Duo: A High-Throughput Reconfigurable Datacenter Network Using Local Routing and Control

Johannes Zerwas
TUM School of Computation, Information and Technology, Technical University of Munich
Munich, Germany
johannes.zerwas@tum.de

Csaba Györgyi
University of Vienna & ELTE Eötvös Loránd University
Austria and Hungary
gycsaba96@inf.elte.hu

Andreas Blenk
Siemens AG
Munich, Germany
andreas.blenk@tum.de

Stefan Schmid
TU Berlin & Fraunhofer SIT
Berlin, Germany
stefan.schmid@tu-berlin.de

Chen Avin
School of Electrical and Computer Engineering, Ben-Gurion University of the Negev
Beer-Sheva, Israel
avin@cse.bgu.ac.il

ABSTRACT

The performance of many cloud-based applications critically depends on the capacity of the underlying datacenter network. A particularly innovative approach to improve the throughput in datacenters is enabled by emerging optical technologies, which allow to dynamically adjust the physical network topology, both in an oblivious or demand-aware manner. However, such topology engineering, i.e., the operation and control of dynamic datacenter networks, is considered complex and currently comes with restrictions and overheads.

We present Duo, a novel demand-aware reconfigurable rack-to-rack datacenter network design realized with a simple and efficient control plane. Duo is based on the well-known de Bruijn topology (implemented using a small number of optical circuit switches) and the key observation that this topology can be enhanced using dynamic ("opportunistic") links between its nodes.

In contrast to previous systems, Duo has several desired features: i) It makes effective use of the network capacity by supporting integrated and multi-hop routing (paths that combine both static and dynamic links). ii) It uses a work-conserving queue scheduling which enables out-of-the-box TCP support. iii) Duo employs greedy routing that is implemented using standard IP longest prefix match with small forwarding tables. And iv) during topological reconfigurations, routing tables require only local updates, making this approach ideal for dynamic networks.

We evaluate Duo in end-to-end packet-level simulations, comparing it to the state-of-the-art static and dynamic networks designs. We show that Duo provides higher throughput, shorter paths, lower flow completion times for high priority flows, and minimal packet reordering, all using existing network and transport layer protocols. We also report on a proof-of-concept implementation of Duo's control and data plane.[1]

SIGMETRICS '23 Abstracts, June 19–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0074-3/23/06.
https://doi.org/10.1145/3578338.3593537

ACM Reference Format:
Johannes Zerwas, Csaba Györgyi, Andreas Blenk, Stefan Schmid, and Chen Avin. 2023. Duo: A High-Throughput Reconfigurable Datacenter Network Using Local Routing and Control. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '23 Abstracts), June 19–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3578338.3593537

1 INTRODUCTION

The performance of many cloud applications, e.g., related to distributed machine learning, batch processing, or streaming, critically depends on the bandwidth capacity of the underlying network topology. Accordingly, over the last years, great efforts have been made to improve the throughput of datacenter networks.

Emerging optical technologies enable what is now known as *topology engineering*, a particularly innovative approach to improve datacenter performance, by supporting dynamic and real-time reconfigurations of the physical network topology [5]. In particular, advanced optical circuit switches enable dynamic physical topologies by providing dynamic input-output ports *matchings* [1, 2, 4, 6]. Reconfigurable datacenter networks (RDCNs) use such switches to establish topological shortcuts (i.e., shorter paths) between racks, hence utilizing available bandwidth capacity more efficiently and improving throughput [4, 6].

RDCNs come in two flavors: *oblivious* and *demand-aware* [5]. Oblivious RDCNs such as Opera [6], Sirius [2] or MARS [1] rely on quickly and periodically changing interconnects between racks, to emulate a complete graph. In contrast, demand-aware RDCNs allow to *optimize* topological shortcuts, that depend on the traffic pattern. Demand-aware networks such as ProjecToR [3] are attractive since datacenter traffic typically features much temporal and spatial structure: traffic is bursty and skewed, and a large fraction of communicated bytes belong to a small number of elephant flows [3]. By adjusting the datacenter topology to support such flows, e.g., by

[1] Extended abstract. The full version of this paper appears in [7].

providing direct connectivity between intensively communicating source and destination racks, network throughput can be increased further.

However, the operation of RDCNs comes with overheads and limitations. In general, existing RDCNs typically rely on a hybrid topology which combines static (electrical) and dynamic (optical) parts. While such a combination is powerful [3], current architectures support only fairly restricted routing. First, communication on the (dynamic) optical topology is often limited to one or two hops. This constrains the possible path diversity, and hence capacity, of the optical network. Furthermore, routing is usually *segregated*: flows are either only forwarded along the static or the dynamic network, but not a combination of both [3, 4]. The restriction to segregated routing also entails overheads as it requires significant buffering while the reconfigurable links are not available, *non* work-conserving scheduling, and a more complex buffer management

This paper is motivated by the desire to overcome these limitations, and to better exploit the available link resources, by supporting a general *multi-hop* and *integrated* (i.e., non-segregated) routing. To this end, we propose a simpler and more efficient control plane for RDCNs which avoids packet forwarding delays by supporting *local and greedy integrated routing*: the forwarding rules depend on local information only, i.e., the set of direct neighbors as well as information in the packet header; they are hence not affected by topological changes in other parts in the network and do not have to be updated under such reconfigurations. This can significantly reduce control plane overheads during topological adjustments, maintaining a simple routing, buffering and control, and is hence well-suited for highly dynamic networks.

2 THE DUO RDCN DESIGN

We present Duo, a novel demand-aware RDCN which leverages such a local control plane using a (static) de Bruijn topology (built from a small number of optical switches). In order to maximize performance, Duo uses dynamic, demand-aware links to provide shorter paths for elephant flows, while other flows are transmitted via the combined (static + dynamic) topology. This *integrated multi-hop* and work-conserving routing across both switch types is a key feature of Duo and distinguishes it from previous works that rely on segregated and single-hop forwarding for demand-aware links [3, 4]. To achieve this, Duo relies on two observations. First, the de Bruijn topology supports greedy routing from a source to a destination based solely on the destination address. In addition, this routing can be realized via a simple forwarding table with longest prefix matching. Second, adding shortcuts to a static de Bruijn topology allows to continue supporting this greedy local routing. This enables a high update rate at low overheads.

Duo uses IP-based logical addressing. Its control plane can be realized using centralized or distributed algorithms. For instance, operating a receiver-based approach for the efficient detection of elephant flows as well as the local and collision-free scheduling of demand-aware links. Due to its simplicity Duo is well-suited to be realized, e.g., using the Sirius [2] architecture which was originally designed to be demand-oblivious, but can potentially support demand-aware link scheduling. Thereby, it benefits from the low cost and power consumption of the Sirius architecture.

3 KEY EVALUATION RESULTS

We evaluate Duo under various network configurations and traffic scenarios using packet-level simulations. These simulations demonstrate that Duo's properties allow us to use TCP out-of-the-box, without the need to develop new transport protocols. Moreover, we obtain the following empirical findings:

- Duo provides a higher throughput compared to the state-of-the-art, static and dynamic, networks.
- Duo preserves competitive flow completion times (FCT). It effectively trades off the higher throughput and better FCTs for small flows against the FCTs of medium-sized flows. Moreover, it has only a moderate amount of packet reordering.
- Assessing the characteristics of used forwarding paths demonstrates that Duo's higher throughput indeed stems from the *integrated multi-hop* routing.
- The advantages of Duo persist also for larger topologies as well as variations of the traffic patterns.

We further report on a proof-of-concept implementation of the control and data plane of Duo that demonstrates the feasibility of Duo with standard network stacks. The implementation builds around a single P4 switch which we use to emulate a scenario with 16 ToRs. As a contribution to the research community, we release our implementation together with this paper: https://github.com/tum-lkn/duo-simulator.

ACKNOWLEDGEMENTS

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, consolidator project Self-Adjusting Networks (AdjustNet), grant agreement No. 864228, Horizon 2020, 2020-2025. The work was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 438892507.

REFERENCES

[1] Vamsi Addanki, Chen Avin, and Stefan Schmid. 2023. Mars: Near-Optimal Throughput with Shallow Buffers in Reconfigurable Datacenter Networks. *Proc. ACM Meas. Anal. Comput. Syst.* 7, 1, Article 2 (mar 2023), 43 pages. https://doi.org/10.1145/3579312

[2] 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 *Proc. ACM SIGCOMM 2020 Conference*. 782–797.

[3] 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 *Proc. ACM SIGCOMM 2016 Conference*. ACM, 216–229.

[4] Chen Griner, Johannes Zerwas, Andreas Blenk, Manya Ghobadi, Stefan Schmid, and Chen Avin. 2021. Cerberus: The Power of Choices in Datacenter Topology Design (A Throughput Perspective). *Proc. ACM Meas. Anal. Comput. Syst.* 5, 3, Article 38 (dec 2021), 33 pages.

[5] Matthew Nance Hall, Klaus-Tycho Foerster, Stefan Schmid, and Ramakrishnan Durairajan. 2021. A Survey of Reconfigurable Optical Networks. *Optical Switching and Networking* 41 (2021), 100621.

[6] William M Mellette, Rajdeep Das, Yibo Guo, Rob McGuinness, Alex C Snoeren, and George Porter. 2020. Expanding across time to deliver bandwidth efficiency and low latency. In *Proc. 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. 1–18.

[7] Johannes Zerwas, Csaba Györgyi, Andreas Blenk, Stefan Schmid, and Chen Avin. 2023. Duo: A High-Throughput Reconfigurable Datacenter Network Using Local Routing and Control. *Proc. ACM Meas. Anal. Comput. Syst.* 7, 1, Article 20 (mar 2023), 25 pages. https://doi.org/10.1145/3579449