Cerberus: The Power of Choices in Datacenter Topology Design - A Throughput Perspective

Cerberus: The Power of Choices in Datacenter Topology Design*

A Throughput Perspective

Chen Griner
School of Electrical and Computer Engineering, Ben-Gurion University of the Negev
Beer-Sheva, Israel
griner@post.bgu.ac.il

Johannes Zerwas
Department of Electrical and Computer Engineering, Technical University of Munich
Munich, Germany
johannes.zerwas@tum.de

Andreas Blenk
Department of Electrical and Computer Engineering, Technical University of Munich & Faculty of Computer Science, University of Vienna
Munich, Germany
andreas.blenk@tum.de

Manya Ghobadi
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
Boston, USA
ghobadi@csail.mit.edu

Stefan Schmid
Faculty of Electrical Engineering and Computer Science, TU Berlin & University of Vienna
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 bandwidth and latency requirements of modern datacenter applications have led researchers to propose various topology designs using static, dynamic demand-oblivious (rotor), and/or dynamic demand-aware switches. However, given the diverse nature of datacenter traffic, there is little consensus about how these designs would fare against each other. In this work, we analyze the throughput of existing topology designs under different traffic patterns and study their unique advantages and potential costs in terms of bandwidth and latency "tax". To overcome the identified inefficiencies, we propose Cerberus, a unified, two-layer leaf-spine optical datacenter design with three topology types. Cerberus systematically matches different traffic patterns with their most suitable topology type: e.g., latency-sensitive flows are transmitted via a static topology, all-to-all traffic via a rotor topology, and elephant flows via a demand-aware topology. We show analytically and in simulations that Cerberus can improve throughput significantly compared to alternative approaches and operate datacenters at higher loads while being throughput-proportional.

ACM Reference Format:
Chen Griner, Johannes Zerwas, Andreas Blenk, Manya Ghobadi, Stefan Schmid, and Chen Avin. 2022. Cerberus: The Power of Choices in Datacenter Topology Design: A Throughput Perspective. In *Abstract Proceedings of the 2022 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS/PERFORMANCE '22 Abstracts), June 6–10, 2022, Mumbai, India.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3491050

*The first two authors contributed equally to this work.

SIGMETRICS/PERFORMANCE '22 Abstracts, June 6–10, 2022, Mumbai, India.

ACM ISBN 978-1-4503-9141-2/22/06.
https://doi.org/10.1145/3491050

1 INTRODUCTION

Datacenter networks have become a critical backbone of our digital society. The traffic these networks need to serve is growing explosively, and researchers are exploring novel *optical* datacenter networks, including both *static* and *dynamic* topologies. The architectural design choices of operators are expanding. In particular, there exist several fundamentally different optical datacenter topologies, relying on different switching technologies. We classify these topologies along the two independent dimensions, *static vs dynamic* and *demand-oblivious vs demand-aware* and we identify *three* main topology types: (i) Traditionally, datacenter networks are based on *static* and *demand-oblivious* topologies, e.g., Clos and expander graphs [1], (ii) More recent proposals also explore *dynamic* but *demand-oblivious* topologies, e.g., relying on rotor switches that periodically reconfigure the topology [2].(iii) Furthermore, there are *dynamic* and *demand-aware* topologies that can be reconfigured according to the current traffic pattern.However, there is little consensus in the networking community on how these different designs fare against each other [1], in particular when it comes to *throughput* [3]. What is more, we currently lack a unified model and analytical tools to close this gap.

Matching traffic patterns to topology designs. This paper is motivated by the observation that existing datacenter network designs in many cases provide a *mismatch* between some common traffic patterns and the switching technology used in the network topology to serve them. Different optical topologies provide different tradeoffs and the used topology should depend on the demand. For example, mice flows that are time-sensitive should be served on a static topology since reconfiguration times may violate their latency constraints. Elephant flows, may benefit from dynamic demand-aware topologies. Since the reconfiguration time is relatively small compared to the transmission time of such large flows, the reconfiguration can be amortized, and throughput improved by establishing direct links between frequently communicating pairs.

A key concern: throughput. Our main metric of interest is the end-to-end *throughput* supported by systems in a fluid-flow model. We follow the throughput definition by Jyothi et al. [3] and focus on the network topology. The *throughput* of a system for a traffic demand matrix T is the highest scaling factor $\theta(T)$ for which the traffic is feasible in the system. That is, we seek the maximum scaling factor for which there exists a feasible multi-commodity *network flow* assignment that routes the traffic in the matrix $\theta(T) \cdot T$ through the network from each source to each destination. A feasible flow means that the flow rates are subject to link capacities and to flow conservation at each intermediate node. The throughput of a topology, denoted by θ^*, is defined as the worst-case throughput among all traffic matrices [3].

Potential inefficiencies: bandwidth and latency tax. To compare the throughput of different topologies, we propose to quantify their inefficiencies in terms of *taxes*. Static topologies require *multi-hop* forwarding. This can be problematic, especially for large flows: the more hops a flow must traverse, the more network capacity is consumed. As noticed in prior work [2], inefficiency arising from multi-hop routing can be seen as a "bandwidth tax" (BW-Tax). In contrast, in dynamic networks, the topology may be reconfigured to provide *direct* connectivity to elephant flows at the cost of a reconfiguration time. In general, we can regard the reconfiguration time as a "latency tax" (LT-Tax). Regarding the design choice between static vs dynamic topology, we therefore observe that whereas *dynamic topologies introduce a latency tax, static topologies introduce a bandwidth tax*. In this paper, we present a *systematic* approach to study the throughput of both static and dynamic topologies for different traffic types. We make the following contributions.

Contribution 1: Throughput analysis including bandwidth and latency tax. We first extend the throughput definition by Jyothi et al. [3] from a demand matrix T to our general traffic generation model $\mathcal{T}$. We further extend the definition to apply also to *dynamic* network topologies rather than only static topologies. In turn, we present a mathematical framework that allows us to evaluate analytically the performance and trade-offs of arbitrary demands using different optical switches, considering real traffic distributions. In particular, based on our models, we formally show that all three topology *types* have a unique raison d'etre. In contrast to previous work , we propose to study the throughput via *demand completion time* (DCT), which allows us to incorporate *both* bandwidth and latency tax into our analysis. We further show that the efficiency of different topology types depends on the *skewness* of the traffic distribution in a non-trivial manner. This enables us to provide novel insights into an abstract version of existing architectures such as *rotor-net* for RotorNet [2] and *expander-net* for Xpander [1], including their throughput.

Contribution 2: A unified network model. We formalize the forte of three popular optical topologies from prior work: a static expander-based [1] , a rotor-based [2], and a demand-aware topology [4]. Based on these models, we propose a unified two-layer leaf-spine network model, called ToR-Matching-ToR (TMT), that simultaneously contains both *static* and *dynamic* topologies, and in particular, the three topology parts: static, rotor, and demand-aware. This generalizes the existing architectures mentioned above since the spine switches can be of *different* types.

	expander-net	*rotor-net*	Cerberus
BW-Tax	✓	✓	✗
LT-Tax	✗	✓	✓
θ^*	0.53	0.45	Open
Datamining	0.53	0.6	0.8 (+33%)
Permutation	0.53	0.45	≈ 1 (+88%)
Case Study	0.53	0.66	0.9 (+36%)

Table 1: The main topology designs we consider in this paper and their related properties and achieved throughput (with our improvement in %).

Contribution 3: Matching traffic to topologies with Cerberus. Motivated by the identified inefficiencies resulting from a mismatch between traffic and topology, we describe a novel architecture, Cerberus[1], which can serve traffic on the topology that best *matches* the traffic's characteristics. For example, latency-sensitive mice flows can be transmitted via static switches, all-to-all traffic via dynamic rotor switches, and elephant flows via demand-aware switches. Table 1 gives an overview of our corresponding contributions. For example, our simulations show that compared to static and dynamic demand-oblivious topologies, Cerberus improves the throughput by 33% or more for a datamining workload, by 36% or more for a synthetic case study imitating realistic flow-size distributions, and by 88% or more for permutation matrices, i.e., sparse matrices which represent the worst-case input for oblivious designs. We further prove that Cerberus is throughput-proportional [1], namely, that Cerberus is able to utilize its full capacity proportionally, even when only a subset of the servers generates traffic.

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 program (grant agreement No 864228 - AdjustNet). The work was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 438892507. The authors alone are responsible for the content of the paper. This work is part of the PhD thesis of the first two authors. A full version of this paper appears in [5].

REFERENCES

[1] S. Kassing, A. Valadarsky, G. Shahaf, M. Schapira, and A. Singla, "Beyond fat-trees without antennae, mirrors, and disco-balls," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, pp. 281–294, ACM, 2017.

[2] W. M. Mellette, R. McGuinness, A. Roy, A. Forencich, G. Papen, A. C. Snoeren, and G. Porter, "Rotornet: A scalable, low-complexity, optical datacenter network," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, pp. 267–280, ACM, 2017.

[3] S. A. Jyothi, A. Singla, P. B. Godfrey, and A. Kolla, "Measuring and understanding throughput of network topologies," in *SC'16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 761–772, IEEE, 2016.

[4] M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick, and D. Kilper, "Projector: Agile reconfigurable data center interconnect," in *Proceedings of the 2016 ACM SIGCOMM Conference*, pp. 216–229, ACM, 2016.

[5] C. Griner, J. Zerwas, A. Blenk, M. Ghobadi, S. Schmid, and C. Avin, "Cerberus: The power of choices in datacenter topology design-a throughput perspective," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 5, no. 3, pp. 1–33, 2021.

[1] In Greek mythology, Cerberus is a dog with three heads (corresponding to our three topology parts).