Implementation of Optical Burst Switching in Data Centers

Mohammed Y. Sowailem, David V. Plant, and Odile Liboiron-Ladouceur
Photonics Systems Group, Electrical and Computer Engineering Department
McGill University, Montreal, QC H3A 2A7, Canada.

Abstract—We demonstrate the implementation and performance of optical burst switching in data centers. We propose the use of hybrid burst assembly algorithm and compare both the first-fit and horizon scheduling algorithms in a fat-tree topology.

I. INTRODUCTION

The implementation of optical networks in data centers requires cutting edge technologies that concerns researchers in Datacom. In [1], a hybrid electrical/optical architecture for data centers is proposed where optical circuit switching (OCS) scheme is used for the inter-pod switching. The main idea depends on the aggregated traffic characteristics. When the aggregated traffic is stable and continuous, the OCS scheme is used whereas the electrical packet switching is used when the aggregated traffic is bursty.

In this paper, we investigate the use of optical burst switching (OBS) techniques in data centers. OBS combines the advantages of both OCS and optical packet switching (OPS) [2]. The basic data unit in OBS is the burst, which can be as large as a session (as in OCS) or as small as a packet (as in OPS). This feature makes OBS a promising solution to achieve the all-optical switching in the inter-pod communication, since it can handle the stable and the bursty traffic.

II. BURST ASSEMBLY ALGORITHM

The burst comprises a number of packets in OBS networks. The construction of the data burst follows a certain assembly algorithm which assembles the data packets in the data burst according to their destination. There are several burst assembly algorithms; time-based, burst length-based and time/length-based (hybrid) assembly algorithms [3]. In the hybrid algorithm which is adopted in this paper, a timer is started as packets are assembled. If the burst size reaches a burst length threshold ($B_{th}$) before the time is up, the burst is sent. Otherwise, the burst is transmitted when the timer reaches a time threshold ($T_{th}$). The choice of the burst assembly algorithms and setting of its thresholds has a great impact on the network performance and on the shape of the burst arrival traffic at the core switches [2].

III. BURST SCHEDULING ALGORITHM

In this paper, centralized control is adopted [2]. In such data center network, the centralized scheduler has a global knowledge of the core layer switches status and the availability of the wavelengths at the input and output of those switches. The scheduler assigns a wavelength to each burst according to a scheduling algorithm. In this work, we are comparing two scheduling algorithms both exhibiting minimal time complexity [3]; first-fit and horizon scheduling algorithms. In the first-fit scheduling algorithm, the scheduler searches in a fixed order the wavelengths to find a wavelength not used by the source pod at the input link of a core switch and at the output link of that same core switch connected to the destination pod. In the horizon scheduling algorithm, the scheduler searches in a fixed order the wavelengths to find a wavelength not used by the source pod at the input link of a core switch and at the output link of that same core switch connected to the destination pod. The scheduler, in our approach, makes its decision based on what is the earliest wavelength that is free on that link.

IV. DATA CENTER NETWORK ARCHITECTURE

In this section, we describe the necessary modification in a data center architecture for the implementation of the OBS without changing the fat-tree topology proposed in [4]. In [5],...
the implementation of Labeled OBS with home circuits is proposed with significant change in the topology. They also argue that the contention problem in OBS networks impacts their implementation. In our proposed solution, the network topology conventionally used in data centers is maintained with only minor modifications in the architecture (Fig. 1(a)). Additionally, the contention problem is difficult to be resolved due to OBS networks signaling protocols typically based on one way reservation protocol [6]. However in data centers, centralized control is used due to the localization of the pods within the same physical location. This centralized control nature in data centers gives the potential of solving the contention problem with some enhancement in the scheduling algorithm.

The architectures of both the pod switch and the core switch are shown in Fig. 1(b) and Fig. 1(c), respectively. Essentially, pod switches in the data center serve as ingress and egress nodes in the OBS network, and core layer switches serve as core nodes in the OBS network. Thus, a set of \((N-1)\) buffers is needed at each pod switch where \(N\) is the number of pods. Also, a switch controlled by the scheduler is used between the \((N-1)\) buffers and the \((w \times S)\) laser sources (inside the E/O) where \(w\) is the number of wavelengths and \(S\) is the number of switches (Fig. 1(b)). After the burst is constructed, the switch is configured by the scheduler according to its decision about the core switch used and wavelength assigned to a burst. Inside the core switches, an optical cross connect is configured by the scheduler to switch the burst from the source pod to the destination pod according to the assigned wavelength by the scheduler (Fig. 1(c)).

V. SIMULATION DESCRIPTION AND RESULTS

In the simulation (using MATLAB R2010b), we assume that the traffic introduced by each of the 64 pods \((N = 64)\) to the core layer with 2 switches \((S = 2)\) has the same properties. We also assume that Ethernet frames are generated by the pods, such that the inter-arrival times between the Ethernet frames follow an exponential distribution. Finally, we assume that the Ethernet frame length follows an exponential distribution with a mean packet (frame) length of 2400 bits. However, when a frame is generated with a length less than the minimum frame length, its length is adjusted to be equal to the minimum frame length. Similarly, when the frame length is greater than the maximum frame length, it is adjusted to be equal to maximum frame length.

In the hybrid burst assembly algorithm, \(B_{th}\) is set to be ten times the mean packet length and \(T_{th}\) is set to be 1 million bit duration. Fig. 2 shows the effect of the burst assembly on the latency which reveals that even at low traffic when the burst contains on the average 2 or 3 packets, the average latency is still a small portion of the time threshold. This means that the latency is always bounded by \(T_{th}\), and hence, we can use \(T_{th}\) as a design parameter to achieve the QoS requirements for the latency in data centers. Furthermore, we can divide the traffic into classes according to their QoS requirements where \(B_{th}\) and \(T_{th}\) could be adjusted for each class.

The scheduler has two tasks; determine the switch configuration used between two communicating pods and the assigned wavelength \((w = 8)\) to the burst in the communication. To balance the load on both switches, we toggle the start of the search algorithm for an available wavelength after each wavelength allocation. Fig. 3 shows the burst loss probability versus the traffic load for both first-fit and horizon scheduling algorithms. The first-fit algorithm outperforms the horizon algorithm with orders of magnitude. This reveals that the change in the scheduling algorithm affects significantly the data center performance without modifying the data center fat-tree topology.

VI. CONCLUSION

We demonstrate the implementation of OBS in data centers with the required modification in the architecture while maintaining the fat-tree topology. Moreover, we show that the choice of the scheduling algorithm is a key design parameter that impacts the performance significantly. We find that the latency due to the burst assembly can be adjusted by setting the time threshold of the proposed hybrid burst assembly algorithm to meet a required QoS.

REFERENCES