Improvement of Multiprocessing Performance by Using Optical Centralized Shared Bus

Xuliang Han, Ray T. Chen
Microelectronic Research Center, Department of Electrical and Computer Engineering
The University of Texas at Austin
PRC/MER 1.606G, 10100 Burnet Road, Austin, TX, USA 78758

ABSTRACT

With the ever-increasing need to solve larger and more complex problems, multiprocessing is attracting more and more research efforts. One of the challenges facing the multiprocessor designers is to fulfill in an effective manner the communications among the processes running in parallel on multiple multiprocessors. The conventional electrical backplane bus provides narrow bandwidth as restricted by the physical limitations of electrical interconnects. In the electrical domain, in order to operate at high frequency, the backplane topology has been changed from the simple shared bus to the complicated switched medium. However, the switched medium is an indirect network. It cannot support multicast/broadcast as effectively as the shared bus. Besides the additional latency of going through the intermediate switching nodes, signal routing introduces substantial delay and considerable system complexity. Alternatively, optics has been well known for its interconnect capability. Therefore, it has become imperative to investigate how to improve multiprocessing performance by utilizing optical interconnects. From the implementation standpoint, the existing optical technologies still cannot fulfill the intelligent functions that a switch fabric should provide as effectively as their electronic counterparts. Thus, an innovative optical technology that can provide sufficient bandwidth capacity, while at the same time, retaining the essential merits of the shared bus topology, is highly desirable for the multiprocessing performance improvement. In this paper, the optical centralized shared bus is proposed for use in the multiprocessing systems. This novel optical interconnect architecture not only utilizes the beneficial characteristics of optics, but also retains the desirable properties of the shared bus topology. Meanwhile, from the architecture standpoint, it fits well in the centralized shared-memory multiprocessing scheme. Therefore, a smooth migration with substantial multiprocessing performance improvement is expected. To prove the technical feasibility from the architecture standpoint, a conceptual emulation of the centralized shared-memory multiprocessing scheme is demonstrated on a generic PCI subsystem with an optical centralized shared bus.

Keywords: Optical Interconnects, Multiprocessing, Shared Bus, Switched Medium, Electro-Optical Interface, Vertical Cavity Surface-Emitting Laser (VCSEL)

1. INTRODUCTION

Interconnect is becoming an even more dominant factor in the high performance computing (HPC) systems. Electrical interconnects face numerous challenges such as signal integrity, power consumption, electromagnetic interference (EMI), and skin effect at high speed. Currently a typical electrical backplane bus operates at a frequency of less than 400MHz, whereas the speed of the state-of-the-art microprocessors has already surpassed 3GHz. This trend of computing speed outpacing interconnect capacity is becoming more and more prominent. Meanwhile, the next generation networks are envisioned to deliver beyond 10Gbps throughput to terascale grid-based applications. Therefore, a major performance bottleneck is anticipated at the board-to-board hierarchical level. Optics has been well known for its interconnect capability [1], [2]. The success of optical interconnects has already emerged at the machine-to-machine hierarchical level. To prevent the projected bottleneck from throttling the board-to-board data transfers, a new opportunity exists for the further exploitation of optical interconnects to replace the conventional electrical interconnects inside a box [3].

Photonics Packaging and Integration IV, edited by Randy A. Heyler, Ray T. Chen, Proceedings of SPIE
Vol. 5358 (SPIE, Bellingham, WA, 2004) · 0277-786X/04/$15 · doi: 10.1117/12.525754

As multiprocessing comes into the mainstream, the board-to-board interconnects become even more critical. One significant challenge in the design of a multiprocessing system is to fulfill in an effective manner the communications among several processes that are simultaneously running on multiple processors. The shared bus topology is the preferred interconnect scheme because its broadcast nature can be effectively utilized to reduce communication latency, lessen networking complexity, and support cache coherence in a multiprocessing system [4]. However, the physical length, the number of fan-outs, and the operation speed of the backplane bus impose strict limitations on electrical interconnects. Thus, the switched backplane with switch fabrics and simple point-to-point interconnections is currently being employed in the electrical domain. By changing the backplane topology from the shared bus to the switched medium, however, several crucial performance aspects are compromised. The switched medium cannot carry out broadcast as effectively as the shared bus. Besides the additional latency of going through the intermediate switching nodes, signal routing introduces substantial delay and considerable complexity, which has become a throttling factor in the high-end multiprocessing systems [5]. Therefore, an innovative optical technology that can provide sufficient bandwidth capacity, while at the same time, retaining the essential merits of the shared bus topology is highly desirable for multiprocessing performance improvement.

At the board-to-board level, optical implementation techniques can be classified into three basic categories: optical waveguide interconnects, free-space optical interconnects, and substrate-guided optical interconnect, as illustrated in Fig. 1. Similar to metal traces, optical waveguides can be laid out on a board, but essentially for point-to-point interconnects. Although free-space approaches provide some routing flexibility, the backplane topology is still constrained to point-to-point. Meanwhile, the free-space optical data links are open to the environmental noise. This problem can be avoided by confining optical signals within a waveguiding substrate. As illustrated, the optical signal that carries the data at the source is coupled into the substrate by a properly designed holographic grating. At the substrate/air interface, the incident angle is prescribed to be larger than the critical angle. Thus, this light cannot escape from the confinement of the substrate under the total internal reflection (TIR) condition [6]. At the destination, another properly designed holographic grating couples the light back into the free space to the photodiode. In this manner, an optical link from the source to the destination is established. With the appropriate design of the types of the holographic gratings and their relative positions, signal broadcast can be effectively implemented. Thus, it is possible to utilize this method to develop optical backplane buses.

As one of the most significant contributions to the efforts on optical backplane bus, an innovative architecture called optical centralized shared bus was developed [7]. To the best of our knowledge, this is the first architecture that is able to achieve equalized bus fan-outs in the optical domain. Since the critical optical/electrical interface becomes uniform across the entire backplane bus, this merit can considerably save the system power budget to maintain the required bit error rate (BER) and substantially ease the overall system integration. Based on this architecture, we propose in this paper to apply the optical centralized shared bus in the multiprocessing systems for performance improvement. After a brief overview of the centralized shared-memory multiprocessing scheme in Section 2, the architectural features of the optical centralized shared bus will be presented in the context of the centralized shared-memory multiprocessing in Section 3. A preliminary feasibility demonstration on a generic PCI subsystem will be described in Section 4. Finally, a summary is given in Section 5.

2. CENTRALIZED SHARED-MEMORY MULTIPROCESSING

In the centralized shared-memory multiprocessing model, as illustrated in Fig. 2, multiple processors share a single physical memory on a shared bus [4]. The term of shared-memory refers to the fact that the address space is shared, i.e., the same physical address on different microprocessors refers to the same location in the main memory. This shared address space can be used to communicate data implicitly via the load and store operations. The advantages of the shared-memory communications mainly include [4]:

- Ease of programming when the communication patterns among the microprocessors are complex or vary dynamically during execution. Also, this advantage simplifies the compiler design.
- Lower communication overhead and better use of the available bandwidth. This arises from the implicit nature of communication and the use of memory mapping to implement protection in hardware rather than through the operating system.

- Capability of automatic caching of all data, both shared and private. Caching provides both decreased latency and reduced contention for accessing the shared data, and thus the frequency of the remote communications can be minimized.

Caching is a widely applied technique to improve system performance by utilizing the locality of the programs. The centralized shared-memory multiprocessing model supports automatic caching of all data, both shared and private. The private data are accessible only for a single processor, while the shared data for multiple processors. The communications among the processors are essentially carried out via the read and write operations upon the shared data. When a private item is cached, its location is migrated to the cache, reducing the average access time as well as the memory bandwidth required. Since no other processors use the private data, the program behavior is identical to that in a uniprocessing system. When a shared item is cached, the shared value may be replicated in multiple caches. In addition to the reduction in the access latency and required memory bandwidth, this replication also lessens the contentions that may exist for the shared items that are being accessed by multiple processors at the same time. The shared bus topology plays a pivotal role for the correct functioning of the centralized shared-memory multiprocessing scheme. To ensure the consistency of the shared memory seen by each processor, all caches must be retained in coherence. Since only one processor can deliver data on the shared bus at a time, the write operations upon the shared memory are forced in a sequential order. In the centralized shared-memory multiprocessing scheme, this property is called write serialization [4]. Meanwhile, all cache controllers snoop on the shared bus that carries all actual data exchanges. Because of the broadcast nature of the shared bus topology, all processors can simultaneously monitor every access to the shared memory, and quickly determine whether or not they have a cached copy of the item being transferred. Accordingly, the cached copies may be either invalidated or updated with the detected new value. In this manner, cache coherence is consistently maintained across the whole system.

As expected, the performance of a centralized shared-memory multiprocessing system is critically dependent on the shared bus that carries out all inter-processor communications and broadcast actions for maintaining cache coherence. In the electrical domain, the anticipated performance bottleneck essentially originates from the restricted bandwidth capacity of the shared bus. The physical length, the number of fan-outs, and the operation speed of the backplane bus are significantly restricted by the physical limitations of electrical interconnects. Thus, the switched backplane with switch fabrics and simple point-to-point interconnections is currently being employed in the high-end multiprocessing systems. By changing the backplane topology from the shared bus to the switched medium, however, several crucial performance aspects, e.g., interconnect latency, are compromised. The switched medium is an indirect network, and thus cannot carry out broadcast as effectively as the shared bus. Besides the additional latency of going through the intermediate switching nodes, signal routing introduces substantial delay and considerable system complexity. In Ref. [5], the statistics of the memory read latency in a medium and a large size switch-based multiprocessing system shows that the wire delay is only a moderate fraction of the total memory read latency, in contrast, the transactions through switches and the multicast/broadcast actions to maintain cache coherence are a significant fraction, furthermore, the delay associated with switching and cache coherence increases with the system scale more rapidly than the wire delay. Meanwhile, the additional involvement of many expensive devices, such as switch fabrics and transceiver modules, tremendously increases the overall system cost. In consequence, there would be few prominent benefits if simply replacing electrical wires with optical point-to-point interconnections. Without all-optical switching, the additional optical domain overhead, i.e., the optical-to-electrical and electrical-to-optical conversions at the interface of the switch fabric, could even make worse the latency problem in the switch-based multiprocessing systems. Therefore, an innovative optical technology that can provide sufficient bandwidth capacity, while at the same time, retaining the essential merits of the shared bus topology is highly desirable for the performance improvement in a centralized shared-memory multiprocessing system.

3. OPTICAL CENTRALIZED SHARED BUS ARCHITECTURE

Fig. 3 illustrates the architectural concept of the optical centralized shared bus [7]. For simplicity, only five slots (#A1, #A2, #B1, #B2, and #C) are drawn in this schematic. In the context of the centralized shared-memory multiprocessing, a memory board is to be inserted into the central slot (#C), while the other slots (#A1, #A2, #B1, and #B2) on the backplane bus are for processor boards. The electrical backplane provides interconnects for the non-critical paths. The electro-optical transceivers, including VCSELs (vertical-cavity surface-emitting lasers) and photodiodes, are integrated at the bottom of the electrical backplane, and aligned with the underlying optical interconnect layer. Therefore, the insertion/removal of the boards during the normal operations does not affect the critical alignment. Different from other modules, the positions of the central VCSEL and photodiode are swapped as indicated in Fig. 3. The configured optical interconnect layer consists of a waveguiding plate with the properly designed volume holographic gratings integrated on its top surface. The plate provides a turbulence-free medium for optical interconnects, and the waveguide holograms function as optical fan-in/fan-out devices. Underlying the central slot (#C) is an equal-efficiency double-grating hologram, while the others are single-grating holograms. By employing such a unique configuration, both broadcastability and bi-directionality of signal flows on the backplane bus are enabled [7], which are the essential obstacles to achieving equalized bus fan-outs in the other optical shared bus architectures [8], [9].

The optical centralized shared bus well fits in the centralized shared-memory multiprocessing scheme from the architecture standpoint. The memory integrated on the central board functions as the centralized shared memory. For a write operation upon the shared memory, as illustrated in Fig. 3, the VCSEL of the source processor board emits the light that carries the data and projects it surface-normally onto its underlying waveguide hologram. This light is coupled into the optical waveguiding plate by the grating and propagates within the confinement of the plate under the total internal reflection (TIR) condition [6]. Then, it is surface-normally coupled out of the plate by the central double-grating hologram and detected by the central photodiode. Because there is only one photodiode inside the central electro-optical transceiver module, the data deliveries from the processor boards on the backplane to the memory board are forced in a sequential order as in the centralized shared-memory multiprocessing scheme, i.e., write serialization. Subsequently, cache coherence is ensured in a simple broadcast fashion. The central VCSEL generates the outbound optical signal that carries the updated data and projects it surface-normally onto its underlying double-grating hologram. This light is coupled into the plate and equally diffracted into two beams by the hologram, propagating along the two opposite directions within the confinement of the plate under the total internal reflection (TIR) condition [6]. During the propagation, a portion of the light is surface-normally coupled out of the plate by the single-grating hologram underlying each processor board on the backplane and detected by the photodiode. By snooping on the shared bus, all processor boards can immediately obtain the updated data from the centralized shared memory, and then either invalidate or update the cached copies. In this manner, cache coherence is consistently maintained across the whole system.

The volume holographic gratings integrated on the top surface of the waveguiding plate function as optical fan-in/fan-out devices. Their diffraction properties in the Bragg regime can be analyzed with Kogelnik’s Coupled Wave Theory [10]. By balancing the diffraction efficiency of the waveguide holograms in use, the bus fan-outs across the entire optical interconnect layer can be equalized as demonstrated in Fig. 4 [11]. This merit is highly desirable from the system integration standpoint because of the reduced constraint on the dynamic ranges of the electro-optical transceiver modules in use. Compared with electrical interconnects, the most significant benefit of optical interconnects is the tremendous gain in the bandwidth capacity. In Ref. [12], the bandwidth capacity per substrate-guided optical line was experimentally characterized to be approximately 2.5THz. With such an enormous bus bandwidth, the interconnect bottleneck in the centralized shared-memory multiprocessing scheme would be completely eliminated by employing the optical centralized shared bus.

4. DEMONSTRATION ON PCI SUBSYSTEM

The real implementation of the centralized shared-memory multiprocessing scheme by employing the optical centralized shared bus certainly involves too many processor-specific issues. Meanwhile, the advanced microprocessors are upgrading at a rapid pace, probably with different micro-architectures from one generation to the next. With the focus on demonstrating the technical feasibility in a general scenario, a conceptual emulation of the centralized shared-memory scheme was carried out on a generic PCI subsystem that incorporated an optical centralized shared bus.

PCI stands for Peripheral Components Interconnect. It defines a local bus architecture that is not specific to any particular processor [13]. A processor is connected to the root PCI bus through a compatible chipset, which is frequently referred to as the North Bridge. The use of the North Bridge isolates the generic PCI local bus from the specific processor bus. There are two participants in every PCI data transaction: master, or called initiator, and target.

The master is the device that initiates the data transfer, and the target is the device that is addressed by the master for the purpose of performing the data transfer. It is very important to note that the PCI data transfers can be accomplished in the burst mode [14]. A burst transfer consists of a single address phase followed by two or more data phases, and the master has to arbitrate for the bus ownership only one time for the whole block of the data to be transferred. Thus, the arbitration overhead is largely reduced, and the available bus bandwidth may be fully utilized for the actual data transfers. During the address phase, the start address and transaction type are issued in a broadcast fashion on the shared bus. The targeted device latches the start address into an address counter, claims the transaction, and is responsible for incrementing the address from one data phase to the next. As the master is ready to transfer each data item, it informs the target whether or not it is the last one, and the entire PCI burst transaction completes when the final data item has been transferred.

The centralized shared-memory multiprocessing scheme was conceptually emulated on a generic PCI subsystem as shown in Fig. 5. The electrical part of this system is consisted of a passive PCI backplane, a single board computer (SBC) card, a Gigabit Ethernet Network Interface Card (NIC), and a PCI memory card. As illustrated by the connectivity diagram in Fig. 5 (b), the PCI memory card function as the centralized shared memory as in the centralized shared-memory multiprocessing scheme. The SBC card contains a 1.2GHz microprocessor and a North Bridge that controls the interface to the PCI backplane. The microprocessor can access to the PCI memory card through the North Bridge. The NIC card is connected to another workstation through a RJ-45 crossover cable. With the capability to request for the bus ownership, the NIC card can initiate PCI data transactions targeting the PCI memory card without any CPU actions on the SBC card. Thus, the communications between the NIC card and the SBC card can proceed via the PCI memory card on the shared bus in a conceptually equivalent manner to the shared-memory communications among multiple processors in a centralized shared-memory multiprocessing system.

As shown in Fig. 6, the optical centralized shared bus was integrated underneath the PCI backplane, where the equalized bus fan-outs were established across the entire optical interconnect layer. As a preliminary attempt, only PCI bus line AD02 was replaced by the optical interconnection link while the other electrical wires on the passive PCI backplane were remained. In order to incorporate the optical interconnect layer into the generic PCI subsystem, a special extension interface was developed, as shown in Fig. 7, to be integrated with the AD02 pins of the PCI slots on the backplane. It contains an electro-optical transceiver module and the required logic controls in consistent with the PCI protocol. A single electrical PCI bus line carries bi-directional signal transmissions. Meanwhile, the commercial PCI core does not explicitly indicate the actual data transfer direction. The data transaction type, either read or write, is negotiated between the master and the target in an implicit manner involving several PCI bus signals [14]. Thus, a logic-interpreting circuit was implemented, as illustrated in Fig. 7 (b), to generate the RACTIVE and TACTIVE control signal, as illustrated in Fig. 7 (a), to appropriately coordinate the operations of the electro-optical transceiver modules during the PCI data transfers.

From the operating system standpoint, the PCI memory card was actually treated as a RAM disk after mounting a file system. To conceptually emulate the shared-memory communications among multiple processors in a centralized shared-memory multiprocessing system, the same file was transferred from the NIC card to the PCI memory card, and then from the PCI memory card to the SBC card. On the shared bus, the signal waveforms during the PCI data transfers were captured in the real time by the bus analyzer card as shown in Fig. 5 (a). This bus analyzer card was connected to a logic analyzer (HP1660ES) for the logic timing verification. In particular, the signal waveforms presented at the AD02 pins of the NIC card and the PCI memory card were displayed on an oscilloscope for the direct visualization of the implemented optical interconnection during the PCI data transfers. Fig. 8 is one of the captured results during such tests, where Channel 1 displays the signal waveforms at the AD02 pin of the NIC card, which were the modulation inputs to the VCSEL driver inside its extension interface module, and Channel 2 displays the signal waveforms at the AD02 pin of the PCI memory card, which were the outputs from the edge detector inside its extension interface module. The obtained results of these tests verified the correct connectivity of the implemented optical interconnection link.

5. CONCLUSION

The optical centralized shared bus utilizes the enormous bandwidth capacity of substrate-guided optical interconnects, while at the same time, retaining the desirable architectural features of the shared bus. Its unique topological configuration enables the fulfillment of equalized optical bus fan-outs across the entire architecture, and thus a uniform electrical/optical interface can be obtained. This significant achievement is highly desirable from the system integration standpoint. Meanwhile, it is particularly pointed out in this paper that this innovative architecture well fits in the centralized shared-memory multiprocessing scheme. As a preliminary attempt, a conceptual emulation of the centralized shared-memory multiprocessing scheme was carried out on a generic PCI subsystem that incorporated an optical centralized shared bus. Since this research prototype originated from the existing system, the actual data transfers were still at the same standard PCI bus speed (33MHz) as without using optical interconnects. Apparently, the compromise is that it cannot exhibit any performance improvement since the terahertz bandwidth potential of optics is not utilized in the constructed prototype. Nonetheless, the objective of the demonstration presented herein is to prove the technical feasibility from the architecture standpoint. Because there is no doubt on the interconnect capability of optics, which has been confirmed both theoretically and experimentally, it can be projected for sure that the interconnect bottleneck of the shared bus in the centralized shared-memory multiprocessing system would be completely eliminated by employing the optical centralized shared bus.

ACKNOWLEDGEMENT

The authors would like to thank BMDO, DARPA, ONR, AFOSR, and the ATP program of the State of Texas for supporting this work.

REFERENCES

1. 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.
2. E. D. Kyriakis-Bitzaros, N. Haralabidis, M. Lagadas, A. Georgakilas, Y. Moisiadis, and G. Halkias, “Realistic end-to-end simulation of the optoelectronic links and comparison with the electrical interconnections for system-on-chip applications,” *IEEE Journal of Lightwave Technology*, vol. 19, pp. 1532-1542, October 2001.
3. A. F. J. Levi, “Optical interconnects in systems,” *Proceedings of the IEEE*, vol. 88, pp. 750-757, June 2000.
4. D. A. Patterson, J. L. Hennessy, “Computer architecture: a quantitative approach,” 2nd Edition, Chapter 8, Morgan Kaufmann Publishers, August 1995.
5. D. Huang, T. Sze, A. Landin, R. Lytel, and H. Davidson, “Optical interconnects: out of the box forever?” *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 9, pp. 614-623, March/April 2003.
6. K. Brenner, F. Sauer, “Diffractive-reflective optical interconnects,” *Applied Optics*, vol. 27, pp. 4251-4254, October 1988.
7. X. Han, G. Kim, G. J. Lipovski, and R. T. Chen, “An optical centralized shared-bus architecture demonstrator for microprocessor-to-memory interconnects,” *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 9, pp. 512-517, March/April 2003.
8. 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.
9. 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.
10. H. Kogelnik, “Coupled wave theory for thick hologram gratings,” *The Bell System Technical Journal*, vol. 48, pp. 2909-2947, November 1969.
11. 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.
12. G. Kim, R. T. Chen, “Three-dimensionally interconnected bi-directional optical Backplane,” *IEEE Photonics Technology Letters*, vol. 11, pp. 880-882, July 1999.
13. “PCI local bus specification,” Revision 2.2, December 1998.
14. T. Shanley, D. Anderson, “PCI system architecture,” 4th Edition, Addison-Wesley Longman, August 1999.

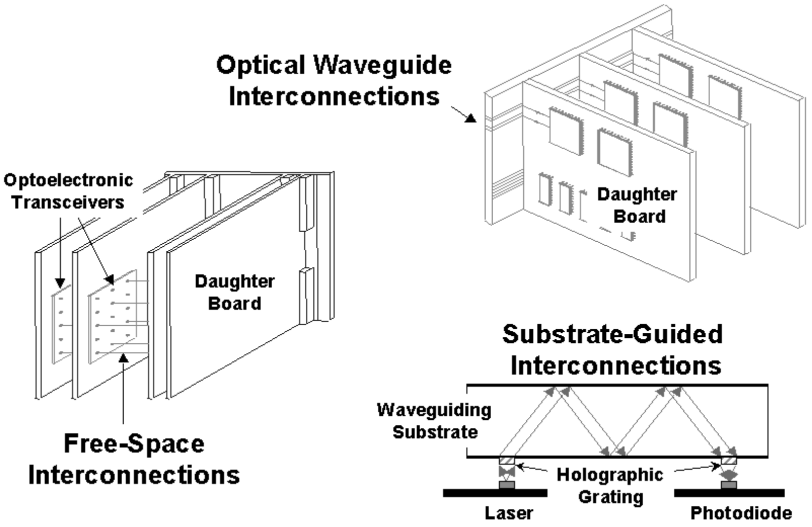

Optical Waveguide Interconnections
Optoelectronic Transceivers
Daughter Board
Daughter Board
Substrate-Guided Interconnections
Waveguiding Substrate
Holographic Grating
Laser
Photodiode
Free-Space Interconnections

Fig. 1 Three Basic Optical Interconnect Methodologies

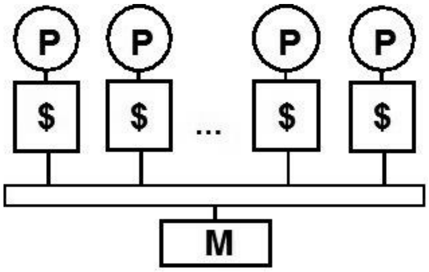


Fig. 2 Centralized Shared-Memory Multiprocessing System (P: Processor, $: Cache, M: Centralized Shared Memory)

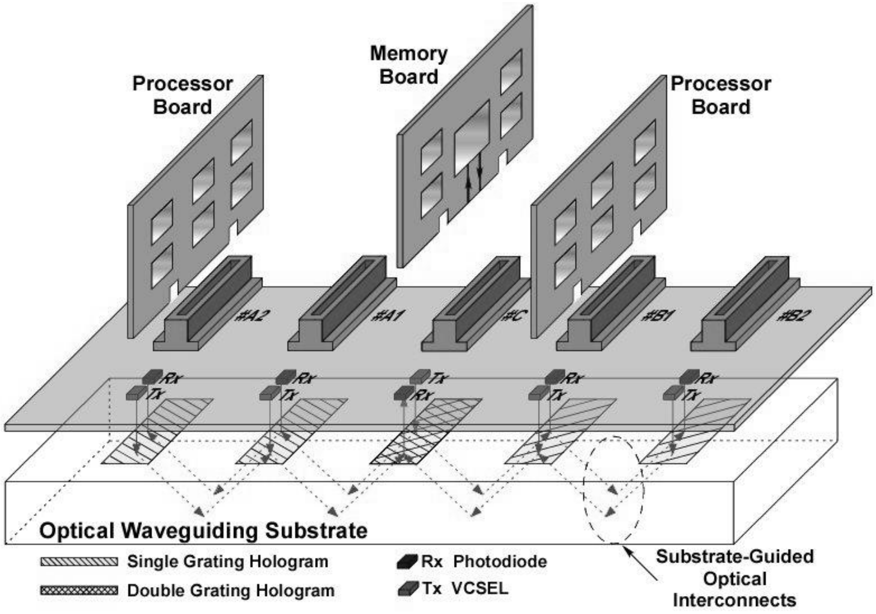


Fig. 3 Optical Centralized Shared Bus Architecture

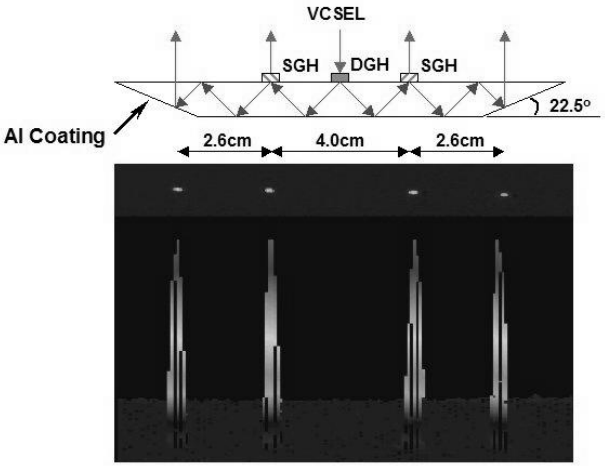
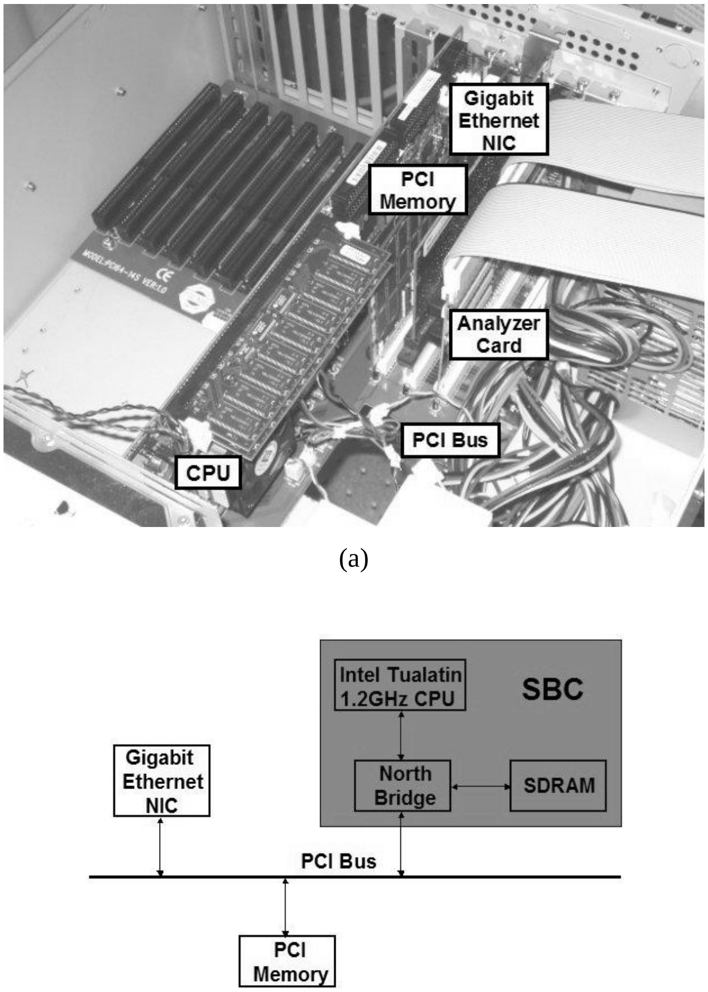


Fig. 4 Demonstration of Equalized Optical Signal Fan-Outs



(a)

(b)

Fig. 5 Centralized Shared-Memory Multiprocessing on PCI Subsystem

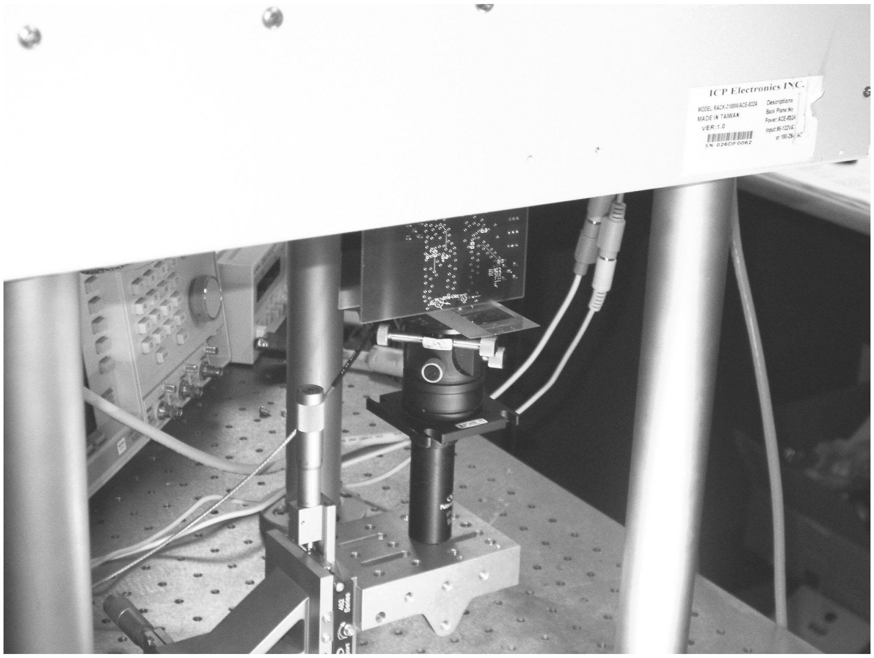


Fig. 6 Optical Centralized Shared Bus as Optical PCI Bus Line

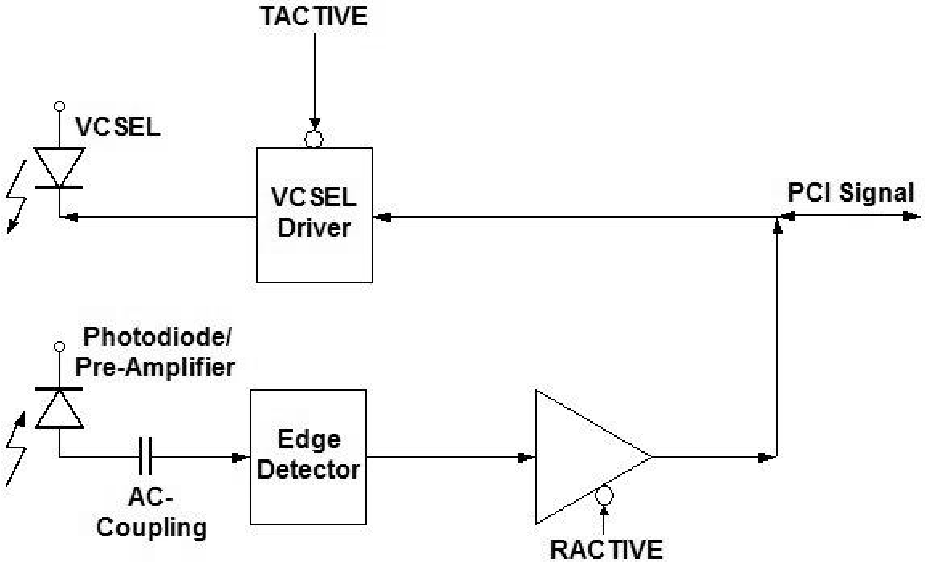


Fig. 7 (a) PCI Electro-Optical Interface Module

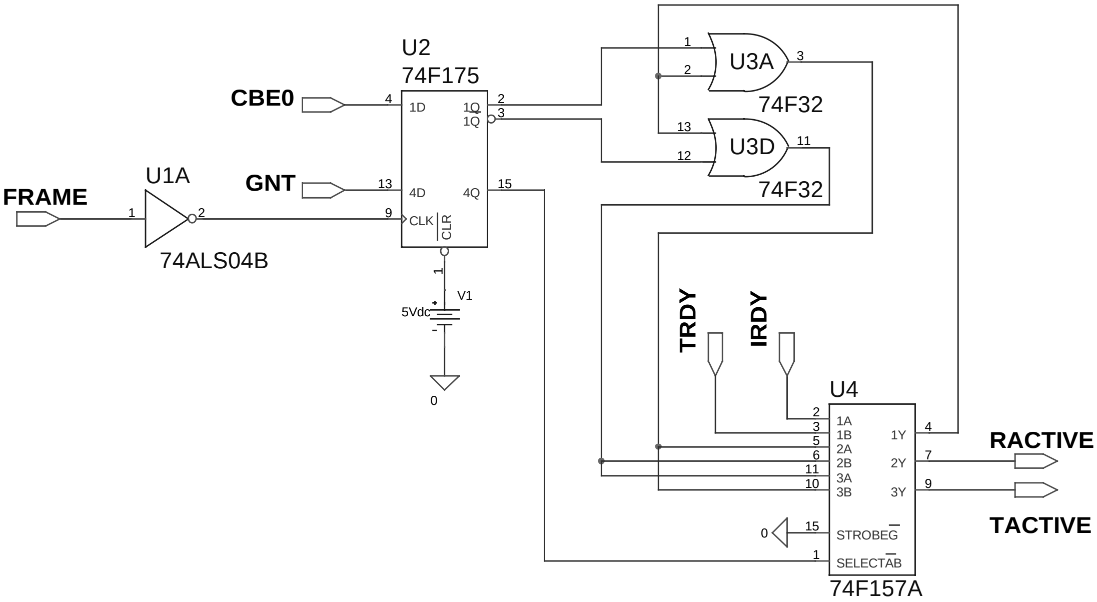
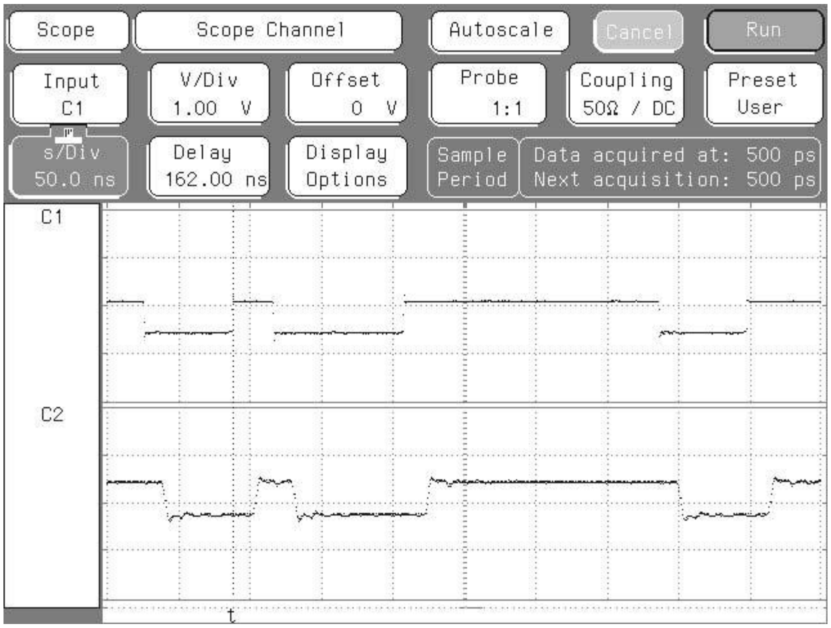


Fig. 7 (b) PCI Extension Interface Logic Generation



Input C1
V/Div 1.00 V
Offset 0 V
Probe 1:1
Coupling 50Ω / DC
Preset User
s/Div 50.0 ns
Delay 162.00 ns
Sample Period
Data acquired at: 500 ps
Next acquisition: 500 ps

Fig. 8 Optical Interconnection Link of PCI Bus Line AD02