An Optical switch Network for Multiprocessor Cluster System of 128 Nodes

Mingcui Cao, Zhixiang Luo, Fengguang Luo, Xinjun Zhou, Jun Xu

National Lab. of Laser Technology,
Huazhong University of Science and Technology,
1037 Luo Yu Road, Wuhan 430074 P. R. China

ABSTRACT

A novel hybrid electrical optical Clos switch network for multiprocessor cluster system was presented. For multiprocessor cluster system of 128 hosts , the novel optical Clos network includes 16 basic modules, a passive optical fiber backplane with (8 x 15) x 16 which has a total of 1920 optical data channels and a signaling control system. The basic module is composed of the input line cards of 8 hosts, a single chip of 16 x 16 crossbar switch, parallel transmitting VCSEL modules for fan-out of (16-1) optical fiber channels and (16-1) x 1 optical combiners. The passive optical fiber backplane of very large capacity and high density, based on linear VCSEL arrays and fiber ribbon technology, is to be used to interconnect between hosts of different sub-clusters. The routing of the optical Clos switch network is decided by a signaling control system. Compared with high performance electronic system, this technology offers a relatively easy and simple means of communicating large amount of information between hosts, and lower delay time.

Keywords: Optical interconnection network, Parallel optical links, Computer, Cluster, Parallel processing, VCSEL

1. INTRODUCTION

The tremendous increase in the speed of data transport on optical fiber network has stimulated a high demand for gigabit multimedia services such as distance learning and video conferencing. As data traffic increases exponentially in the Internet, a need of deploying IP routers with terabit/s capacity and high performance multiprocessor cluster system for Internet service are emerging [1]. It is widely recognized that IP routers with terabit/s capacity and the multiprocessor cluster system for Internet service must overcome the limitations of traditional metallic buses, backplanes and electrical interconnection network to achieve high overall

Active and Passive Optical Components for WDM Communications III, Achyut K. Dutta, Abdul Ahad S. Awwal, Niloy K. Dutta, Kazuo Fujiura, Proceedings of SPIE Vol. 5246 (2003)

performance for the requirements of the exponential growth of multimedia traffic. The ongoing multimedia trend and the required communication amount of the modern information society requires enormous performance of computer networks as well as of the electronic equipment. These demands force semiconductor industry to develop microprocessors operating with significantly higher clock-rates. Microprocessors of the next generation will provide on-chip clock frequencies above 1 GHz and at least 64-bit architectures. In the near future as in the next 10-year, the on-chip clock-rates of about 10GHz seems to be possible in 2011. [2] Taking into account the above mentioned, by combining the strength of both optical and electronic technologies, a high-bandwidth, highly capable optical interconnection technology on board- and system level will be a key technology for future high-performance information and communication equipment[3-6]. Several novel optical interconnection technologies have been presented such as the multi-layer PC board with optical interconnects, the passive optical backplane, and so on. Parallel optical interconnection of the passive optical backplane, based on linear VCSEL arrays and fiber ribbon technology, have great potential for use as high bandwidth interconnects over short distances, such as the board to the board and shelf to shelf level. They offer compact, high bandwidth with low power requirements. It have had increasing commercial acceptance in recent.

This paper presents a novel hybrid electrical optical Clos switch network based on a passive optical fiber backplane for multiprocessor cluster system. The passive optical fiber backplane of very large capacity and high density, based on linear VCSEL arrays and fiber ribbon technology, is to be used to interconnect between hosts of different sub-clusters. For multiprocessor cluster system of 128 hosts, it includes 16 basic modules, an optical fiber interconnection backplane with (8 x 15) x 16 optical data channels and a signaling control system. The routing of the optical Clos switch network is decided by a singling control system. The format of this paper is as follows: Section 2 describes the system configuration of Clos network with a large capacity passive optical backplane. In section 3, the singling control system is presented. Finally, the performances of hybrid optical electrical Clos switch network are discussed in section 4.

2. SYSTEM CONFIGURATION OF THE CLOS NETWORK WITH A LARGE CAPACITY PASSIVE OPTICAL BACKPLANE

Clos network are named for Charles Clos, who introduced them in a paper titled "A Study of Non-Blocking Switching Network," published in the Bell System Technical Journal in March 1953. The Clos network topology is a full-bisection Clos network, it has excellent properties-scaling to large size, modularity, and multiple-path

redundancy that make it an ideal topology for cluster networks. The topology structure of Clos network for multiprocessor cluster system of a 128-host is shown in Fig.1. It is composed of 16 sub-clusters which has 8 hosts, as illustrated in the lower row. This topology provides so many interconnection paths between hosts. There is a unique shortest route to connect hosts in same sub-clusters, which is implemented on a single chip of 16 x 16 crossbar switch. The traffic between hosts of different sub-clusters through the Clos spreader spine network, as shown in the upper row, which may be implemented by use two techniques of an active backplane and a passive optical backplane. A known technique of the active backplane had been described in greater detail in the products of Myrinet-2000 Switches, in which the 8 spine Xbar 16s would be used. The Xbar 16 is build on a single chip of 16 x 16 crossbar switch[7]. It was pointed out in above-mentioned, the traffic between hosts of different sub-clusters should be traversed through three times of Xbar 16 switches in Clos network with known technique of the active backplane. It has been result in more complication of the control system and increasing of delay time.

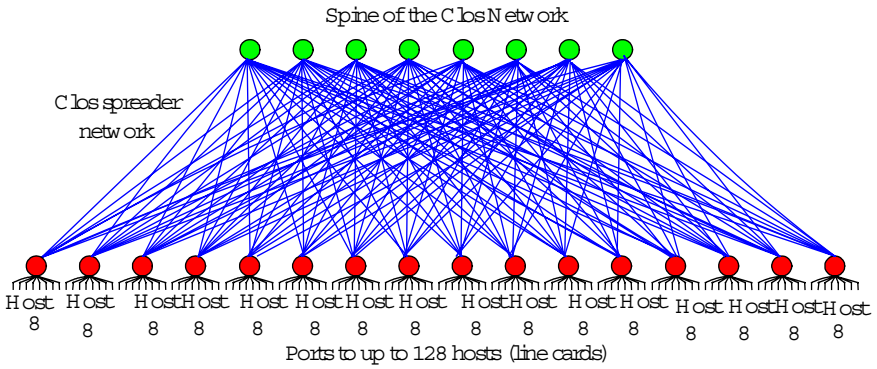

Fig.1 128-hosts Clos network

In order to overcome the above problem, a passive optical backplane of very large capacity and high density, based on linear VCSEL arrays and fiber ribbon technology, instead of an active backplane is to be used for Clos network of multiprocessor cluster system for high-speed transmission. For the convenience of description, taking Clos network of 32-hosts as an example. There are four sub-clusters which has 8-hosts in multiprocessor cluster system of 32-hosts. The traffic between 8 hosts in same sub-clusters is implemented by use of a single chip of 16 x 16 crossbar switch, in which 8 inputs and 8 outputs are used for routing of 8 hosts in same sub-clusters, the other 8 inputs/outputs are to be used for routing from coming hosts in difference sub-clusters to 8 hosts of this sub-cluster. A passive optical fiber backplane provide interconnection paths for the hosts between difference sub-clusters, as shown in Fig. 2. A hybrid optical electrical Clos switch network of 32-hosts is composed of four modules, a passive optical fiber backplane and a singling control system, as as illustrated in Fig. 3. The input line cards

of 8 hosts, a single chip of 16 x 16 crossbar switch, parallel transmitting VCSEL modules for fan-out of (4-1) optical fiber channels and (4-1) x 1 optical combiners are packaged on a module. In order for communication between hosts of difference sub-clusters, a passive optical backplane of high bandwidth optical fiber interconnection network provide high speed data channels for all over 32-hosts. The input port of each host has fan-out of three optical fiber channels to connect into other three modules respectively, as shown in Fig. 3. Parallel transmitting VCSEL module provide the interface of fan-out of three optical fiber channels, in which transmitting optical signal of each cell of VCSEL module can be controlled by a single chip of 1 x (4-1), that means, it is decided to connect the host which module is belonged in. If the connected host belongs to module 3, the third cell of VCSEL transmitting module is controlled to send optical signal, the optical signal is sent to the third module with third fiber. Each module has (4-1) x 8 optical channels coming from hosts of other sub-clusters. They are divided into 8 groups, in which there are three optical channels in each group. The first group is coming from number 1 host of other sub-clusters respectively, The second group is coming from number 2 host of other sub-clusters respectively, and so on. Then the optical channels of each group are recombined into one optical channel respectively in (4-1) x 1 combiners and sent to the eight output of a 16 x 16 switch chip. The above described will show that very high-speed optical data channels in the passive optical backplane are provided for the interconnecting communication between hosts of difference sub-clusters. The routing election and prevention of congestion are decided by a set of simple signaling system, it is described in section 3.

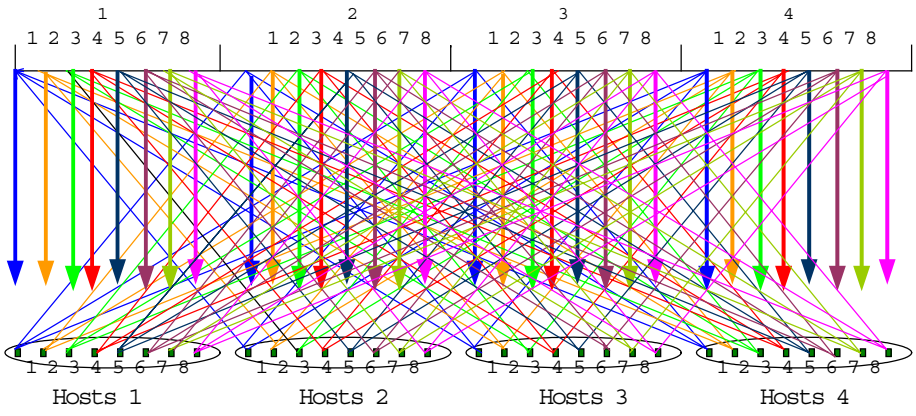

Fig.2, The passive optical fiber backplane for Cluster of 32-hosts

This hybrid optical electrical Clos switch network, as mentioned above, can be applied to a Clos switching network for multiprocessor cluster system of 128-hosts. The Clos switching network is composed of 16 modules, a passive optical fiber interconnection backplane and a signaling control system. There are the input line

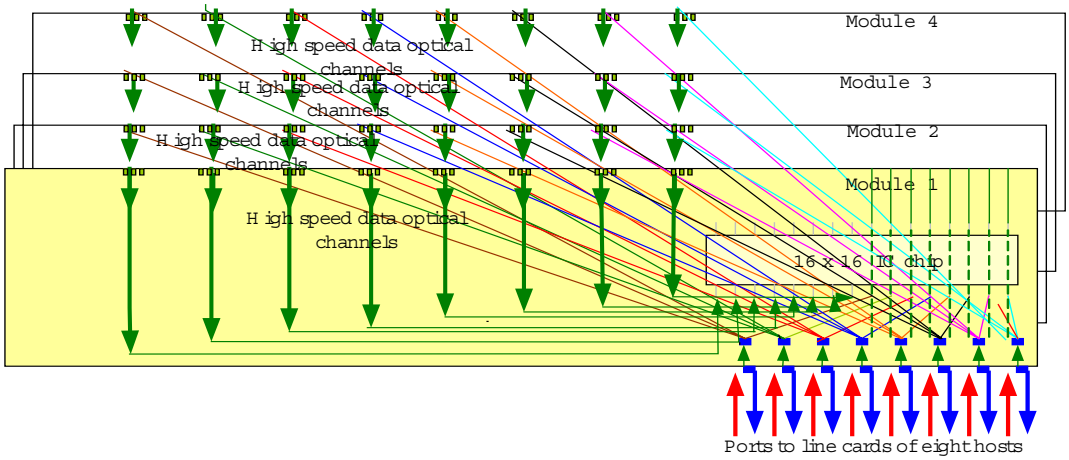

Fig.3. A hybrid electrical optical Clos switch network for multiprocessor cluster system of 32-hosts

cards of 8 hosts, a single chip of 16 x 16 crossbar switch, parallel transmitting VCSEL modules for fan-out of (16-1) optical fiber channels and (16-1) x 1 optical combiners on a module. The input port of each host has fan-out of 15 passive optical fiber channels to connect into other 15 modules respectively. An optical fiber interconnection backplane with (8 x 15) x 16 which a total of 1920 data channels is to be used for the communication between the hosts in difference sub-clusters of this system, as shown in Fig. 4. The high-speed optical data channels in the passive optical backplane are provided for the interconnecting communication between hosts of difference sub-clusters for 128 hosts. The routing election and prevention of congestion are decided by a set of simple signaling system, it is described in section 3. Compared with high performance electronic system, this technology offers a relatively easy and simple means of communicating large amount of information between hosts and lower delay time.

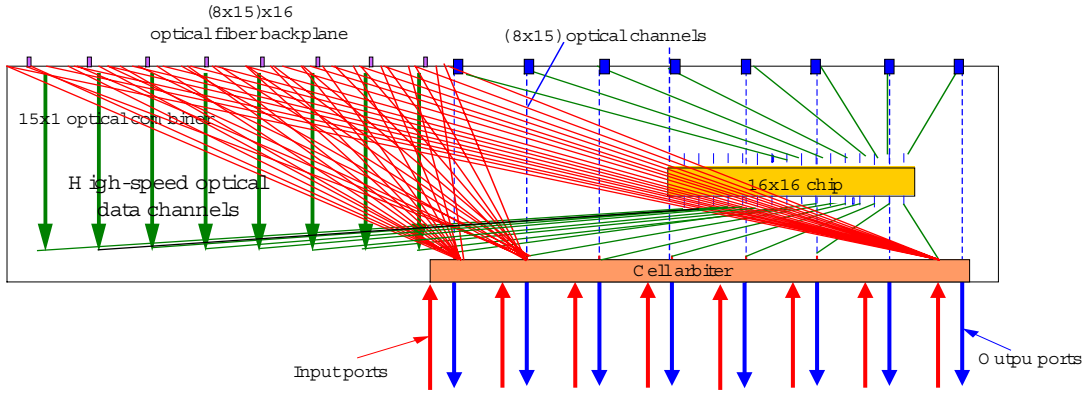

Fig.4. The architecture of a module in hybrid optical electrical Clos network for 128 hosts

3. SIGNALING CONTROL SYSTEM OF HYBRID OPTICAL ELECTRICAL CLOS SWITCH NETWORK

Communication protocol of a Clos switching network for multiprocessor cluster system can be user-defined. For the all optical switching transport network, different communication protocols can be adopted, such as 1) the out-of-band control communication protocol; 2) the in-band control communication protocol.

The architecture of the computer clusters comprising 4 groups and 8 hosts in each group is shown in Fig. 5.

Each group consists of interface processing unit(IPU), 16 x 16 crossbar IC chip and intra-cluster routing and controlling unit(IntraRCU) and inter-cluster routing and controlling unit(InterRCU). Signal messages are transferred between IntraRCU and InterRCU, and then paths are selected. The optical fibers are used to transfer high speed data to the other 3 clusters.

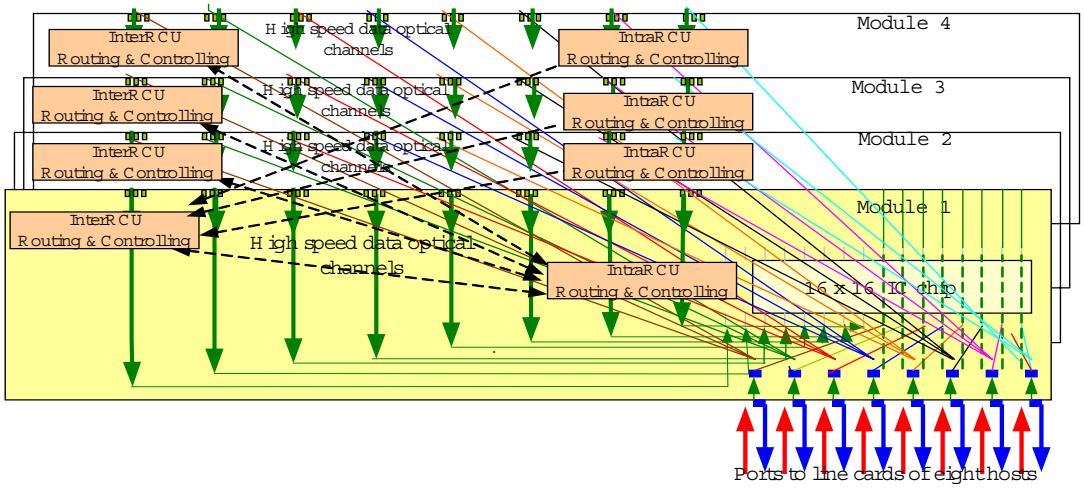

Fig.5. signaling control system of hybrid optical electrical Clos switch network

The format of the signaling message is defined as follows:

Message Format

Type	Class	Flag	Resource Address	Destination Address

Type: 1bit—"0" denotes "Request Message"
"1" denotes "Response Message"
Flag: 1bit—"0" denotes "Request Denied"
"1" denotes "Request Accepted"
Resource/Destination Address: 1Byte, 1 bit denotes the destination host in or out of the local computer group, the other 7 bit denote the 128 hosts.

The block diagram of IntraRCU is shown in fig.6. IntraRCU takes charge of processing the routing messages from the local 8 hosts and the other 8 hosts coming

from other clusters and outlying the contenders. For every host, if routing is successful, IntraRCU will feed back a routing successful signal to the IPU, otherwise it will feed back routing failure signal to the interface handle unit.

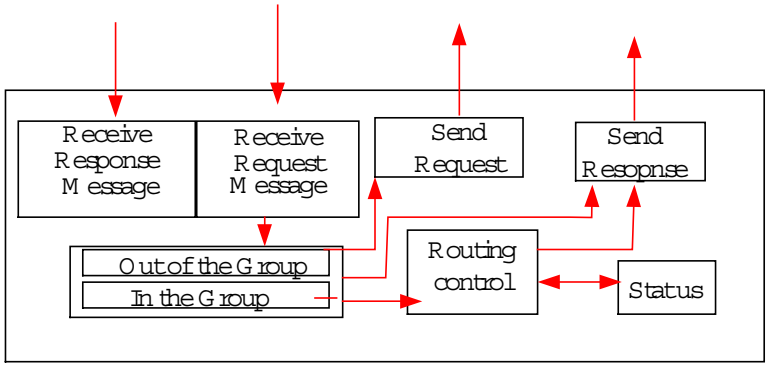

Fig.6. The block diagram of the IntraRCU

Fig.7 InterRCU mainly takes charge of managing routing message come from other 3 clusters. Each interface only permits one route successful at the same time, and 8 interfaces can route parallelly.

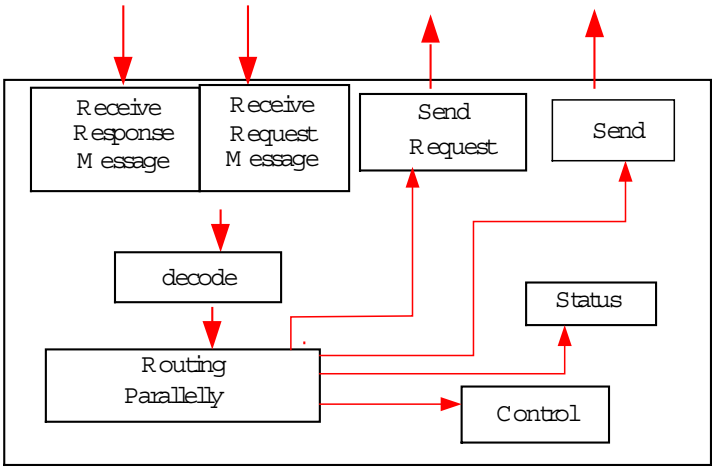

Fig.7 The block diagram of the InterRCU

If a routing request is accepted, the InterRCU will send a new request message to the local cluster's IntraRCU to select the destination host. When the InterRCU receives the success response message, it will control the IPU and send a response message to the original IntraRCU telling the request is acceptable, the original IntraRCU will control the original IPU and tell the source host to send data. Otherwise, the InterRCU will return a failure response message to the original IntraRCU, and the original IntraRCU will tell the source host not to send data and try to send request message again.

Some special functions are included in the user-defined communication protocol:

1) In the transmitter port, host firstly sends a request signal when some data need

to be send to the other hosts. The request signal included the destination address(7 bit), source address(7 bit), class(2 bit), message type(1 bit), flag(1 bit) and so on. The destination address is used to select route. It controls one of the seven VCSEL driver circuit and sends the request signal to the destination cluster through the signal fiber. 2) In the receiver port, based on the information type in the signal, it judges that the signal is request signal or respond signal when the receiver port get the signal through the signal fiber. If the signal is the request one, we arbitrate the request and amend the corresponding symbol according to the rank. When the corresponding data are allowed to receive, set the symbol "1". It means "request permit". Otherwise, set it "0". It means "request denied". At the same time, based on the source address, we control one of the seven VCSEL driver circuit and send the respond signal to the source cluster by using the signal fiber. Then send a massage to the selective receiver circuit to receive the data. 3) When the source cluster received the signal through the signal fiber, based on the information type in the signal, it judges that the signal is request signal or respond signal. If the signal is the request, it detects the correspond symbol in the signal. If it is "1", it means "request permit". We control one of the seven VCSEL driver circuit and send the data to the destination cluster by using the data fiber. If it is "0", it means "request denied". We store the data and send a request signal after a while. 4) When the receiver port get the data through the data fiber, it makes the OE transform, amplifying by using the PIN. After these, it makes the physical layer process, such as CRC. Finally, by user-defined protocol processing, it becomes the user data and is sent to the cluster.

4. CONCLUTIONS

A novel hybrid electrical optical Clos switch network for multiprocessor cluster system was presented. For multiprocessor cluster system of 128 hosts, the novel optical Clos network includes 16 basic modules, a passive optical fiber backplane with (8 x 15) x 16 which has a total of 1920 optical data channels and a signaling control system. The basic module is composed of the input line cards of 8 hosts, a single chip of 16 x 16 crossbar switch, parallel transmitting VCSEL modules for fan-out of (16-1) optical fiber channels and (16-1) x 1 optical combiners on a module. The passive optical fiber backplane of very large capacity and high density, based on linear VCSEL arrays and fiber ribbon technology, is to be used to interconnect between hosts of different sub-clusters. The routing election of the optical Clos switch network is controlled by a signaling control system. Compared with high performance electronic system, this technology offers a relatively easy and simple means of communicating large amount of information between hosts and lower delay time.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the supports by the Chinese National Natural Science Foundation of China under contract: 60177023, and the 863 High-Tech Program of China under contracts: 2001AA111030, 2001AA120203, 2002AA103064.

REFERENCES

[1] J. Esmith and W. Weingarten, Eds., "Research challenges for the next generation Internet: Computing Res. Assoc., May 12-14,1997.

[2] Elmar Griese, "Optical Interconnection on Printed Circuit Boards", In Optics in Computing 2000, SPIE Vol.4089(2000)0277-786X/00

[3] H. J. Chao and Tishiang Wang, "An optical interconnection network for Terabit IP routers" Journal of Lightwave Technology, Vol.18, No.12, pp. 2095-2112, 2000.

[4] K. Kate, M. Ishii, and Y. Inoue, "Packaging of large-scale planar light-wave circuits," in Proc. IEEE Components Technol. Conf., pp.37-45, 1997.

[5] T. H. Szymanski, A. Au, et. al., "Terabit optical local area networks for multiprocessing systems", Applied Optics Vol.37, No.2, pp. 264-275, 1998

[5 Mingcui Cao, Hongpu Li, Fengguang Luo, Ai Jun, and Da Liu, "Free-space regular optical interconnections: a mathematical analysis," Appl. Opt., 33, pp. 2960-2967, 1994.

[6] Fengguang Luo, Mingcui Cao, Anjun Wan, Jun Xu, "New free-space multistage optical interconnection network and its matrix theory," in Optoelectronic Interconnects VII, Proceedings of SPIE Vol. 3952, pp. 296-302, 2000.

[7] http://www.myrinet.com/myrinet/m3switch/guide/myrinet-2000_switch_guide.pdf