sDNA: Software Define Network Accelerator based on Optical Interconnection Architecture

En Shao, Guangming Tan, Zhan Wang, Guojun Yuan, Ninghui Sun
Institute of Computing Technology, Chinese Academy of sciences, University of Chinese Academy of Sciences
Beijing, China
shaoen@ict.ac.cn

1 INTRODUCTION

Many operational supercomputers are designed as multi-tenant systems [1][2][3] with running characteristics similar to those of cloud-computing centers. To reduce resource competition of each job and ensure optimal load balance, a job scheduler adopts the discrete scheduling strategy, making the computing nodes occupied by each job are separated. However, the communication distance, which is caused by the discrete scheduling strategy, worsens the issue of communication delay between two processes. As a result, inter process communication (IPC) between different irrelevant jobs increases the probability of routing paths being shared, as this usually leads to severe congestion of a large-scale system.

A recent trend in HPC design is to build communication shortcuts by using optical circuit switching (OCS) because these optical devices are highly reconfigurable [4–8]. The optical links between racks are reconfigured according to the traffic patterns of different jobs. However, there is a dilemma when we choose one OCS to build large-scale parallel systems.

Table 1 summarizes two mainstream OCS devices: microsecond circuit switching(WSS) and 3D Micro-Electro-Mechanical System (3D-MEMS)/DirectLight Beam-Steering (DBS). They have different properties of switching latency and port count. On one hand, the WSS devices, like Mordia [7] and Projector [8], have been deployed to accelerate the communication in data center. However, the small number of ports limits their use in a very large scale system like the planned exascale supercomputer in 2022. For example, in our projected exascale supercomputer built on 6D-torus network, the number of switches is 4608, and the maximum hop count is 14.

Table 1: The device characteristics of different optical circuit switching device

	Fast reconfigure	Slow reconfigure
Optical device	WSS(Mordia)	3D-MEMS/DBS
Switch time	11.5 μs	20 to 200ms
Port count	24	320/384
Subnet scale	192 switches	12 to 15 switches
Hop count	7 hops	3 hops

On the other hand, both 3D-MEMS and DBS are mature commercial OCS devices that contain more than 300 ports in one device while the optical circuit switching time is as slow as $20 - 200$ms. With these slow switching optical devices, the traffic offloading methods driven by congestion response [4–6] only support highly aggregated traffic with stability in date center. The result is that the traffic offloading method by congestion response cannot keep each optical link of OCS with high utilization in exascale computer as changing traffic.

With respect to the practice application of optical switch devices, an exascale computer will prefer to either 3D-MEMS or DBS. The key issue is how to overcome the shortcoming of slow switching operations. The work introduces an interconnection communication accelerated system–Software Defined Network Accelerator (sDNA) to alleviate the side effect of slow reconfiguration in 3D-MEMS or DBS. *The key idea is to leverage traffic pattern to schedule jobs so that a proper optical link is built and allocated for any job just before it is scheduled to execute.* Thus, sDNA leverages the flexibility of optical interconnection to dynamically make full use of the separated resources. Two steps must be carefully designed and implemented for communication acceleration: (i) *build optical links*: we connect optical links to offload the congested traffic. (ii) *use optical links*: we use the optical links and route the traffic in an effective way. We take the operating environment into account for calculating any optical link candidate's revenue. After the links are connected, our routing method provides an efficient mechanism to achieve high usability and utilization of optical link. Meanwhile, a deadlock-free routing algorithm adjusts the utilization of optical link guided by prior knowledge.

2 MOTIVATION

A communication bottleneck occurs when the scheduling continuity is broken, threatening the computing efficiency of the multi-tenant system. However, as the limitation of slow switching optical device, the traffic offloading method by congestion response cannot always keep each optical link of OCS with high utilization in several milliseconds.

Fig. 1 show the performance comparison of throughput between two building optical link methods when we use slow switching optical device to offload traffic. We choice all-to-all as the traffic pattern. Only the method according to longterm traffic pattern can effectively improve network performance by offloading traffic. As a result, to effectively offload traffic from electrical network in several milliseconds or more long time, it is necessary to combine the characteristic of HPC traffic pattern with building optical links.

In response to the limitation of slow switching optical device, we propose a new traffic offloading method called extended edge forwarding index (E-EFI). Instead of changing the topology of optical network in response to congestion, E-EFI instead leverages the characteristic of HPC traffic pattern. We use a theory of network measurement to compute the traffic offloading revenue of each optical link candidates. E-EFI chooses these optical link candidates with high revenue to rapidly multiplex circuits across a set of endpoints.

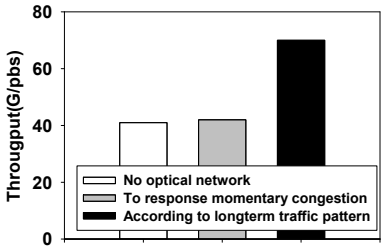

Figure 1: Comparison between two traffic offloading methods.

In this way, we can keep optical network with high utilization as routing path during each job's running. In addition, we need to overcome three challenges which have been shown as below.

- It is hard to take the suitable characteristics of HPC traffic into account. We select scheduling location of each job and communication ratio (*CR*) between each node per each job as two key characteristics of HPC traffic. Because the scheduling location limits the participants of communication in each job. On another hand, the communication ratio is able to quantify the aggregated degree of traffic between different nodes.
- The ready-made solution to calculate the traffic offloading revenue has not been proposed so far. The original edge forwarding index (EFI) is a classics network measurement theory. Inspired by original EFI, we proposed a new way of network measurement to calculate revenue of each optical link candidate with two characteristics of HPC traffic.
- As offloading influence of each selected optical link candidates is ignored, the calculation of later candidates' revenue may be inaccuracy for repeatedly offloading same traffic. We optimize the process of revenue calculation by updating the topology of optoelectronic network when one optical link candidate has been selected for building. In addition, we take the updated topology into account for calculating the revenue of later optical link candidates.

3 BUILDING OPTICAL LINKS

Inspired by the original EFI, we propose a new network measurement as the mathematical foundation of building optical links. First, according to the basic definition of EFI, we define $EFI_{CR}(c_q)$ as the congestion evaluation of link c_q. Second, we design an algorithm for calculating the revenue of each optical link candidate.

3.1 Extended EFI

In order to formulate our algorithm, we first several mathematical symbols used in the new network measurement.

Definition 1 (Network *I*). We define an interconnection network as a multi-graph $I := G(N, C)$ of the node set N and the link set C. Each pair of network switches $n \ni N$ is connected by several duplex links $c \ni C$. Unlike other networks, the link set in C contains optical link subset C^O and electronic link subset C^E as $C = C^O \cup C^E$.

Definition 2 (Routing path set $P_{a,b}$). Routing path set $P_{a,b}$ is defined as a sequence of links $(c_a,...,c_b) := P_{a,b}$. Each link in routing path $P_{a,b}$ refers to the electronic link as $\{c_a,...,c_b\} \subset C^E$, with the link starts from node a as $c_a := (n_a, \cdot)$ and the link end at node b as $c_b := (\cdot, n_b)$. The link in the middle of a routing path $c_{a_1,a_2} := (n_{a_1}, n_{a_2})$ represents the link between node n_{a1} and n_{a2}.

Definition 3. The original Edge Forwarding Index (original EFI) of link c_q is given by:

$$EFI(c_q) := \sum \left| \{P_{a,b} | a, b \in N \wedge c_q \in P_{a,b}\} \right| \quad (1)$$

The $EFI(c_q)$ acts as the counting result of link c_q. When c_q acts as a routing path between two nodes, the counter of EFI adds *1* to the $EFI(c_q)$ of link c_q. In addition, the original EFI indicates the utilization of one link.

Definition 4. The extended EFI with the scalar of application communication ratio *CR* is given by:

$$EFI_{CR}(c_q) := \sum_j \left| \{P_{a,b} \cdot CR_{(a,b)} | a, b \in N_j \wedge c_q \in P_{a,b}\} \right| \quad (2)$$

In the above formula, the job j is scheduled onto a subset of nodes, $N_j \subseteq N$. Each job occupies those nodes during a finite time. An over-low $EFI_{CR}(c_q)$ indicates that the link c_q is utilized at low efficiency. On the contrary, an over-high $EFI_{CR}(c_q)$indicates that the link c_q is congested with high probability. The application's communication ratio (*CR*) between node a and node b is given by:

$$CR_{(a,b)} = \sum_j CR_{(j,a \to b)} + \sum_j CR_{(j,b \to a)} \quad (3)$$

3.2 Revenue Evaluation

The calculating principle of the evaluation algorithm is to calculate the max and sum of $EFI_{CR}(c_q)$ covering every electrical link. Different optical link candidates have a different influence on the result of $EFI_{CR}(c_q)$. The revenue of one optical link candidate is higher if the result of $EFI_{CR}(c_q)$ is lower. The max of $EFI_{CR}(c_q)$, which is the result of the evaluation, shows the congestion level of the network.

The main component of the evaluation algorithm contains three components: First, we set every link's $EFI_{CR}(c_q)$ to zero. Second, we build a logical topology with the optical link between the node pair(N_i,N_j) and generate the valid routing path. Third, we calculate the max and sum of $EFI_{CR}(c_q)$ for every electrical link. We learned from the previous idea of SAR[1] and removed the invalid routing paths between different jobs in our design. Because of the irrelevance of communication, routing paths between different jobs are never used.

The basic algorithm of revenue evaluation contains extend EFI. Comparing the original EFI, another optimization of revenue evaluation contains the topology updating. To achieve this optimization, we compared three different methods of revenue evaluation, namely by distance, by original EFI and by extended EFI that includes topology updating.

4 CONCLUSION

Our sDNA is able to solve the two key issues of network accelerator, namely the connection and the usage of the optical links, according to the operational situation and prior knowledge. Also, this work demonstrates that the optical interconnection can act as a capable and flexible network accelerator in the exascale computer.

REFERENCES

[1] J. Domke and T. Hoefler. Scheduling-Aware Routing for Supercomputers. Nov. 2016. Accepted at The International Conference for High Performance Computing, Networking, Storage and Analysis (SC'16).

[2] Kevin J Barker, Alan Benner, Ray Hoare, Adolfy Hoisie, Alex K Jones, Darren K Kerbyson, Dan Li, Rami Melhem, Ram Rajamony, Eugen Schenfeld, et al. On the feasibility of optical circuit switching for high performance computing systems. In *Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, page 16. IEEE Computer Society, 2005.

[3] Zhiyang Guo and Yuanyuan Yang. Multicast fat-tree data center networks with bounded link oversubscription. In *INFOCOM, 2013 Proceedings IEEE*, pages 350–354. IEEE, 2013.

[4] Christopher Batten, Ajay Joshi, Vladimir Stojanov, and Krste Asanovi. *Designing Chip-Level Nanophotonic Interconnection Networks*. Springer New York, 2013.

[5] Guohui Wang, David G Andersen, Michael Kaminsky, Konstantina Papagiannaki, TS Ng, Michael Kozuch, and Michael Ryan. c-through: Part-time optics in data centers. In *ACM SIGCOMM Computer Communication Review*, volume 40, pages 327–338. ACM, 2010.

[6] Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Helios: a hybrid electrical/optical switch architecture for modular data centers. *ACM SIGCOMM Computer Communication Review*, 40(4):339–350, 2010.

[7] George Porter, Richard Strong, Nathan Farrington, Alex Forencich, Chen Sun Pang, Tajana Rosing, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Integrating microsecond circuit switching into the data center. *ACM SIGCOMM Computer Communication Review*, 43(4):447–458, 2013.

[8] Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. Projector: Agile reconfigurable data center interconnect. In *Conference on ACM SIGCOMM 2016 Conference*, pages 216–229, 2016.