Improved Scalability of Demand-Aware Datacenter Topologies With Minimal Route Lengths and Congestion*

Extended Abstract

Maciej Pacut
Faculty of Computer Science
University of Vienna, Austria
maciej.pacut@univie.ac.at

Wenkai Dai
Faculty of Computer Science
University of Vienna, Austria
wenkai.dai@univie.ac.at

Alexandre Labbe
Institut Polytechnique de Paris
ENSTA Paris, France
alexandre.labbe@ensta-paris.fr

Klaus-Tycho Foerster
Department of Computer Science
TU Dortmund, Germany
klaus-tycho.foerster@tu-dortmund.de

Stefan Schmid
TU Berlin, Germany
University of Vienna, Austria
Fraunhofer SIT, Germany
stefan.schmid@tu-berlin.de

ABSTRACT

The performance of more and more cloud-based applications critically depends on the performance of the interconnecting datacenter network. Emerging reconfigurable datacenter networks have the potential to provide an unprecedented throughput by dynamically reconfiguring their topology in a demand-aware manner. This paper studies the algorithmic problem of how to design low-degree and hence scalable datacenter networks optimized toward the current traffic they serve. Our main contribution is a novel network design which provides asymptotically minimal route lengths and congestion. In comparison to prior work, we reduce the degree requirements by a factor of four for sparse demand matrices. We further show the problem to be already NP-hard for tree-shaped demands, but permits a 2-approximation on the route lengths and a 6-approximation for congestion. We further report on a small empirical study on Facebook traces.

Keywords

Network design, reconfigurable networks, demand-awareness, congestion and route lengths, approximation algorithms

1. INTRODUCTION

As the performance of many data-centric and cloud-based applications increasingly depends on the underlying networks, datacenter networks have become a critical infrastructure of our digital society. Indeed, current application trends introduce stringent performance requirements and a demand for datacenter networks providing ultra-low latency and high bandwidth. For example, emerging distributed machine learning applications which use highspeed computational devices, periodically require large data transfers during which the network can become the bottleneck.

Another example is today's trend of resource disaggregation in datacenters, which introduces a need for very fast access to remote resources (GPU, memory, disk) [23]. Traces of jobs from a Facebook cluster reveal that network transfers on average account for a third of the execution time [26].

Demand-aware networks are particularly motivated by empirical studies showing that communication patterns feature much structure. Indeed, traffic matrices (a.k.a. demand matrices) are often sparse and skewed in datacenters [5, 17, 21]. This introduces optimization opportunities, which stands in stark contrast to traditional, demand-oblivious datacenter network designs [18, 22, 29].

Emerging reconfigurable datacenter topologies, enabled by novel optical technologies, introduce new opportunities to significantly improve datacenter performance [16, 19]. In particular, by dynamically establishing topological shortcuts, reconfigurable datacenter networks allow to overcome the cost (or "tax" [24]) of multihop routing [9, 25], or to improve the flow completion time of elephant flows by directly connecting frequently communicating racks, in a demand-aware manner [4, 8, 9, 11, 12, 17, 20, 30].

This paper studies a fundamental algorithmic problem underlying such reconfigurable networks: how to design a demand-aware topology which, given a demand matrix, provides short topological routes between frequently communicating nodes (e.g., top-of-rack switches [6]), also minimizing congestion. For scalability reasons and as reconfigurable hardware consumes space and power, the interconnecting network should be of *low degree*, ideally a small constant.

The prior research to this problem is by Avin et al. [2, 4] who investigate demand-aware network designs of bounded degree, providing several interesting approximation algorithms, in particular a constant-approximation for the weighted route length objective for sparse demands. The paper already had several followups, e.g., a robust demand-aware network has been proposed in [7], and a version which also minimizes congestion in [3].

2. OUR CONTRIBUTION

Our contributions revolve around the design of demand-aware networks (BNDs) under a degree restriction, which asymptotically minimize communication cost and congestion, especially when the demand matrix induces a sparse graph or tree. In particular, we present an algorithm to de-

*Supported by the European Research Council (ERC), grant agreement No. 864228 (AdjustNet), Horizon 2020, 2020-25.

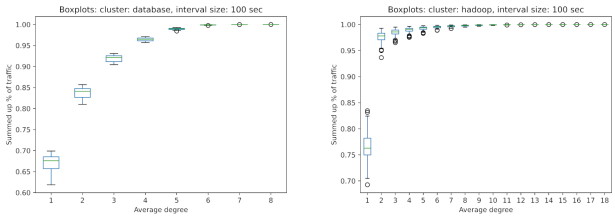

Figure 1: Boxplots showing % of traffic covered by different avg. degrees for Facebook's database (left) and hadoop (right) cluster at the pod level, with 110 and 109 nodes.

sign a network of maximum degree $3\Delta_{\text{avg}} + 8$ with asymptotically optimal route lengths and congestion, when the demand matrix is induced by a sparse graph of an average degree Δ_{avg}. This reduces the required maximum degree of the network by a factor of $4\times$ compared to previous work [2, 3].

We also show that the demand-aware network design problem is NP-hard, already when ignoring congestion and if both the demand itself and the network topology are restricted to be trees; prior work only established the hardness for general demands [1], respectively in hybrid [13, 14, 15] or geometric [10] settings. We moreover prove that optimizing for congestion, independent of route lengths, is NP-hard as well. On the positive side, we show that for tree-demands, one can jointly 2-approximate the optimal route lengths and 6-approximate the minimum congestion. Our results significantly improve the approximation ratio of route length from $\log^2(\Delta_{\max} + 1)$ to 2, where $\Delta_{\max}$ is the maximum degree of the designed network [3]. Comparing to similar approaches [2, 3, 4], which proposed *ego-trees* to reduce the degree while preserving distances, we present a tree called *Round Robin Tree* that is particularly well suited to jointly minimize weighted route length and congestion, and which we can interconnect with other trees in a low-degree manner.

Finally, we provide empirical insights into the practicality of our approach, considering traffic traces from Facebook [28]. As shown in, e.g., Fig. 1, nearly all traffic can be covered by *sparse* demand graphs of low average degree.

The full version of this paper has been published in Performance Evaluation (PEVA) [27].

3. REFERENCES

[1] R. Andrade et al. Minimum linear arrangements. *Elect. Notes in Discr. Mathematics*, 62:63 – 68, 2017.

[2] C. Avin et al. Demand-aware network designs of bounded degree. In *Proc. DISC*, 2017.

[3] C. Avin et al. Demand-aware network design with minimal congestion and route lengths. In *Proc. IEE INFOCOM*, 2019.

[4] C. Avin et al. Demand-aware network designs of bounded degree. *Distributed Comput.*, 33(3-4):311–325, 2020.

[5] C. Avin et al. On the complexity of traffic traces and implications. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(1):20:1–20:29, 2020.

[6] C. Avin et al. An online matching model for self-adjusting tor-to-tor networks. *arXiv:2006.11148*, 2020.

[7] C. Avin, A. Hercules, A. Loukas, and S. Schmid. *rDAN*: Toward robust demand-aware network designs. *Inf. Process. Lett.*, 133:5–9, 2018.

[8] N. H. Azimi et al. Firefly: a reconfigurable wireless data center fabric using free-space optics. In *Proc. ACM SIGCOMM*, 2014.

[9] H. Ballani et al. Sirius: A flat datacenter network with nanosecond optical switching. In *Proc. ACM SIGCOMM*, 2020.

[10] E. Ceylan et al. Demand-aware plane spanners of bounded degree. In *Proc. IFIP Networking*, 2021.

[11] K. Chen et al. OSA: an optical switching architecture for data center networks with unprecedented flexibility. *IEEE/ACM Trans. Netw.*, 22(2), 2014.

[12] W. Dai et al. Load-optimization in reconfigurable networks: Algorithms and complexity of flow routing. *SIGMETRICS Perform. Eval. Rev.*, 48(3):39–44, 2020.

[13] T. Fenz et al. Efficient non-segregated routing for reconfigurable demand-aware networks. *Comput. Commun.*, 164:138–147, 2020.

[14] K. Foerster et al. Characterizing the algorithmic complexity of reconfigurable data center architectures. In *Proc. ACM ANCS*, 2018.

[15] K. Foerster et al. On the complexity of non-segregated routing in reconfigurable data center architectures. *Comput. Commun. Rev.*, 49(2):2–8, 2019.

[16] K. Foerster and S. Schmid. Survey of reconfigurable data center networks: Enablers, algorithms, complexity. *SIGACT News*, 50(2):62–79, 2019.

[17] M. Ghobadi et al. Projector: Agile reconfigurable data center interconnect. In *Proc. ACM SIGCOMM*, 2016.

[18] C. Guo et al. Bcube: a high performance, server-centric network architecture for modular data centers. In *Proc. ACM SIGCOMM*, 2009.

[19] M. N. Hall et al. A survey of reconfigurable optical networks. *Opt. Switch. Netw.*, 41:100621, 2021.

[20] S. Kandula et al. Flyways to de-congest data center networks. *Proc. ACM HotNets*, 2009.

[21] S. Kandula et al. The nature of data center traffic: measurements & analysis. In *Proc. ACM IMC*, 2009.

[22] S. Kassing et al. Beyond fat-trees without antennae, mirrors, and disco-balls. In *Proc. SIGCOMM*, 2017.

[23] Y. Li et al. Hpcc: high precision congestion control. In *Proc. ACM SIGCOMM*, 2019.

[24] W. M. Mellette et al. Rotornet: A scalable, low-complexity, optical datacenter network. In *Proc. ACM SIGCOMM*. ACM, 2017.

[25] W. M. Mellette et al. Expanding across time to deliver bandwidth efficiency and low latency. In *Proc. USENIX NSDI*, 2020.

[26] J. C. Mogul and L. Popa. What we talk about when we talk about cloud network performance. *Comput. Commun. Rev.*, 42(5):44–48, 2012.

[27] M. Pacut, W. Dai, et al. Improved scalability of demand-aware datacenter topologies with minimal route lengths and congestion. *Performance Evaluation*, 152:102238, 2021.

[28] A. Roy et al. Inside the social network's (datacenter) network. In *Proc. ACM SIGCOMM*, 2015.

[29] A. Singh et al. Jupiter rising: a decade of clos topologies and centralized control in google's datacenter network. *Comm. ACM*, 59(9):88–97, 2016.

[30] X. Zhou et al. Mirror mirror on the ceiling: flexible wireless links for data centers. In *Proc. ACM SIGCOMM*, 2012.