Optics Communications 324 (2014) 147–151

Contents lists available at ScienceDirect

Optics Communications

journal homepage: www.elsevier.com/locate/optcom

Scheduling and performance evaluation of high line-rate space–wavelength routed switch for datacenter

Qirui Huang*, Xiaofei Cheng, Luying Zhou

*Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632, Singapore*

ARTICLE INFO

Article history:
Received 2 February 2014
Received in revised form
27 February 2014
Accepted 11 March 2014
Available online 28 March 2014

Keywords:
Datacenter
Optical cross-connect (OXC)
Optical packet switching

ABSTRACT

In this paper, we investigate the performance of a high line-rate space–wavelength routed switch to assess its potential for use in large-scale datacenters. A scheduling technique is introduced for the switch to address challenges in packet arbitration taking into account a high line rate and large port-count. The traffic performance in terms of packet loss rate and latency of the switch is evaluated by simulations under various traffic patterns. The results show that even under burst traffic conditions, the packet loss rate and latency of the switch are almost independent of the switch size, which indicates that the switch could support a large number of nodes for switching without significant performance deterioration for both fixed and variable packet sizes. It is also found that a slight increase in the number of tunable transmitters in each node could lead to a great reduction in packet loss and latency for the switch.

1. Introduction

Datacenters have become indispensable infrastructures for many businesses and organizations today. Driven by the increase of many new applications in the Internet, such as cloud computing, virtual machines, and storage area network (SAN), the number of servers in datacenters is increasing correspondingly to accommodate the explosion of data traffic [1]. Unlike telecommunication networks, the switching fabric of a datacenter network needs to support hundreds or thousands of nodes/racks. Today, a large-scale datacenter may comprise thousands of racks, each of which could host around 40 servers, and it requires a high bandwidth and low latency network to support fast data exchange between racks and core switches.

Currently, optical fiber cables have already been extensively used in datacenters for building high bandwidth interconnections between racks and core switches, while the data switching between these devices is still performed by electronic switches whereby the data signals have to undergo electrical–optical–electrical (E/O/E) conversion repeatedly. The use of electronic switches consumes a large proportion of energy utilized by datacenters, and this will keep increasing along with the growth of the datacenter. A large number of electronic switching devices also increase the cabling complexity and operating cost of the datacenters. These challenges limit the scalability and capacity of the future datacenters [2].

In recent years, optical switching techniques have been employed to replace some of the switching tasks originally performed by electronic switches to reduce the cabling complexity and power consumption of datacenters [3]. Optical circuit switching (OCS) is suitable for end-to-end connections between nodes over a long period of time. It can support large scale switching systems due to its low power consumption and implementation cost [4,5]. Currently, the switching time of commercial OCS based on MEMS is 10–20 ms. Due to its slow switching time, it cannot respond quickly to small-granularity traffic that requires high switching speed, and this results in a low utilization of bandwidth in the fiber interconnects. To address the limitation of switching speed, some optical packet-switched architectures have been proposed for datacenters and supercomputing applications. Ref. [6] proposed a buffer-less optical switch with three-stage Clos network that could achieve nearly 100% throughput, but it is difficult for practical implementation due to immaturity of tunable wavelength converters with large wavelength tuning range. The OSMOSIS switch in [7] used a simple broadcast-and-select scheme to realize low-latency and scalable switching architecture, but its implementation cost increases rapidly with the port count because of the broadcast-and-select architecture. Wavelength-routed switch architecture based on arrayed waveguide grating router (AWGR) and wavelength tuning technique is a very attractive scheme of realizing large switching speed [8,9]. Ref. [10] reported a 448×448 space–wavelength routed switch utilizing a parallel array of 32×32 cyclic AWGRs in conjunction with optical space

* Corresponding author.
E-mail address: qhuang@i2r.a-star.edu.sg (Q. Huang).

http://dx.doi.org/10.1016/j.optcom.2014.03.025
0030-4018/

switches. The large scalability combined with high switching speed makes the switch a promising solution for building high-throughput switching networks for large-scale datacenter networks. We have experimentally shown the feasibility of supporting 100 Gbit/s per wavelength in this switch fabric using coherent technology [11]. In this paper, to assess the potential application of the space–wavelength routed switch, we design a scheduling technique for the switch to perform packet arbitration, and then present a detailed investigation on its traffic performance.

The rest of the paper is organized as follows. In Section 2, the architecture of the space–wavelength routed switch is presented, followed by a scheduling technique for packet arbitration. In Section 3, the traffic performance of the space–wavelength routed switch is evaluated by simulation under various traffic patterns. Finally, the conclusions of this work are summarized in Section 4.

2. Architecture and operation

2.1. Space–wavelength routed switch

In order to support a datacenter with a large number of nodes, an optical switch should have large scalability. Wavelength-routed switches that employ fast tunable lasers (or wavelength converters) and AWGR to achieve packet switching are usually considered as a promising solution for large-scale switching. However, from practical implementation point of view, their scalability might be limited by the number of available wavelengths of the tunable laser (or wavelength converters) and the scalability of the AWGR. Due to the high inter-channel crosstalk and insertion loss, it is rather difficult to fabricate an AWGR with over 100×100 ports. So far, most of commercial AWGRs are 32×32 and 64×64. Meanwhile, fast tunable laser that can cover a broad wavelength range over 60 nm is also difficult to achieve.

To address these limitations, Fig. 1 depicts a space–wavelength routed switch utilizing a parallel array of smaller AWGRS. The switch architecture consists of N space switching planes and one wavelength switching plane that is composed of M number of $N \times N$ cyclic AWGRs and a set of wavelength division multiplexing (WDM) couplers. Each space switching plane is composed of $2M$ number of $1 \times M$ optical switches and M number of $2M \times 1$ optical combiners. Before entering the space switching planes, the optical signals are tuned to the appropriate wavelengths by the tunable transmitters at inputs. Then by configuring fast optical switches in the space switching planes, data from the input ports are switched to the appropriate AWGRs of the wavelength switching plane. Due to the routing characteristics of the cyclic AWGR, the optical signals assigned in different free spectral ranges (FSR) would emerge at a particular AWGR output port. To separate these signals before being received, a WDM coupler with two-band FSR is used to act as a waveband filter. With this approach, a 448×448 switch could be realized by a parallel array of 32×32 cyclic AWGRs [10], and larger scalability could be achieved by a slight increase in the number or size of AWGRs.

2.2. Pipelined scheduling technique

To solve the arbitration issues of the switch, we design a kind of pipelined scheduling technique. Note that the switch allows incoming packets to be stored in the electrical buffer in the tunable transmitters. To support a large number of nodes in a datacenter, the number of input/output ports of a switch would be very large. Most of the heuristic scheduling algorithms based on virtual output queues (VOQs) such as i-SLIP [12], DRRM [13] or PIM [14] need to carry out $O(\log_2 N)$ iterations to converge on a maximal matching between input and output ports for adequate performance. Although a number of iterations could be implemented within a short timeslot by current VLSI (very-large-scale integration) microelectronic chips and FPGAs (field-programmable gate array), there is a challenge to complete enough iterations as the timeslot is being shortened when the line rate of switch increases. For example, a 512×512 switch at 40 Gbit/s line rate at each input port needs to carry out nine iterations within12.8 ns if the packet size is 64 byte. Motivated by these, we adopt a pipelined arbitrator that consists of parallel sub-schedulers to address the challenge above. Fig. 2 illustrates the operation of four sub-schedulers. In this scheme a sub-scheduler performs only part of arbitration and then forwards the matching results to the next sub-scheduler. The matching results are obtained in an iteration that includes the following steps

Step 1: All sub-schedulers receive arbitration request from all of the unmatched VOQs at the input ports.
Step 2: In each sub-scheduler, only the request that has the highest priority at input and output pointers will be accepted for generating a matched input–output pair.

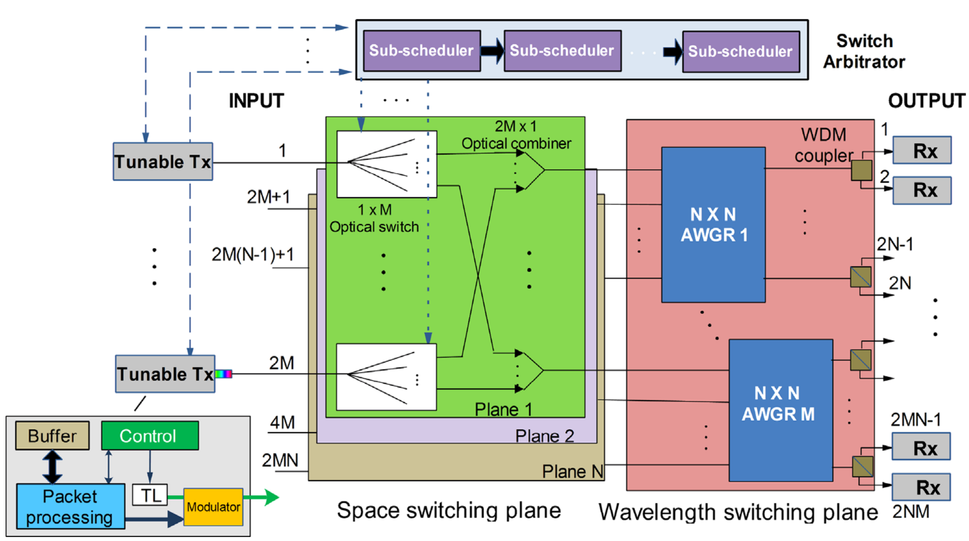

Fig. 1. Architecture of the space–wavelength routed switch. Tx: transmitter; Rx: receiver; N: number of space switching planes; M: number of $N \times N$ cyclic AWGRs.