Extreme Data-rate Scheduling for the Data Center

Neelakandan Manihatty Bojan, Noa Zilberman, Gianni Antichi, and Andrew W. Moore
Computer Laboratory, University of Cambridge
firstname.lastname@cl.cam.ac.uk

CCS Concepts

•Networks → Bridges and switches; Hybrid networks; Data center networks; •Hardware → Programmable logic elements; Emerging optical and photonic technologies;

Keywords

Data center networks; Optical networks; Switching; Scheduling

1. INTRODUCTION

Designing scalable and cost-effective data center interconnect architectures based on electrical packet switches (EPS) is very challenging [3]. Researchers have tried to explore optics and its advantages (in terms of bandwidth scaling, transmission speed, energy efficiency etc.) to address the challenges in data center environment [2]. This has resulted in various hybrid switching architectures [2, 5] for data centers, wherein an optical circuit switch (OCS) is used alongside an EPS. The OCS is used to serve long bursts of traffic and the EPS is used to serve the remaining traffic and short bursts.

The performance of the hybrid switching architecture depends on its scheduler [4]. The scheduler estimates demand based on the incoming traffic, computes switch configuration and maps the traffic on to either the EPS or OCS in a optimal fashion. Existing software based schedulers [2, 5] lack the speed and flexibility to cope with the faster switching technologies and increasing network demands [3]. Slow schedulers can negatively impact the performance of the data center network due to poor resource utilization. As networking requirements increase, the above problem is going to escalate, motivating the need for faster and scalable schedulers. One approach to do this would be to use a hardware based scheduler. We argue that the first step to achieve a good hybrid-switch scheduler is to have a framework for fast prototyping and evaluation of new hardware-based switch schedulers.

SIGCOMM '15 August 17-21, 2015, London, United Kingdom

ACM ISBN 978-1-4503-3542-3/15/08.
DOI: http://dx.doi.org/10.1145/2785956.2790019

2. MOTIVATION

With increasing data rates and emerging fast optical switching technologies [3][1], software based scheduling can no longer sustain the requirements (i.e., fast demand estimation and efficient schedules computation) of hybrid networks. Software based schedulers used in hybrid switching architectures [2, 5] operate in the order of milliseconds due to their inherent latency (delays during demand estimation, schedule calculation, Input/Output (IO) processing, propagation delay between host and switch). Software based schedulers also requires tight synchronization between the host and switch, which is difficult to achieve at faster switching times and higher transmission rates [3].

The optical switching time is the time taken by the optical device to configure its input and output ports based on the schedule, so that the incoming optical signals (packets) can be routed towards their destination. During the switching time (which can vary from nanoseconds to milliseconds based on its construction), no packets can be sent through the switch and hence need to be buffered.

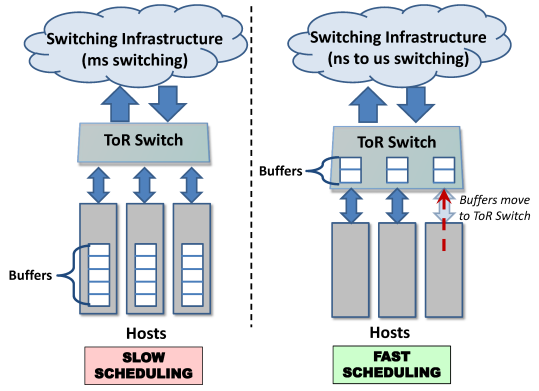

Figure 1: Host buffering vs Switch buffering

As an example, a switching infrastructure containing 64x64 input-queued switch (operating at a rate of 10 Gbps per port) with a millisecond switching time results in approximately gigabytes of buffering memory requirement in order to sustain bursts of traffic without losses. Such an amount of memory was not available in the Top of Rack (ToR) switches, forcing packets to be stored at the hosts (Figure 1, Slow Scheduling). As a result, packets stored in the host can be passed to the switch only at appropriate times, upon a grant from the scheduler. This increases design complexity,

end-to-end latency, and forces tight synchronization requirements between hosts and the switch. This can increase the overall traffic latency and jitter of widely used applications (i.e., VOIP, multiuser gaming etc.) and decrease the user quality of experience. As we move towards faster switching times, memory requirements diminish. Under the same configuration, a nanosecond switching time requires only kilobytes of buffering memory. This enables buffering packets directly in the ToR switch (see Figure 1, Fast Scheduling) and would remove issues relating to synchronization between the host and switch, thereby decreasing design complexity.

The scheduler is a key element that determines the performance of the data center network. With the availability of fast optical switches [3, 1] and increasing network demands, rapid scheduling is a necessity and not an option. Compared to its software counterparts, hardware based schedulers can match the speeds of fast optical switches and can be quick in responding to the dynamically varying network demand. This is inherent due to their hardware design: allowing quick demand estimation, fast schedule computation and rapid communication of computed schedules to the switch.

3. PROPOSED DESIGN

In the previous section we motivated the need for hardware based schedulers. Hardware may not be fast by default, but with proper implementation fast, high performance operation can be achieved. To this aim, we argue that the path towards the design and implementation of an optimal hardware scheduler requires a flexible framework for rapid prototyping, exploration and evaluation of novel hybrid schedulers. We aim to prototype the framework using a reconfigurable platform, NetFPGA-SUME [6]. The NetFPGA-SUME platform was designed for data center research, and enables the evaluation of new designs under real traffic workloads and with comparable performance.

We partition our design into processing logic, switching logic and scheduling logic as shown in Figure 2. The processing logic and switching logic are part of the infrastructure that is constant (yet configurable), and the users implement novel design in the scheduling logic module. Incoming packets from hosts H1, H2, .., Hn are sent to the processing logic. There, packets are classified into flows based on configurable look-up rules and places them into their respective Virtual Ouput Queue (VOQ). As the status of a VOQ changes, the subsystem generates scheduling requests and transmits packets upon receiving transmission grants from the scheduling logic. The scheduling logic processes the incoming requests, estimates the demand matrix, and runs the scheduling algorithm, generating corresponding transmission grants. Before providing a grant to the processing logic, the scheduler sends the grant matrix to the switching logic to configure the circuits in the OCS to match the grant matrix. Once the grant message is received by the processing logic, it dequeues packets from the respective VOQ and sends them to the OCS (that has already been configured according to the grant matrix) to be delivered to the respective destination. Based on the scheduling mechanism, residual traffic can be sent through the EPS. The scheme allows for multiple VOQs to be served at once, matching the port dimensions of the switching logic.

The design contains network interfaces, memory interfaces and various logical elements, omitted from the discussion for clarity. Individual partitions can be designed as separate entities and then integrated to realize a setup that emulates or uses commodity network devices. The resulting testbed enables us to explore scheduling architectures for hybrid switching, hybrid topologies for data center networks, synchronization issues, scalability and latency requirements in heterogeneous networks etc. It also allows to detect and analyse transient effects that may not be visible under simulation environments. The proposed architecture has the advantage of supporting both centralized and distributed implementations. A large testbed can be assembled, using tens of processing elements, a centralized scheduling entity and a commercial OCS. This implementation also allows to explore SDN practices over the hybrid network.

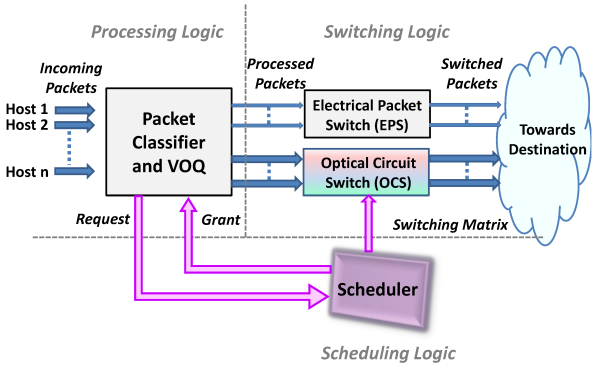

Figure 2: Proposed implementation

4. CONCLUSION

This paper motivates the need for hardware based schedulers in hybrid switches in order to meet emerging data center requirements. We have shown the main drawbacks that arise when using software based schedulers. We argue that the first step to achieve an optimal hybrid switch scheduler is to have a framework for rapid prototyping and assessment of new hardware-based scheduling algorithms. Finally, we show the architecture of the proposed framework, serving as an enabler for new scheduling algorithms.

5. ACKNOWLEDGEMENTS

This project is supported by the EPSRC INTERNET Project EP/H040536/1.

6. REFERENCES

[1] Epiphotonics. Nano-second speed plzt switch. http://www.epiphotonics.com/products3.htm. Accessed: 2015-05-08.

[2] N. Farrington et al. Helios: A hybrid electrical/optical switch architecture for modular data centers. In *SIGCOMM*. ACM, 2010.

[3] H. Liu et al. Circuit switching under the radar with reactor. In *NSDI*. USENIX, 2014.

[4] C. Raffaelli et al. Evaluation of packet scheduling in hybrid optical/electrical switch. *Photonic Network Communications*, 2012.

[5] G. Wang et al. c-through: part-time optics in data centers. *SIGCOMM CCR*, 2010.

[6] N. Zilberman et al. NetFPGA SUME: Toward 100 Gbps as Research Commodity. *Micro*, 2014.