Machine learning for dynamic network resource allocation

Frédéric Giroire

frederic.giroire@cnrs.fr

CNRS/Université Côte d'Azur

2019

With the latest wave of Cloud and IoT adoption, a sweeping technological change has been affecting our daily uses and opening up new opportunities for people and businesses. According to Cisco [9], it is estimated that there will be 28.5 billion connected devices by 2022, up from 18 billion in 2017. Furthermore, the appearance of these new heterogeneous devices is leading to a wide range of applications (e.g., wearable activity monitors, autonomous cars, industrial robotics) that can be built to enable smart healthcare, intelligent transportation and smart logistics. However, these kinds of applications are typically latency-sensitive and require intensive computation resources as well as high energy consumption for processing. For instance, interactive applications require ultra-low network latencies (e.g., end-to-end delays under 20 ms for augmented reality) whereas latencies to the closest Data Center are 20-30 ms using wired networks and up to 50-150 ms on 4G mobile networks. Also, throughput-oriented applications require local computations (e.g., distributed real-time video surveillance is relevant only close to the cameras). To mitigate these shortcomings, Edge Computing [7,8] has been proposed as the cloud close to the ground in order to leverage distributed and near-edge computation by putting intelligence closer to the data source. As an intermediate overlay, this new paradigm provides several benefits from location awareness, mobility support to geo-distribution and allow preliminary local computation before involving the central-cloud.

In addition, several emerging approaches promise to simplify the management of the network and increase network capability, primarily Software-Defined Network (SDN) and Network Function Virtualization (NFV). By enabling scaled on demand instantiation of Virtual Network Functions (VNFs), they facilitate cost efficiency of the network and deliver more agile programmable networks. As a solution to ossification of the Internet, these trends are expected to continue unabated and play an important role in next-generation networks.

Since Edge computing enables its applications to benefit from SDN and NFV technologies, it has inherited the challenges of resource provisioning and service placement. The latter have been studied using various optimization methods such as approximation algorithms [1,4], and column generation [2] with theoretically proven performance. Still, due to the high dynamicity of the demands and resources within the IoT environment, traditional algorithms are limited by new arising network reconfiguration challenges. To this end, recent trends in networking are proposing the use of machine learning approaches [5,6] for the control and operation of networks.

The goal of this thesis is to address these challenges through the following steps:

(i) Propose good machine learning-based models of traffic that adapt to the availability and dynamicity of the services after deciding where and on which level to perform resource computation and data storage.

(ii) Find efficient optimization algorithms for dynamic routing, resource allocation and VNFs placement based on predefined metrics (e.g., delay, energy consumption, resource utilization).

(iii) Deploy and test the proposed solutions on real topologies using SDN and NFV technologies for network control.

References:

[1] A. Tomassilli, F. Giroire, N. Huin, and S. Perennes, "Provably Efficient Algorithms for Placement of Service Function Chains with Ordering Constraints", in Proceedings of IEEE Infocom, 2018.

[2] N. Huin, B. Jaumard, F. Giroire, "Optimization of Network Service Chain Provisioning", In IEEE/ACM Transactions on Networking (ToN), vol. 26, nb. 3, pp. 1320-1333, 2018.

[3] F. Giroire, N. Huin, A. Tomassilli, S. Perennes, "When Network Matters: Data Center Scheduling with Network Tasks", in Proceedings of IEEE Infocom, 2019.

[4] R. Cohen, L. Lewin-Eytan, J. S. Naor, and D. Raz, "Near optimal placement of virtual network functions", in Proceedings of IEEE Infocom, 2015.

[5] T. Ouyang, R. Li, X. Chen, Z. Zhou, X. Tang, "Adaptive User-managed Service Placement for Mobile Edge Computing: An Online Learning Approach", in Proceedings of IEEE Infocom, 2019.

[6] S. Wang, T. Tuor, T. Salonidis, KK. Leung, C. Makaya, T. He, K. Chan, "When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning", in Proceedings of IEEE Infocom, 2018.

[7] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, 2016.

[8] M. Satyanarayanan, ''The emergence of edge computing,'' Computer, vol. 50, no. 1, pp. 30–39, Jan. 2017.

[9] Cisco visual networking index: Forecast and methodology, 2017–2022.