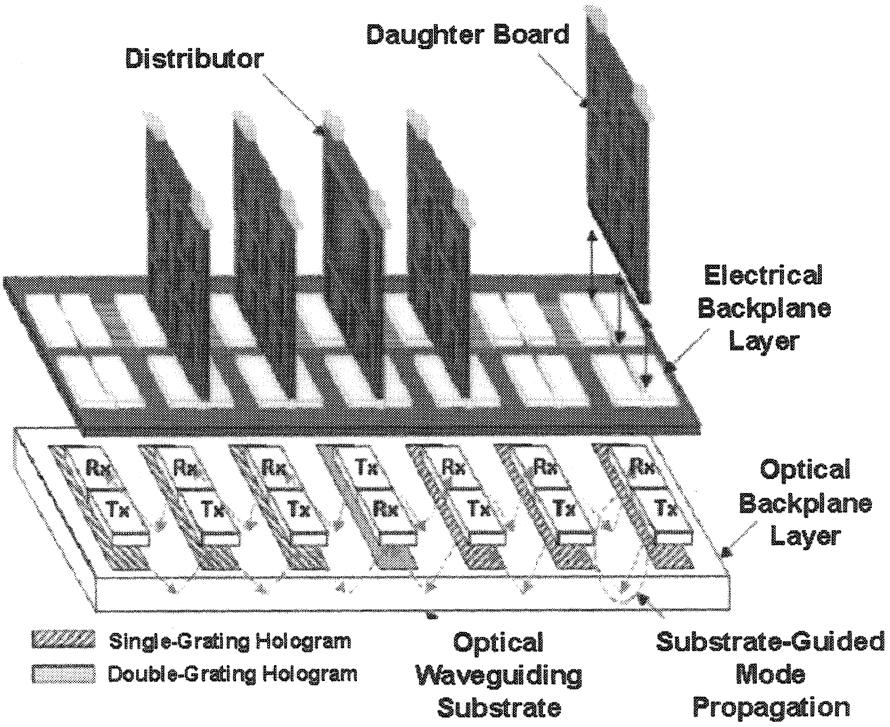
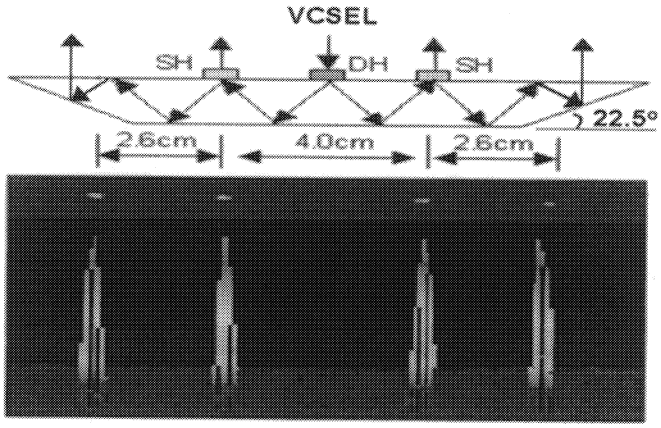
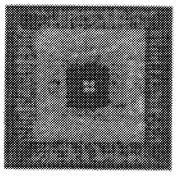


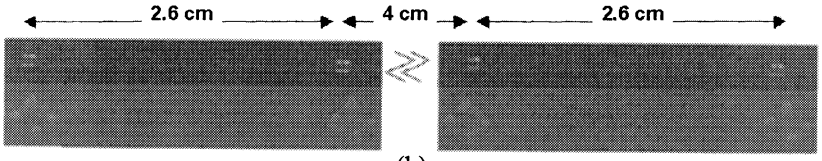
Fig. 4 Optical centralized shared-bus architecture



(a)



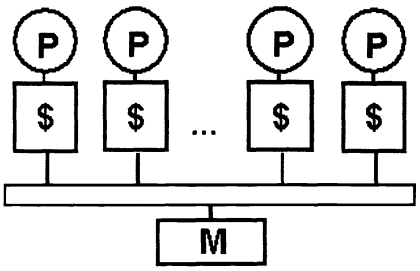
Microlens Array Packaged
VCSEL Array



(b)

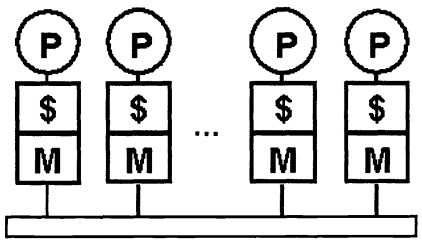
Fig. 5 Uniform optical signal fan-outs with (a) single bus line, (b) 2-D bus lines

P: Microprocessor $: Cache M: Memory



Centralized Shared Memory

(a)



Distributed Memory

(b)

Fig. 6 Two different multiprocessor model: (a) centralized shared memory multiprocessor, (b) distributed memory multiprocessor

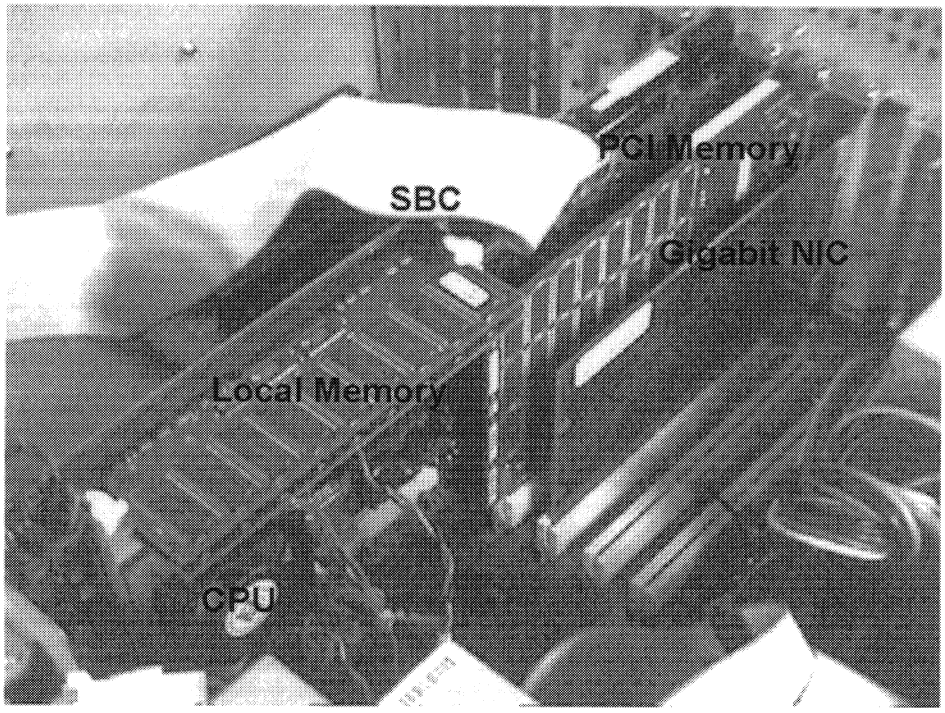


Fig. 7 PCI implementation of the centralized shared memory multiprocessor system