Job Scheduling for Data-Parallel Frameworks with Hybrid Electrical/Optical Datacenter Networks*

Zhuozhao Li and Haiying Shen
Department of Computer Science, University of Virginia
{zl5uq,hs6ms}@virginia.edu

ABSTRACT

In spite of many advantages of hybrid electrical/optical datacenter networks (Hybrid-DCN), current job schedulers for data-parallel frameworks are not suitable for Hybrid-DCN, since the schedulers do not aggregate data traffic to facilitate using optical circuit switch (OCS). We propose SchedOCS, a job scheduler for data-parallel frameworks in Hybrid-DCN that aims to take full advantage of the OCS to improve the job performance.

CCS CONCEPTS

• **Networks → Data center networks**;

KEYWORDS

Optical circuit switch, parallelism, traffic aggregation

ACM Reference Format:
Zhuozhao Li and Haiying Shen. 2017. Job Scheduling for Data-Parallel Frameworks with Hybrid Electrical/Optical Datacenter Networks. In *Proceedings of SoCC '17, Santa Clara, CA, USA, September 24–27, 2017,* 1 pages.
https://doi.org/10.1145/3127479.3132694

1 INTRODUCTION

Several studies [1, 2, 4, 5] propose to augment the traditional electrical packet switch (EPS) datacenter network with an on-demand *rack-to-rack* network using the OCS (namely Hybrid-DCN), which has low capital expenditures (CapEx) and low operating expenditures (OpEx). However, OCS can be used only for large data transfers (e.g., 1.125GB) between racks, so that the overhead (on the order of μs-to-ms) used to *reconfigure* the input-to-output connections of OCS is negligible.

Current job schedulers for the data-parallel frameworks [3, 6] are not designed for the Hybrid-DCN and fail to use OCS to accelerate the data transfer. To take full advantage of Hybrid-DCN, we could aggregate the data to be transferred by placing the tasks of a job (e.g., map and reduce tasks in MapReduce) in only a few racks. However, it may sacrifice the basic principle of data-parallel frameworks – parallelism (i.e., the tasks of a job running concurrently). There is a tradeoff between parallelism and traffic aggregation. If a rack does not have sufficient available resources to run all the assigned tasks concurrently, it increases the latency of the job (i.e., the duration from the start of a job until its completion). In this paper, we propose a job scheduler called SchedOCS that aims to efficiently leverage OCS in Hybrid-DCN to improve job performance by finding the optimal tradeoff between the parallelism and traffic aggregation. SchedOCS attempts to aggregate the shuffle data transfers of a job in order to use OCS effectively.

2 SCHEDOCS DESIGN

SchedOCS consists of an offline scheduler and a real-time scheduler.
Offline scheduler The job profiler explores the tradeoff of shuffle-heavy recurring jobs based on the estimated job characteristics [3] and outputs all feasible schedules (i.e., number of racks to run the tasks) of a job that can leverage the OCS efficiently while achieving sufficient parallelism.

Then, the job manager enumerates all the feasible schedules of the recurring job and finds out a global schedule including the sequence to run the map/reduce tasks of recurring jobs in each rack that yields the best performance (i.e., high throughput for batch jobs and short completion time for online jobs).
Real-time scheduler Based on the schedule from offline scheduler, the real-time cluster scheduler places the input data and schedules the recurring jobs to the racks accordingly. The non-recurring jobs then use the idle resources that are not assigned to the recurring jobs. As the recurring jobs can finish earlier by more efficiently utilizing OCS, it leaves more computing resources and network bandwidth to ad-hoc jobs for them to complete earlier.

3 CONCLUSION AND FUTURE WORK

We propose SchedOCS, a job scheduler for data-parallel frameworks in Hybrid-DCN that aggregates the data transfers of a job to fully takes advantage of the OCS to improve job performance. As the future work, we plan to implement and evaluate SchedOCS in simulation and real cluster.

REFERENCES

[1] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen. 2012. OSA: An optical switching architecture for data center networks with unprecedented flexibility. In *Proc. of NSDI.*
[2] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat. 2010. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. In *Proc. of SIGCOMM.*
[3] V. Jalaparti, P. Bodik, I. Menache, S. Rao, K. Makarychev, and M. Caesar. 2015. Network-Aware Scheduling for Data-Parallel Jobs: Plan When You Can. In *Proc. of SIGCOMM.*
[4] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat. 2013. Integrating Microsecond Circuit Switching into the Data Center. In *Proc. of SIGCOMM.*
[5] G. Wang, D.G. Andersen, M. Kaminsky, K. Papagiannaki, TS Ng, M. Kozuch, and M. Ryan. 2010. c-Through: Part-time optics in data centers. In *Proc. of SIGCOMM.*
[6] M. Zaharia, D. Borthakur, S. Sen, K. Elmeleegy, S. Shenker, and I. Stoica. 2010. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In *Proc. of EuroSys.*

*This research was supported in part by U.S. NSF grants ACI-1719397 and CNS-1733596, and Microsoft Research Faculty Fellowship 8300751.

SoCC '17, September 24–27, 2017, Santa Clara, CA, USA

ACM ISBN 978-1-4503-5028-0/17/09.
https://doi.org/10.1145/3127479.3132694