Analysis and Optimization of Data Center Networks for Flexibility

Johannes Zerwas
Technical University of Munich, Germany
johannes.zerwas@tum.de

***Abstract*—Today‘s communication networks are facing diverse demands which create the necessity for more flexible networks. In particular, data center networks are confronted with a plethora of time-varying requirements. A recent trend in data center networking research is the development of architectures that can adapt their topology at runtime to account for changing demands, e.g., using Optical Circuit Switches. This thesis will evaluate how such architectures can explicitly optimize the flexibility of data center networks.**

I. Introduction

The emerging digitalization of society and industry turns data centers (DC) into a crucial part of the infrastructure. The rise of Cloud Computing, Network Virtualization and a broad range of applications from web browsing to Internet of Things along with increased user mobility lead to time-varying requirements for DC networks. On one hand, low performance and inefficiency results in revenue loss for operators [1]. On the other hand, many DC networks have oversubscribed physical resources to reduce CAPEX. To satisfy the dynamic requirements, operators might have to adapt networks, e.g., by re-routing flows, re-locating functions or scaling capacities.

While these operations work on logical (virtual) level a recent trend in DC networking research are architectures, that use flexible links and adapt the network topology at runtime to accommodate new requirements, e.g., [2]–[4]. Most of such architectures leverage Optical Circuit Switches (OCS) to augment the electrical packet switched network and add direct connections between pairs of racks. Such approaches have the potential to increase the flexibility of the network, i.e., its capability to satisfy new demands, but might also counter-intuitively complicate network design and management by inducing additional configuration possibilities.

Many of the dynamic DC network topology designs claim to provide more flexible networks. However, there is no common notion established in this area of how to measure the flexibility. Two possible metrics for flexibility in DC networks could be the bisection bandwidth of the network or throughput proportionality [5]. Bisection bandwidth measures the worst-case bandwidth between any two equal-size partitions in the network [6]. Throughput proportionality says that there is a proportional relationship between the throughput per server and the fraction of servers that participate in the communication [5].

None of the two considers the case of adaptive topologies over time where reconfigurations lead to additional problems: As users expect high quality and seamless services, adaptations must be performed in a timely manner. Thus, highly flexible networks need resource managements with short decision and execution times. Furthermore, adaptions might lead to packet loss and re-transmissions of packets whose path changed with the new topology.

Both issues are captured by the Network Flexibility metric [7] which measures how many requests a networking system can fulfill under an adaptation time and cost constraint.

The remainder of this statement introduces Network Flexibility more thoroughly and gives a brief overview of existing work on dynamic DC topologies. Finally, open questions and the contributions of this thesis are sketched.

II. Network Flexibility

Kellerer et al. [7] argue that many works in networking research claim to improve the flexibility of communication networks but evaluate this only from a qualitative point of view. Accordingly, they propose a quantitative measure for network flexibility that can compare specific systems or network designs. To measure its flexibility a system is faced with a number of requests which it must accommodate within a certain amount of time (adaptation time) and cost (adaptation cost). The normalized number of fulfilled requests is the flexibility of the system. In contrast to bisection bandwidth and throughput proportionality, this approach explicitly addresses adaptive systems such as adaptive DC topologies and also covers the problem of timely adaptations.

III. Dynamic Data Center Topology Designs

Traditional DC topologies such as Fat-Trees [8], BCube [9] or VL2 [10] do often not meet the time-varying requirements of today’s applications or only at a very high cost. To overcome this, several approaches to adaptive DC network topologies have been presented. Many of them augment an existing packet-switched DC network and create a hybrid network with a packet-switched and a circuit-switched part. The latter one can establish one hop connections between racks or hosts that have high communication demand. An early example of such an approach is Helios [2] which uses OCS in the core layer to connect the different pods in addition to a fixed packet-switched network part. The algorithm collects rate matrices of flows in the network and uses the Edmonds algorithm to determine the pod-to-pod connectivity. xWeaver [11] and DeepConf [4] use this basic architecture but employ

Supervised Learning and Reinforcement Learning methodologies to solve the NP-hard problem of determining the configuration of the OCS. The issues due to reconfigurations are not addressed in these works.

ProjecToR [3] tackles the scalability problem (centralized network control and high fan-out of OCS required) of this hybrid network approach with free-space optics that are mounted on top of the racks and connect the racks via mirrors at the ceiling and fully replace the electrical network. Its control logic matches the racks in a distributed way to transmit bursts of packets. FlatTree [12] places multiple circuit switches at different locations in the network to (locally) convert the topology between a traditional one, e.g., Fat-Tree and a random graph. Also Larry [13] implements the reconfigurability of the topology on a more local scope by connecting only certain groups of racks with an OCS. RotorNet [14] does not set circuits based on demand but periodically rotates through all possible configurations of the OCS and mixes in Valiant Load Balancing to improve the performance of the network.

Besides these adaptive topologies, there are also new proposals for static topologies based on random [6] or expander graphs [15] which criticize the adaptive approaches [5].

This brief overview illustrates the variety of adaptive topology designs which all show a performance increase in terms of throughput or flow completion time. However, it remains open, which solution performs best and maximizes network flexibility.

IV. Open Questions and Contributions

Most of the work described in Section III does not explicitly address the inherent drawbacks of reconfiguring network topologies, namely service interruption and potential retransmissions. Furthermore, the impact of the topology on quantitative network flexibility is unknown. This thesis tackles these gaps with three steps:

A. Flexibility Evaluation of Data Center Topologies

In the first step of our research agenda, we start by quantifying network flexibility and try to gain insights on the impact of the network‘s structure, e.g., different DC network topologies, on its provided flexibility. The performance comparison of different adaptive DC topology designs shall also answer the question under what circumstances certain designs are more favorable than others and also analyze the impact of reconfigurations. The comparison will build on flow-level and packet-level simulations. Therefore, an extension of the NS3 packet-level simulator to support OCS will be implemented.

B. Optimizing Network Topologies for Flexibility

In the second step, while considering these insights, we want to develop algorithms that design and adapt networks to maximize their flexibility. Depending on the outcomes of the previous step, these algorithms will account for the cost of the adaptation process and explicitly optimize for network flexibility. To maintain short adaptation times, we envision to enhance our algorithms based on machine learning and data-driven approaches.

C. Joint Optimization of Workload Allocation and Topology

Apart from the network‘s structure, we also want to look deeper into rising challenges when putting flexible network into effect, e.g., we want to focus on the management of physical resources in virtualized environments. This means to also consider the allocation of computing workload, e.g., in terms of virtual machines, and to aim at a joint optimization of workload allocation and topology adaptation. For instance, we envision a system that predicts future workload, pro-actively adapts the topology towards this workload and allocates workload while being aware of future adaptations.

Acknowledgment

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 647158 - FlexNets). The authors alone are responsible for the content of the paper.

References

[1] N. Shalom. Amazon found every 100ms of latency cost them 1% in sales. [Online]. Available: https://blog.gigaspaces.com/amazon-found-every-100ms-of-latency-cost-them-1-in-sales/

[2] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, “Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers,” *ACM SIGCOMM CCR*, vol. 41, no. 4, pp. 339–350, 2011.

[3] M. Ghobadi, D. Kilper, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, and M. Glick, “ProjecToR: Agile Reconfigurable Data Center Interconnect,” in *Proc. ACM SIGCOMM 2016*, pp. 216–229.

[4] S. Salman, C. Streiffer, H. Chen, T. Benson, and A. Kadav, “DeepConf: Automating Data Center Network Topologies Management with Machine Learning,” in *Proc. ACM Workshop on Network Meets AI & ML 2018*, pp. 8–14.

[5] S. Kassing, A. Valadarsky, G. Shahaf, M. Schapira, and A. Singla, “Beyond fat-trees without antennae, mirrors, and disco-balls,” in *Proc. ACM SIGCOMM 2017*, pp. 281–294.

[6] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, “Jellyfish: Networking Data Centers Randomly,” in *NSDI 2012*, pp. 1–14.

[7] W. Kellerer, A. Basta, P. Babarczi, A. Blenk, M. He, M. Klugel, and A. M. Alba, “How to measure network flexibility? - A proposal for evaluating softwarized networks,” p. 7.

[8] M. Al-Fares, A. Loukissas, and A. Vahdat, “A scalable, commodity data center network architecture,” *ACM SIGCOMM CCR*, vol. 38, pp. 63–74.

[9] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, S. Lu, and G. Lv, “BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers,” in *Proc. ACM SIGCOMM 2009*, pp. 63–74.

[10] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, “VL2: A Scalable and Flexible Data Center Network,” in *Proc. ACM SIGCOMM 2009*, pp. 51–62.

[11] M. Wang, Y. Cui, S. Xiao, X. Wang, D. Yang, K. Chen, and J. Zhu, “Neural Network Meets DCN: Traffic-driven Topology Adaptation with Deep Learning,” in *Proc. ACM SIGMETRICS 2018*, pp. 1–25.

[12] Y. Xia, X. S. Sun, S. Dzinamarira, D. Wu, X. S. Huang, and T. S. E. Ng, “A Tale of Two Topologies: Exploring Convertible Data Center Network Architectures with Flat-tree,” in *Proc. ACM SIGCOMM 2017*, pp. 295–308.

[13] A. Chatzieleftheriou, S. Legtchenko, H. Williams, and A. Rowstron, “Larry: Practical Network Reconfigurability in the Data Center,” in *Proc. NSDI 2018*, pp. 141–156.

[14] 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 *Proc. ACM SIGCOMM 2017*, pp. 267–280.

[15] A. Valadarsky, G. Shahaf, M. Dinitz, and M. Schapira, “Xpander: Towards Optimal-Performance Datacenters,” in *Proc. ACM CoNEXT 2016*, pp. 205–219.