Fast-Reconfigurable Optical Interconnect Architecture Based on Time-Synchronized Node Coordination for High Performance Computing

Yufang Yu, Nan Hua, Zhizhen Zhong, Jialong Li, Ruijie Luo, Zelin Zheng, Xiaoping Zheng
Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
{huan, xpzheng}@mail.tsinghua.edu.cn

Abstract: We propose a Fast-Reconfigurable Optical Interconnect (FROI) architecture enabled by time-synchronized node coordination for high performance computing. Experimental results show that an ultra-low reconfiguration time of 45.6μs can be achieved after traffic pattern changes.

OCIS codes: (060.4250) Networks; (060.4256) Networks, network optimization;

1. Introduction

High performance computing (HPC) systems have been playing a crucial role in various scientific and engineering areas, such as climate predict, earth surface modeling and sky simulation [1]. Within large-scale HPC systems, various kinds of applications coexist, and may have different traffic patterns. Even for a specific scenario, its traffic pattern also varies dynamically over time [2]. Variations in traffic patterns requires the changes of communications connections to provide sufficient bandwidth and acceptable reconfiguration latency. So, HPC systems have to change their internal communication connections with agile responses accordingly.

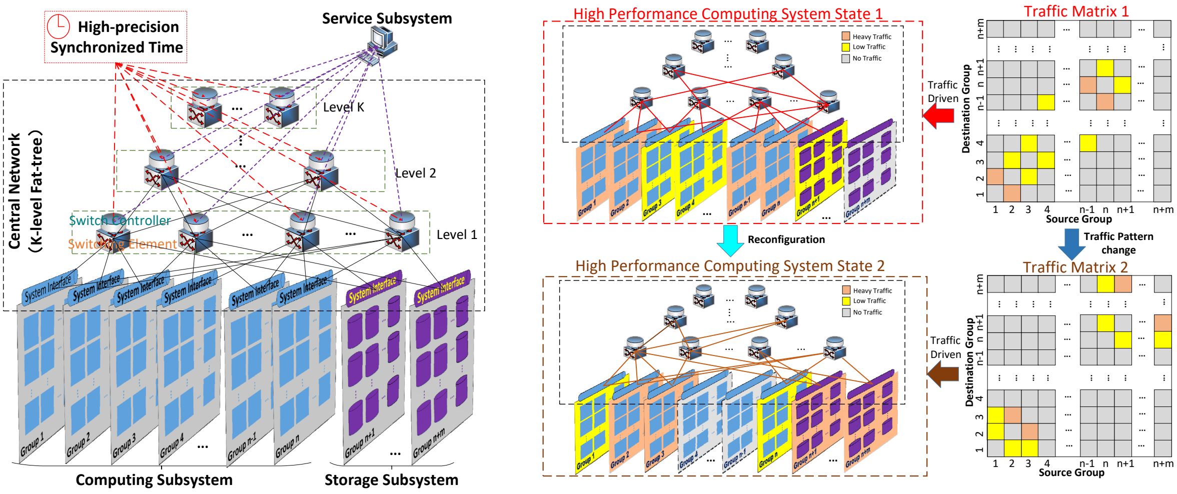
During the fast evolution of HPC, optical interconnect was introduced into HPC to meet the increasing demand of high bandwidth and low latency. However, constrained by the nature of optical switching, HPC with optical interconnect endures a slow switching process to adapt to traffic pattern change. Such switching process generally consists of control signaling, lightpath reconfiguration and communication establishment, etc.. This long pattern reconfiguration time severely compromises computing efficiency, limits the performance of HPC as a key bottleneck. For the past several decades, the conventional way to reconfigure optical connections in HPC systems is to use Micro-Electro-Mechanical Systems (MEMS) switch which takes several milliseconds to reconfigure its connections [3]. Then, from the perspective of the whole HPC systems, the reconfiguration time will be even longer. Obviously, such method is not suitable for future HPC systems. Hence, it is crucial to explore new solutions to achieve optical interconnect with agile reconfiguration. Recently, time-sliced resource allocation with synchronization [4] provides us a promising way for HPC requirements, and the advantages of this approach was first analyzed theoretically against conventional flexi-grid optical networks [5].

In this paper, we proposed a new paradigm to achieve a Fast-Reconfigurable Optical Interconnect (FROI) with time synchronization for HPC. An enabling algorithm for routing and time-slice allocation is introduced, and an experimental demonstration is carried out to verify the superiority of FROI in HPC systems.

2. Fast-reconfigurable optical interconnect architecture

Generally, HPC systems contain three parts: computing subsystem, storage subsystem and service subsystem [3]. Specifically, service subsystem refers to network controlling switching elements. Nodes inside computing subsystem and storage subsystem are further classified into several groups, named computing groups and storage groups, respectively. These groups can access the central network through system interface (SI) and communicate with each other through the central network. In practical HPC systems, like *Sunway Taihulight* Supercomputer [6], the central network is organized in a k-level fat-tree way in order to achieve all-to-all communications while leaving good potentials for future upgrade. Besides, fat-tree topology is efficient for HPC system [7].

In FROI, as shown in Fig. 1 (a), configuration rules are programmed into the switch controllers (SC) previously. A configuration rule contains configuration time, the number of switch controller, ingress port and egress port. The switch controllers translate the rules and control the switching elements (SE) to tear down original lightpath and set up new lightpath at special time points enabled by precise time synchronization. In different time slice, the state (including the topology of central network and the communication state of both computing groups and storage groups) of HPC system is different. The topology is controlled to change accordingly to the traffic pattern (presented by traffic matrix), as shown in Fig. 1 (b). Driven by different traffic patterns, the HPC system reconfigures its internal connections among different states.



(a) (b)

Fig. 1, FROI for HPC: (a) Architecture of the FROI for HPC; (b) An example of HPC reconfiguration through FROI

3. Traffic patterns and proposed algorithm

In an HPC system, there are mainly two kinds of traffic: communications between two computing groups, and communications between computing group and storage group. For different applications, variable factors, such as the number of computing nodes, the number of storage nodes, data exchange between the nodes, etc., are influencing HPC traffic, resulting in different traffic patterns under different applications, as studied in [2].

To support all kinds of applications, we propose a pattern-based algorithm for routing and time slice allocation which is shown as Table 1. The algorithm has two considerations, one is to ensure the algorithm can get a valid routing and time slice allocation for any traffic patterns, another is to make full use of the link resource. Before the algorithm, we should determine the length of time slice and compute the time slice matrix. The element of time slice matrix show the number of time slice for the communication between any two groups.

Tab. 1.Pattern-based routing and time slice allocation algorithm.

```
Input: topology $G(V,E)$, time slice matrix ($V$ to $V$ matrix) $T$.
Output: the set of time slice allocation $\Phi$, the set of routing path $\Psi$.
Initial: $\Phi=\{\o\}$, $\Omega$(the set of topology)$=\{\o\}$, $\Psi=\{\o\}$.
1. For $k=1$ to $V^2$
2.    $T(i, j)=max(T)$, If$(T(i, j)==0)$ break;
3.    For $L=1$ to $/\Omega/$
4.        $R=$Dijkstra $(\Omega(L)$, from group $i$ to group $j)$;
5.        If$(R!=NULL)$    $\Psi(L)=\{\Psi(L), R\}$, $\Phi(L)=\{\Phi(L),(i, j)\}$,$\Omega(L)= \Omega(L) - R$, $T(i,j)=0$, break;
6.    If$(R==NULL)$    $R=$Dijkstra$(G$,from $i$ to $j)$, $\Phi=\{\Phi,(i, j)\}$,$\Omega=\{\Omega,G- R\}$, $\Psi=\{\Psi, R\}$, $T(i,j)=0$;
7. Return $\Phi,\Psi$.
```

We further analyze the computation complexity of the algorithm. Generally, the algorithm contains three steps: selecting the element of time slice matrix, computing the routing path, allocating the time slice. We select the biggest number in the matrix (line 2) and compute the routing path by Dijkstra algorithm (line 4).For the crucial step--time slice allocation, we try to allocate the existing time slice to the selected element. If this is not possible, we then allocate a new time slice to it (line 5-6). The time complexity of our algorithm is $O(V^2Elog(V))$.

4. Experimental setup and results

A prototype experiment is carried out to evaluate the performance of FROI for HPC systems. As shown in Fig. 2, the central network is constituted of eight magneto-optic switches (MO-SW) and organized in a two-level fat-tree way. The central network support communication of eight groups. The communication data from the server is regarded as output of the groups. To simulate the HPC system, two different traffic patterns are generated and their traffic matrixes are shown as Fig. 3. The time slice is 400us, and the time slot of heavy traffic contain two time slices while low traffic contains one. The routing of the system is calculated by the above-mentioned algorithm. One period of Traffic Pattern 1 is calculated to contain ten time slices and that of Traffic Pattern 2 contains seven.

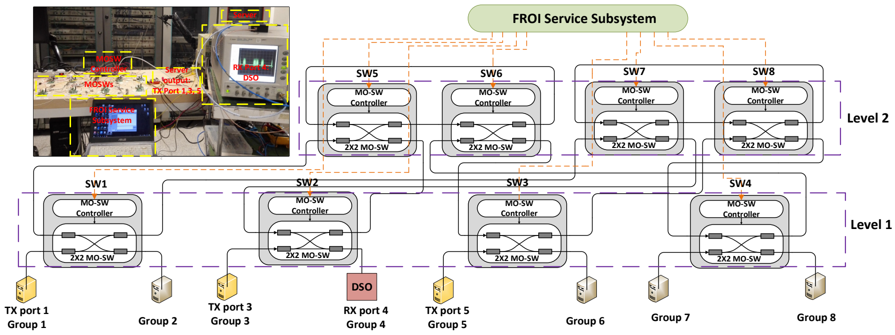


Fig. 2. Experimental setup.

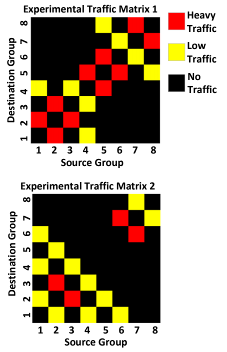
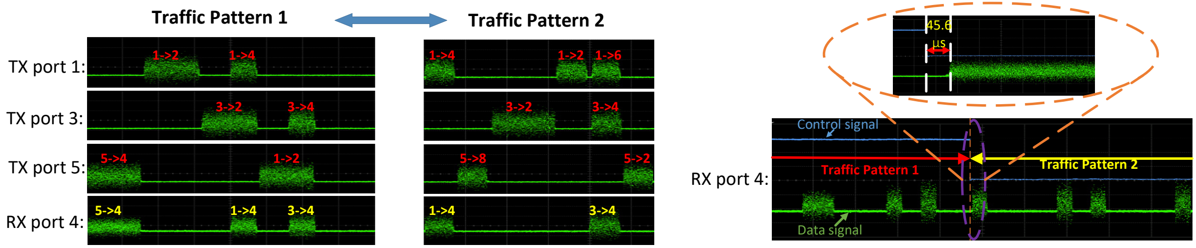


Fig. 3. Experimental traffic patterns.

To observe the changing of the traffic patterns, we use DSO to observe the data received by Group 4. In Traffic Pattern 1, the data received is from Group 1, Group 3 and Group 5. And in traffic pattern 2, the received data is from Group 1 and Group 3. So, we observe the data sent by Group 1, Group 3 and Group 5 by DSO to explore whether the network work correctly. And the results are shown as Fig. 4 (a). Group 4 receives the correct data at the correct time in two traffic patterns. To measure the system reconfigurable time, the SW Controller will generate an electrical signal (the level inversion in Fig. 4 (b)) when the traffic matrix changes from Pattern 1 to Pattern 2. In this way, the reconfigurable time can be obtained by measuring the time interval between the time points of the level inversion and the start of the first time slice of Traffic Pattern 2. As shown in Fig. 4 (b), the reconfiguration time of the system is 45.6μs.



(a) (b)

Fig. 4. Experimental results: (a) Data exchange through FROI ; (b) Reconfigurable time.

4. Conclusion

In this paper, we propose a Fast-Reconfigurable Optical Interconnect (FROI) architecture for HPC systems. The FROI architecture can achieve fast traffic pattern change by time-synchronized node coordination. We further propose a pattern-based routing and time-slice allocation algorithm for network resource allocation under FROI architecture. The fast reconfiguration performance of the proposed architecture is validated through a prototype experiment. Results shows that an ultra-low reconfiguration time of 45.6μs can be achieved, which may help the HPC systems meet the future requirements for fast reconfiguration.

5. Acknowledgement

This work was supported in part by projects under National 973 Program grant No. 2014CB340104/05, and NSFC under grants No. 61621064, 61435006.

6. References

[1] H. Fu, *et al.,* "The Sunway TaihuLight supercomputer: system and applications," *Science China Information Sciences*, 2016.
[2] K. Wen, *et al.*, "Flexfly: Enabling a reconfigurable dragonfly through silicon photonics," in *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, 2016
[3] K. J. Barker, *et al.*, "On the feasibility of optical circuit switching for high performance computing systems," in *Proc. ACM/IEEE conference on Supercomputing*, 2005.
[4] N. Hua, *et al.*, "Optical time slice switching (OTSS): An all-optical sub-wavelength solution based on time synchronization," in *Proc. ACP*, 2013.
[5] N. Hua, *et al.*, "Enabling low latency at large-scale data center and high-performance computing interconnect networks using fine-grained all-optical switching technology," in *Proc. ONDM*, 2017.
[6] C. Yang, *et al.*, "10M-core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics," in *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, 2016.
[7] C. E. Leiserson, "Fat-trees: universal networks for hardware-efficient supercomputing," Computers, IEEE transactions, on 1985, 100(10):892-901.