Poster: Near Non-blocking Performance with All-optical Circuit-switched Core

Sushovan Das
Rice University

Arlei Silva
Rice University

T. S. Eugene Ng
Rice University

ABSTRACT

All-optical circuit-switched (OCS) core is the holy grail for the future generation datacenter architectures. However, such proposals consist of a common operational abstraction termed as round-robin circuit scheduling, which heavily suffers from a) high traffic skewness, and b) high volume of inter-rack traffic. To address this issue, we propose a novel architecture: round-robin OCS-core equipped with OCS-based reconfigurable edge for joint Skewness and Inter-rack traffic Volume (SV) minimization. Our architecture significantly improves the performance of all-optical cores, making it very close to a non-blocking network.

CCS CONCEPTS

• **Networks** → **Network architectures**.

KEYWORDS

All-optical, Datacenter, Network Architecture, Skewness

ACM Reference Format:
Sushovan Das, Arlei Silva, and T. S. Eugene Ng. 2023. Poster: Near Non-blocking Performance with All-optical Circuit-switched Core. In *ACM SIGCOMM 2023 Conference (ACM SIGCOMM '23), September 10, 2023, New York, NY, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3603269.3610868

1 MOTIVATION

Traditional packet-switched network cores in today's Datacenter networks (DCNs) are not sustainable in the long run as CMOS-based electrical packet switches face the challenge posed by the end of Moore's Law [1, 16]. Under such energy-critical situations, optical circuit-switching (OCS) technology equipped with several fundamental properties such as, a) agnostic to data-rate, b) negligible power consumption, c) negligible forwarding latency etc., seems to be the most promising alternative. This, in turn, fuels the necessity to

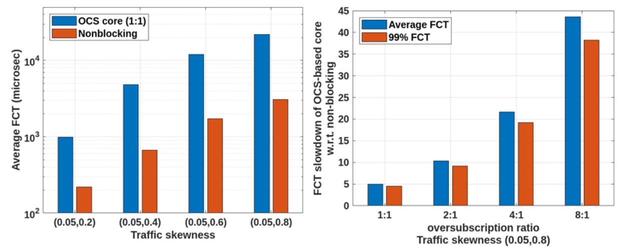

Figure 1: Round-robin OCS cores significantly suffer from (a) traffic skewness, and (b) oversubscription.

envision the all-optical circuit-switched cores [1, 11–13, 15] for designing the future-generation DCN architectures.

In spite of using diverse underlying OCS technologies, those DCN architectures share a common operational abstraction termed as round-robin circuit scheduling [5]. The OCSes are connected to a subset of ToR switches and periodically cycle through a predefined set of circuit configurations, to achieve reconfigurably non-blocking connectivity. Such abstraction can theoretically achieve 100% throughput only if the traffic is uniform. But, DCN workloads show a) high skewness i.e., a small subset of ToR pairs exchange a significant amount of traffic [6–10, 14, 18], and b) high inter-rack traffic volume [2–4, 14]. As a result, circuits between the hot rack pairs are heavily utilized, while the abstraction cannot leverage underutilized bandwidth of the cold circuits. The situation can become even worse if the core is oversubscribed, as the high volume of skewed inter-rack traffic would contend for bandwidth and face severe congestion.

To quantify this issue, we perform packet-level simulations emulating a round-robin OCS-based core such as Sirius [1], alongwith an ideal non-blocking network, in presence of highly skewed and inter-rack traffic, while varying the oversubscription (os) at the core, as shown in Figure 1(a) and 1(b). We define the skewness parameter (x, y) where x fraction of hot-rack pairs exchange y fraction of the traffic. For $1:1$ os, the average flow completion time (FCT) slowdown is 7.2×. The performance degrades rapidly with higher os ratio due to heavy inter-rack traffic volume. For example, at $8:1$ os, the average FCT slowdown is more than 43×.

2 OUR STRATEGY

In this work, we envision a fundamentally different approach: regroup the edge traffic intelligently so that a) most of the

ACM SIGCOMM '23, September 10, 2023, New York, NY, USA

ACM ISBN 979-8-4007-0236-5/23/09.
https://doi.org/10.1145/3603269.3610868

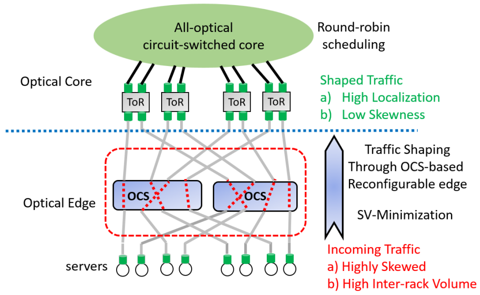

Figure 2: Round-robin OCS-based core with OCS-based reconfigurable edge, for joint SV-minimization.

traffic gets localized within a ToR which reduces the inter-rack traffic volume, and b) the remaining inter-rack traffic accessing the network core becomes almost uniform, a favorable scenario for round-robin OCS-core. Hence, we propose a novel architecture: round-robin OCS-core equipped with OCS-based reconfigurable edge between servers and ToR switches, as shown in Figure 2. OCS-based edge complements the traffic-agnostic round-robin OCS-core by periodically reconfiguring itself, thus reshaping the incoming traffic in order to jointly minimize traffic Skewness and inter-rack traffic Volume, which we call SV-minimization.

However, jointly optimizing for these objectives while finding the right balance is non-trivial and challenging. Example in Figure 3 intuitively demonstrates that naïve traffic localization could reduce inter-rack traffic volume significantly, although the remaining inter-rack traffic could be heavily skewed, affecting performance. Consider a OCS-core with 3 ToRs (with 2 servers each) and one edge OCS. For config (a), the traffic is completely inter-rack (localized traffic = 0) with high skewness (Jain's Fairness Index, JFI, of the inter-rack traffic is 0.18). Aggressive traffic localization would localize most of the flows (config (b)), minimizing the inter-rack traffic volume (localized traffic = 102000). However, it leaves both flows f_2 and f_3 between ToR 0 and ToR 2, leading to high traffic skewness (JFI = 0.17). But in config (c), only flow f_1 is localized, and other flows are deliberately not localized, rather reorganized between ToRs 1 and 2 in a uniform manner. This configuration sacrifices localization by a little (localized traffic = 100000) while reducing the inter-rack traffic skewness significantly (JFI = 0.67).

3 EVALUATION

Figure 4(a) and 4(b) show that, the performance of baseline architectures can be severely affected by higher core oversubscription in presence of high traffic skewness and heavy inter-rack traffic volume. At os ratio 8 : 1, round-robin OCS core can perform 43.6× worse compared to non-blocking network in terms of average FCT. At such high os ratio, VLB [17] and Opera [11] both improve upon round-robin OCS-based core, but still have average FCT slowdown of 13.12×

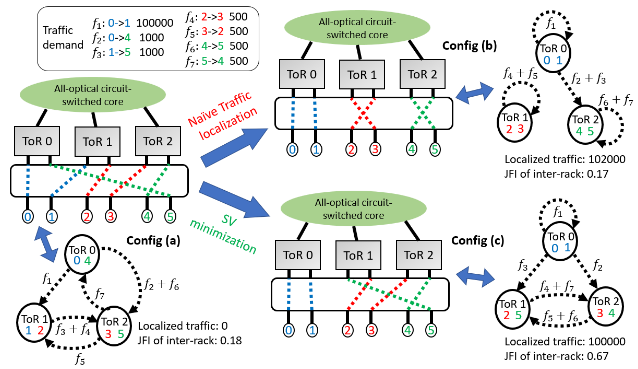

Figure 3: Example to demonstrate the difference between naïve traffic localization vs. SV-minimization.

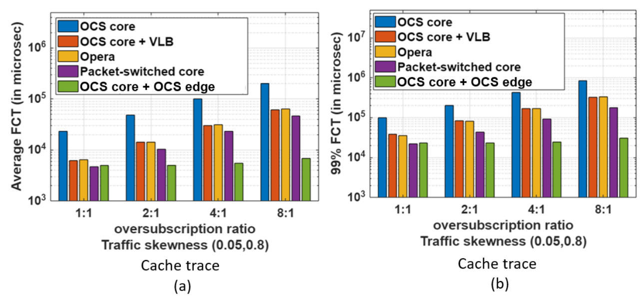

Figure 4: (a) Average FCT and (b) 99% **FCT (both in** μ**sec) of the architectures at different core os ratio for high traffic skewness** (0.05, 0.8) **and high network load** (80%).

and 13.8× respectively. A traditional packet-switched network can outperform all of these baseline architectures at high os ratio, due to path diversity and no OCS-based downtime. However, our proposed architecture, with one OCS at the edge having 10 μsec reconfiguration downtime and 500 μsec reconfiguration interval can achieve closest to the non-blocking performance under 1 : 1 os ratio. The average and 99% FCT slowdown are only 9.7% and 5.1% respectively w.r.t. non-blocking packet-switched core. Equipped with novel SV-minimization, our architecture also outperforms traditional packet-switched cores under high os ratios, e.g., it can effectively reduce the core os ratio from 8 : 1 to 1.4 : 1.

4 CONCLUSION

We propose a traffic-agnostic optical core alongwith a reconfigurable optical edge that jointly minimizes traffic skewness and inter-rack traffic volume. Therefore, it can significantly reduce the performance gap between today's all-optical core and ideal packet-switched nonblocking network, making all-optical cores widely acceptable to the community.

5 ACKNOWLEDGEMENT

We thank the anonymous reviewers for their insightful feedback. This work is partially supported by the NSF under CNS-2214272 and CNS-1815525.

REFERENCES

[1] Hitesh Ballani, Paolo Costa, Raphael Behrendt, Daniel Cletheroe, Istvan Haller, Krzysztof Jozwik, Fotini Karinou, Sophie Lange, Kai Shi, Benn Thomsen, et al. 2020. Sirius: A flat datacenter network with nanosecond optical switching. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 782–797.

[2] Theophilus Benson, Aditya Akella, and David A Maltz. 2010. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 267–280.

[3] Theophilus Benson, Ashok Anand, Aditya Akella, and Ming Zhang. 2009. Understanding data center traffic characteristics. In *Proceedings of the 1st ACM workshop on Research on enterprise networking*. ACM, 65–72.

[4] Peter Bodík et al. 2012. Surviving failures in bandwidth-constrained datacenters. In *SIGCOMM*. ACM.

[5] Sushovan Das, Weitao Wang, and TS Eugene Ng. 2021. Towards all-optical circuit-switched datacenter network cores: The case for mitigating traffic skewness at the edge. In *Proceedings of the ACM SIGCOMM 2021 Workshop on Optical Systems*. 1–5.

[6] Peter X Gao, Akshay Narayan, Sagar Karandikar, Joao Carreira, Sangjin Han, Rachit Agarwal, Sylvia Ratnasamy, and Scott Shenker. 2016. Network requirements for resource disaggregation. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 249–264.

[7] Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre-Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. 2016. Projector: Agile reconfigurable data center interconnect. In *Proceedings of the 2016 ACM SIGCOMM Conference*. 216–229.

[8] Qi Huang, Helga Gudmundsdottir, Ymir Vigfusson, Daniel A Freedman, Ken Birman, and Robbert van Renesse. 2014. Characterizing load imbalance in real-world networked caches. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*. 1–7.

[9] Simon Kassing, Asaf Valadarsky, Gal Shahaf, Michael Schapira, and Ankit Singla. 2017. Beyond fat-trees without antennae, mirrors, and disco-balls. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. 281–294.

[10] Zaoxing Liu, Zhihao Bai, Zhenming Liu, Xiaozhou Li, Changhoon Kim, Vladimir Braverman, Xin Jin, and Ion Stoica. 2019. Distcache: Provable load balancing for large-scale storage systems with distributed caching. In *17th USENIX Conference on File and Storage Technologies (FAST 19)*. 143–157.

[11] William M Mellette and Rajdeep Das. 2020. Expanding across time to deliver bandwidth efficiency and low latency. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*.

[12] William M Mellette, Rob McGuinness, Arjun Roy, Alex Forencich, George Papen, Alex C Snoeren, and George Porter. 2017. Rotornet: A scalable, low-complexity, optical datacenter network. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. 267–280.

[13] George Porter, Richard Strong, Nathan Farrington, Alex Forencich, Pang Chen-Sun, Tajana Rosing, Yeshaiahu Fainman, George Papen, and Amin Vahdat. 2013. Integrating microsecond circuit switching into the data center. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 447–458.

[14] Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C Snoeren. 2015. Inside the social network's (datacenter) network. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 123–137.

[15] Vishal Shrivastav, Asaf Valadarsky, Hitesh Ballani, Paolo Costa, Ki Suh Lee, Han Wang, Rachit Agarwal, and Hakim Weatherspoon. 2019. Shoal: A network architecture for disaggregated racks. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*. 255–270.

[16] Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, et al. 2015. Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network. *ACM SIGCOMM computer communication review* 45, 4 (2015), 183–197.

[17] Leslie G. Valiant. 1982. A scheme for fast parallel communication. *SIAM journal on computing* 11, 2 (1982), 350–361.

[18] Weitao Wang, Dingming Wu, Sushovan Das, Afsaneh Rahbar, Ang Chen, and TS Eugene Ng. 2022. {RDC}:{Energy-Efficient} Data Center Network Congestion Relief with Topological Reconfigurability at the Edge. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. 1267–1288.