Enabling Low Latency at Large-Scale Data Center and High-Performance Computing Interconnect Networks Using Fine-Grained All-Optical Switching Technology

Nan Hua[*†], Zhizhen Zhong[*†], and Xiaoping Zheng[*†]
[*]Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China
[†]Department of Electronic Engineering Tsinghua University, Beijing 100084, China
huan@mail.tsinghua.edu.cn

***Abstract*—In large-scale data center (DC) and high performance computing (HPC) interconnect networks, end-to-end latency becomes a fatal problem due to the processing and queuing delays of electronic packet switching (EPS) at intermediate switching points. Introducing optical switching into DC and HPC networks can provide a potential solution to the latency problem by establishing low-latency optical bypass (end-to-end lightpath). However, the number of connections that can be provided by current coarse-grained optical circuit switching (OCS) technology is far less than the required amount for all-to-all communication in a hundreds-of-thousands-nodes large-scale system, and this will weaken its effect in reducing latency. Optical packet switching has a much finer granularity compared with OCS; however, the lack of adequate technologies for optical buffering makes it difficult to avoid packet collision. In this paper, we investigate the relationship among network scale, granularity and latency, and introduce the use of our proposed fine-grained optical time slice switching (OTSS) in DC and HPC networks over arbitrary topologies. Simulation results under 6x6 2-D Torus topology demonstrate the advantage of OTSS in end-to-end latency compared with conventional EPS and spectrum-flexible wavelength switching (WS).**

***Index Terms*—end-to-end latency; electronic packet switching; fine-grained optical switching; hyper-scale interconnect network; data center; high-performance computing.**

I. Introduction

TOWARDS the year 2020 and beyond, 5G will be able to provide users with ubiquitous coverage, fiber-like access rate and "zero" latency, which creates the need for the deployment of hyper-scale data center (DC) and high-performance computing (HPC) interconnect networks, containing more than hundreds of thousands of interconnected servers/nodes. The hyper-scale DC and HPC networks require highly scalable interconnection networks which can provide low-latency and high-bandwidth connections with low energy consumption [1]. Current designs for such interconnection networks are based on electronic packet switching (EPS), which is gradually exposing the weakness in capacity, energy consumption and end-to-end latency, making the underlying switching fabric a fundamental bottleneck for further expansion of DC and HPC networks [2].

Introducing optical technologies into DC and HPC networks provides a possible alternative to electronic switching, which has the potential benefit of breaking through the bottleneck of electronic packet switching, and enabling flexible network topology reconfiguration [3]. However, in current commercial systems optics are almost utilized for point-to-point data transmission, while data switching at intermediate switching points still remains electronic, leaving the problem of latency and energy consumption still unsolved. For this reason, all-optical switching technology has become more and more attractive in DC and HPC networks which can eliminate the need for optical–electrical–optical (O/E/O) conversion and reduce electronic queuing and processing delays.

In general, optical switching technologies can be classified into two types: unbuffered optical circuit switching (OCS), and buffered optical packet switching (OPS). OCS can reduce the buffering and switching capacity requirements for the router by enabling low-latency optical bypass [4]. However, current OCS is not able to provide fine-grained connections, leading to inefficient network utilization. OPS has a much finer granularity (at packet level) compared with OCS, but it is facing a major challenge of the lack of adequate technologies for optical buffering [5]. Time-division-multiplexed (TDM) wavelength routing was proposed in literatures [6] and [7], which can establish fine-grained lightpaths by dividing time into frames of slots. Since the TDM wavelength routing networks operate in circuit-switched mode, no optical buffer is required. However in such networks, time slot collision becomes a difficult problem that remains unsolved for arbitrary topologies other than star-based or ring-based ones. In this paper, we investigate the relationship among network scale, granularity and latency, and introduce the use of our proposed fine-grained optical time slice switching (OTSS) in DC and HPC networks over arbitrary topologies, with some results from prototype experiments. We also present simulation results under 6x6 2-D Torus topology

that demonstrate the advantage of OTSS in end-to-end latency compared with conventional EPS and spectrum-flexible wavelength switching (WS).

II. Network Scale, Granularity and Latency

In current electronic-switching-based DC and HPC networks, when a packet travels from one server/node to another it will experience significant queuing and processing delay at each switch, and the number of switches it traverses (hops) will increase with the expanding network scale, leading to higher end-to-end latency. Taking for example the "direct network" architecture [8], it can be proved that the maximum hop count for a k-Degree and N-node network is no less than O($log_k N$). Replacing electronic switches with optical ones will significantly reduce electronic-layer hop (virtual hop) count by establishing optical bypass, thus shortening end-to-end latency. In an ideal situation where a dedicated lightpath (LP) can be provisioned to every node pair, the maximum virtual hop count is only 1.

Fig. 1 gives the estimated results of the maximum number of end-to-end lightpaths that can be offered in different topologies (i.e., 6-D Torus, 2-D torus, 4-degree Butterfly and Fat Tree) with single wavelength link between neighboring nodes. As shown in this figure, whatever the topology used, there exists a significant gap between the number of lightpaths required for all-to-all communication (dashed line) and that can be offered, and it can be observed that this gap will become wider as the network scales up.

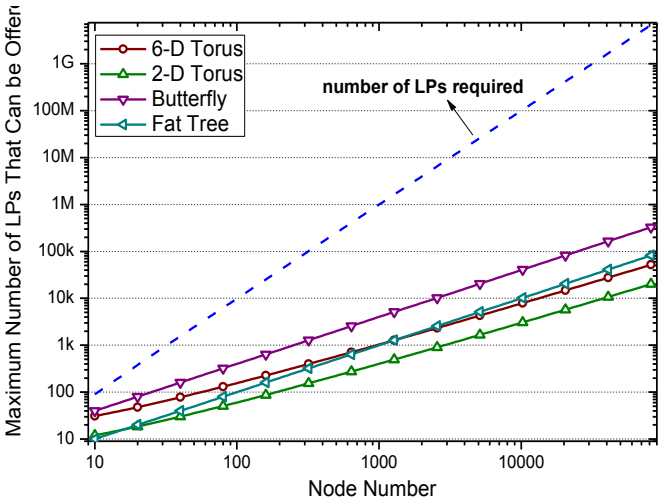

Fig. 1. The gap between the number of lightpaths required and that can be offered with single wavelength for different network topologies and scales.

In order to offer more lightpaths to meet the demand for all-to-all communication, fine granularity (defined as the maximum possible number of channels that can be established on a physical link) is needed. As shown in Fig. 2, for all-to-all communication of a ten-thousands-nodes network, the maximum offered lightpath number can meet the requirement if the switching granularity is able to reach the level of ten thousands. However, current dense wavelength division multiplexing (DWDM) systems can only provide several hundreds of wavelength channels in a fiber due to the physical restriction of optical components (e.g. the passband resolution of WSS) [9], which is too coarse to reach the requirement. In this scenario (inadequate granularity), a conflict will exist between latency (virtual hop count) and network throughput, that is, when considering low latency, the virtual hop count should be minimized. In this case, the network throughput will be limited when coarse-granularity wavelength switching is adopted, even for the minimum 6.25-GHz spectrum slot sizes (equivalent to 32 wavelengths in total 200-GHz spectrum bandwidth) at high blocking levels (Fig. 3); but if we are more concerned with network throughput, traffic grooming at the electronic layer is required. In this case, end-to-end latency can not be reduced effectively.

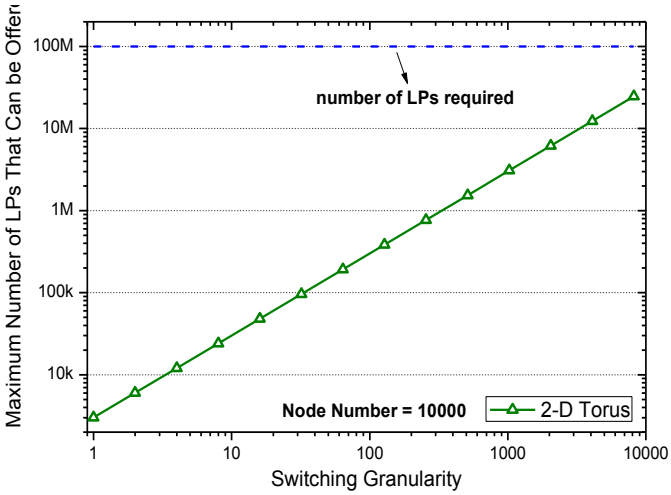

Fig. 2. The gap between the number of lightpaths required and that can be offered at different switching granularities (channel/wavelength numbers).

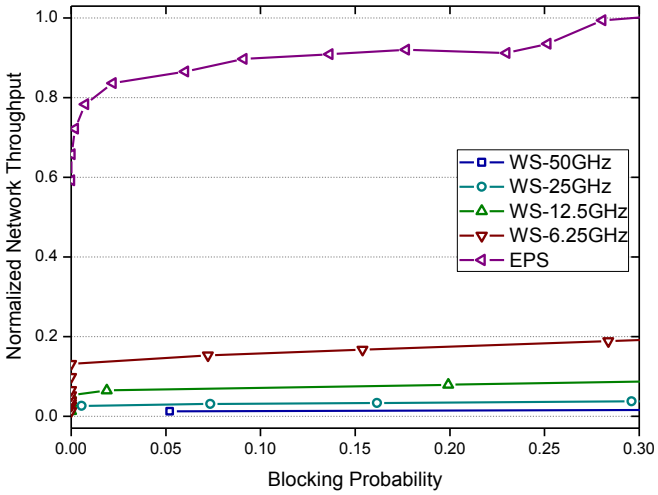

Fig. 3. Network throughput at different blocking probability levels: spectrum-flexible wavelength switching (WS) vs. electronic packet switching (EPS).

In order to resolve the conflicting issue to balance latency with network throughput, fine-grained optical switching technology is required.

III. Fine-Grained All-Optical Switching: Optical Time Slice Switching (OTSS)

In our previous work [10, 11], we proposed a sub-wavelength optical switching solution - optical time slice switching (OTSS). Enabled by mature high-precision time synchronization technology [12] and nanosecond fast optical switches, OTSS is able to offer over one thousand sub-channels on a single wavelength channel.

A. Concept

Figs. 4(a) and 4(b) compare the basic concepts of spectrum-flexible wavelength switching and OTSS. It can be observed that the two switching mechanisms are very similar except for the switching domains and switching fabrics.

In OTSS, the optical transmission channels are organized into repetitive OTSS frames in time domain. Each OTSS frame contains one or several variable-length time slice(s) for data transmission and each time slice occupies one time slot. When a time slice arrives at a switching node, the pre-set (periodic) control signals are sent to the OTSS fabric at the precise time to direct the time slice to the expected output port. To guarantee high-precision timing, time synchronization of all OTSS nodes is required. Literature [12] reported a high-precision network time synchronization result with an accuracy of 65*ns* realized under 13 synchronization hops over commercial transport networks. It is hopeful that this accuracy will further reduce to below 10*ns* in the next few years.

The high-precision network time synchronization and nanosecond fast optical switches make it possible for OTSS to achieve much finer granularity than wavelength switching. When employing the two technologies (i.e., OTSS and WS) together, the granularity is able to reach over ten thousands which is enough for ten thousands nodes' all-to-all communication.

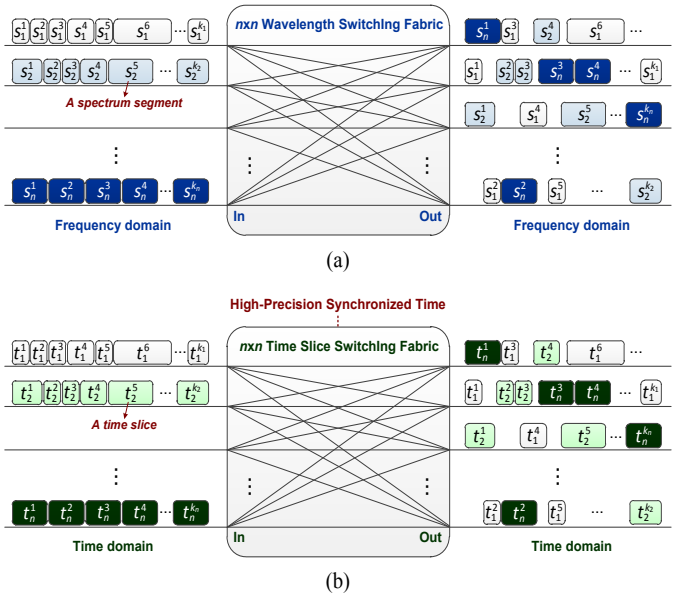

Fig. 4. Optical switching fabrics: (a) spectrum-flexible wavelength switching; (b) optical time slice switching (OTSS).

B. Prototype Experiments

We carried out prototype experiments to introduce OTSS into intra-DC networks [13, 14]. In the experiments, fine-grained all-optical circuit establishment was realized among ToR switches under a 3-Tier Fat-Tree architecture. The minimum time slice length, period of OTSS frame and guard interval between adjacent time slices were 1μs, 100μs and 100*ns*, respectively. All the time slices were switched at the pre-set precise time points by nanosecond PLZT switches. In the experiments, collision-free time slice assignment was also realized by a multi-controller collaboration mechanism.

IV. Simulation Results

In order to evaluate the latency performance of OTSS, we also conduct a simulation under 6x6 2-D Torus topology, as shown in Fig. 5. The length of the fiber link between adjacent nodes is set to 50 meters. Connection requests are generated between all node pairs uniformly and independently, characterized by Poisson arrivals with negative exponential holding times. In the simulation the average end-to-end latencies of EPS, OTSS and spectrum-flexible WS (with different spectrum slot sizes 6.25GHz, 12.5GHz, 25GHz and 50GHz, total 200-GHz spectrum bandwidth) are measured and compared. The spectrum efficiency is set to 1 bit/s/Hz. Traffic grooming at electronic layer is enabled for spectrum-flexible WS. The queuing and processing delay at each intermediate electronic switch is set to 5μs. For OTSS, the widths of a time slot and an OTSS frame are set to 100*ns* and 20μs, respectively.

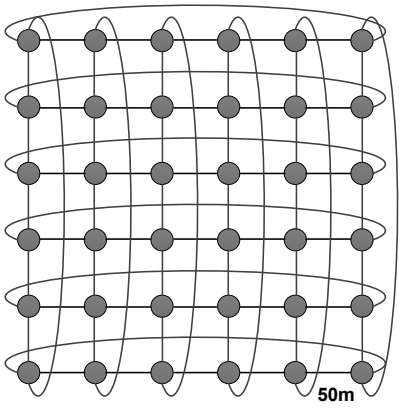

Fig. 5. 6x6 2-D Torus topology.

Fig. 6 gives the results of the average latencies of different switching mechanisms and granularities for two load scenarios. It can be observed that spectrum-flexible WS (with electronic-layer grooming) are able to reduce the end-to-end latency of the pure EPS, especially at light load. However, because of the restricted granularity, this improvement is limited, even for the finest granularity (6.25GHz). In comparison, OTSS can achieve significantly lower latencies in both the light and heavy loads scenarios due to its finer granularity.

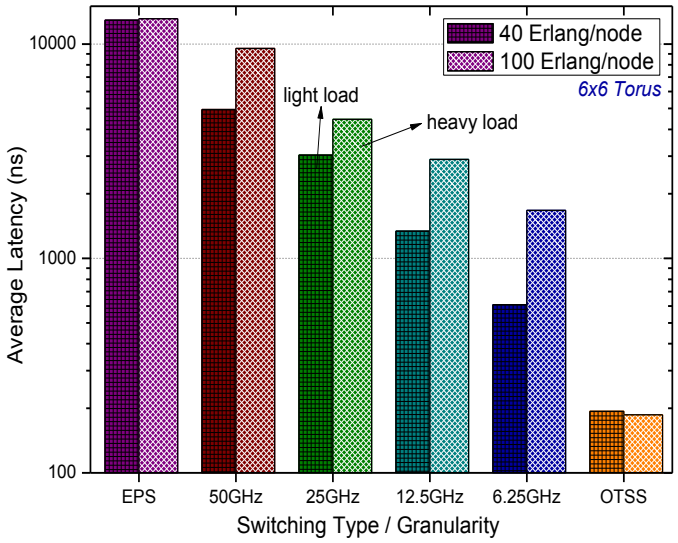

Fig. 6. Average latencies under different switching mechanisms and granularities.

V. Conclusions

As the scale of the DC and HPC networks increases, end-to-end latency is becoming a fatal problem for conventional electronic packet-switched network architecture due to the processing and queuing delays at intermediate switching points. Introducing optical circuit switching by establishing low-latency optical bypass can provide a potential solution to the latency problem; however, its effect is limited due to the restricted OCS granularity. In this paper, we investigate the relationship among network scale, granularity and latency, and indicate the necessity of granularity for reducing end-to-end latency. According to this conclusion, we attempt to introduce our proposed fine-grained optical time slice switching (OTSS), which is able to offer over one thousand sub-channels on a single wavelength channel, into DC and HPC networks. The same conclusion is drawn from the simulation result under 6x6 2-D Torus topology, that OTSS can achieve significantly lower latency compared with conventional EPS and spectrum-flexible WS.

Acknowledgement

This work is supported in parts by projects under National 973 Program grant No. 2014CB340104/05 and NSFC under grant No. 61621064.

References

[1] C. Kachris, I. Tomkos, "Power consumption evaluation of all-optical data center networks," *Cluster Comput.* 16(3): 611–623, 2013.

[2] O. Liboiron-Ladouceur, P.G. Raponi,N. Andriolli,et al., "A scalable space-time multi-plane optical interconnection network using energy-efficient enabling technologies," *IEEE/OSA J. Commun. Net.*, 3(8): 1-11, 2011.

[3] Cyriel Minkenberg, German Rodriguez, Bogdan Prisacari, et al., "Performance benefits of optical circuit switches for large-scale dragonfly networks," in *Proc. IEEE/OSA OFC*, Mar. 2016.

[4] M. Fiorani, M. Casoni, S. Aleksic, "Hybrid optical switching for energy-efficiency and QoS differentiation in core networks," *IEEE/OSA J. Commun. Netw.*, 5(5): 484–497, 2013.

[5] S. Yao, B. Mukherjee and S. Dixit, "Advances in photonic packet switching: An overview," *IEEE Commun. Mag.*, 38(2): 84-94, 2000.

[6] I. P. Kaminow, C. R. Doerr, C. Dragone, et al., "A wideband all-optical WDM network," *IEEE J. Sel. Areas Commun.*, 14(5): 780-799, 1996.

[7] S. Subramaniam, E. J. Harder, H. A. Choi, "Scheduling multirate sessions in time division multiplexed wavelength-routing networks," *IEEE J. Sel. Areas Commun.*, 18(10): 2105-2110, 2000.

[8] W. J. Dally, B. P. Towles, *Principles and practices of interconnection networks*, Elsevier, 2004.

[9] G. Shen, Q. Yang, "From coarse grid to mini-grid to gridless: how much can gridless help contentionless?" in *Proc. IEEE/OSA OFC*, Mar. 2011.

[10] N. Hua, X. Zheng, "Optical time slice switching (OTSS): an all-optical sub-wavelength solution based on time synchronization," in *Proc. IEEE/OSA/SPIE ACP*, Nov. 2013.

[11] N. Hua, X. Zheng, "All-optical time slice switching method and system based on time synchronization," *US Patent*, US 2016/0036555A1.

[12] L. Han, H. Li, L. Wang, et al., "First national high-precision time synchronization network with sub-microsecond accuracy over commercial optical networks for wireless applications," in *Proc. IEEE/OSA/SPIE ACP*, PDP AF4B.6, Nov. 2014.

[13] Y. Li, N. Hua and X. Zheng, "Fine-grained all-optical switching based on optical time slice switching for hybrid packet-OCS intra-data center networks," in *Proc. IEEE/OSA OFC*, Mar. 2016.

[14] Y. Jia, N. Hua, Y. Yu, et al., "Experimenting with multi-montroller mollaboration for large-scale intra-data center networks," in *Proc. IEEE/OSA OFC*, Mar. 2017.