2021 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN) | 978-1-6654-3983-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/NFV-SDN53031.2021.9665118

Application- and reliability-aware service function chaining to support low-latency applications in an NFV-enabled network

Mohammad Mohammadi Erbati
(PhD Candidate)
Fixed Line Device Engineering
Deutsche Telekom Technik GmbH
Darmstadt, Germany
Mohammad.mohammadi.erbati@telekom.de

Gregor Schiele
(PhD Supervisor)
Embedded Systems Group
University of Duisburg-Essen
Duisburg, Germany
Gregor.schiele@uni-due.de

***Abstract*— Future network architectures supporting 5G, 6G, Internet of Things (IoT), and fixed line networks face diverse service requirements in various scenarios, particularly by growing Ultra Low-Latency and Reliable Applications (ULLRA) such as: autonomous driving, remote surgery, augmented reality, virtual reality, IoT applications, etc. Network Function Virtualization (NFV) and Software Defined Networking (SDN) are two promising complementary solutions to meet diverse service requirements. In this study, we propose a novel application- and reliability-aware SFC embedding algorithm aiming to (a) decrease the latency and (b) increase the reliability of ULLRA. In this regard, we prioritize ULLRA as high priority SFC requests and reserve an amount of physical resources (bandwidth, memory, CPU) for embedding only high priority SFC requests to improve their latency and reliability. We formulate our proposed algorithm in the form of an Integer Linear Programming (ILP) optimization model to obtain optimal results. We also offer a heuristic algorithm to obtain near optimal results to reduce the execution time and make it usable for real-world networks. We consider constraints on maximum tolerable end-to-end delay, physical resource utilization (bandwidth, memory, CPU), and reliability. We examine our proposed SFC embedding algorithm in different test scenarios such as changing the number and the length of SFC requests, the proportion of reserved physical resources for high priority SFC requests, and the proportion of low-latency traffic flows to the total traffic flows. The results show that our proposed algorithms effectively improve the latency and reliability of ULLRA with minimal side effects on other applications compared to state of the art algorithms.**

Keywords*— *service function chaining (SFC), network function virtualization (NFV), software defined network (SDN), reliability, quality of service (QoS), virtual network function (VNF).

I. INTRODUCTION

In traditional computer networks, network functions such as Network Address Translation (NAT), Traffic Monitor (TM), and Firewall (FW) are deployed on dedicated hardware named *middleboxes*. This leads to flexibility, scalability, manageability, and agility challenges. One solution to overcome the limitations of traditional networks is to adopt Network Function Virtualization (NFV) and Software-Defined Networking (SDN). Network functions are realized as software implementation and placed on general purpose virtual machines (VMs), containers, and commercial off-the-shelf equipment. Network services are provided in the form of *Service Function Chains* (SFCs). Each SFC is a sequence of *Virtual Network Functions* (VNFs) in a predefined order. This promises to provide elastic network functions, which offers many advantages such as improving network flexibility, scalability, manageability, QoS, innovation time to market, and decreasing capital expenses and operational expenses. With the development of NFV and SDN, Internet Service Providers (ISPs) are increasingly placing VNFs at the network edge [1] [2] [3].

Latency and reliability are gaining increasing importance for growing *Ultra Low-Latency and Reliable Applications* (ULLRA) in 5G, 6G, IoT, and fixed line networks (see Fig. 1), such as autonomous driving, remote surgery, augmented reality , virtual reality, IoT applications, etc. These applications require extremely high sensitivity, accuracy, and reliability. Many studies have been done to enhance the latency and reliability, but they still need more improvements in NFV environments. Our goal in this study is to decrease the latency and increase the reliability of ULLRA. These two requirements often impact each other. Increasing the reliability may lead to an increase of latency, further complicating the problem. Our approach to decrease the latency and increase the reliability is using SFC request prioritization. In this study, we use prioritization in SFC embedding problem. We mark ULLRA as high priority SFC requests and other applications as low priority SFC requests (see Fig. 2). Then, we reserve an amount of physical resources (bandwidth, memory, CPU) for embedding only high priority SFC requests to obtain more optimal deployment paths, and consequently lower latency and higher reliability can be achieved for ULLRA. We also study the impact of changing the amount of physical resource reservation for high priority SFC requests. The main contributions of this study are as follow:

- We offer a novel and efficient SFC embedding algorithm with respect to the priority of each SFC request and possible QoS constraints.

978-1-6654-3983-1/21/$31.00 ©2021 IEEE

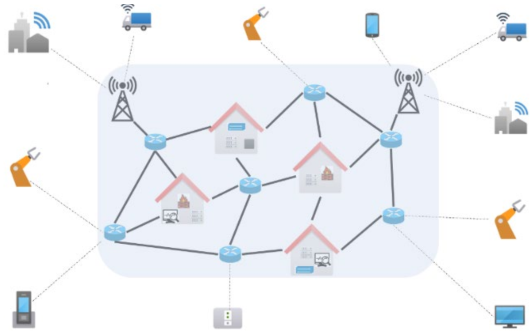

Fig. 1. An example of network architecture

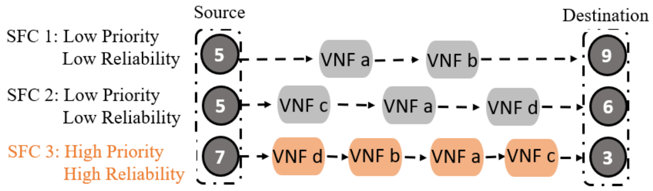

Fig. 2. An example of an SFC request with priority

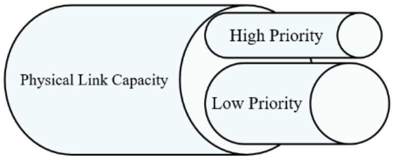

Fig. 3. An example of bandwidth reservation

- We formulate the SFC embedding problem as an ILP optimization model aiming to minimize the latency and improve the reliability of ULLRA with minimal side effects on other applications by taking maximum tolerable end-to-end delay and reliability of low priority SFC requests into account.
- We offer a heuristic algorithm with respect to the priority of each SFC request to obtain near optimal results to make it usable for real-world networks.
- We conduct a detailed analysis of our proposed algorithms and present the enhancements in terms of end-to-end delay, reliability, bandwidth utilization, and path length.

The rest of the paper is organized as follows. In Section II, we discuss related work. Section III details the problem statement. In Section IV we present the current status of our study. Section V introduces remaining challenges. Finally, we conclude this paper in Section VI.

II. Related work

NFV is a promising technology to revolutionize the telecommunication industry. In the following we look at some of the recent studies in this field. We divide the related works into two subsections. First, we look at recent studies, which mostly address the routing optimizations, and then we summarize recent studies, which focus on the reliability issues.

A. Routing Optimization

G. Sun et al. [4] did a study on low-latency and resource efficient SFC orchestration in NFV environment. They proposed an SFC embedding algorithm based on a breadth-first search (BFS) to optimize the SFC deployment path. They optimized the end-to-end delay and bandwidth resource consumption. They also studied [5] cost efficient SFC orchestration for low-latency applications. They proposed Closed-Loop Feedback (CLF) algorithm, which is a heuristic algorithm to find the shortest path to embed an SFC request. In [6], they did a study on energy-efficient provisioning for SFCs to support delay-sensitive applications and proposed an energy-efficient routing algorithm for dynamic SFC deployment.

M. M. Tajiki et al. [7] proposed an ILP optimization model, which focuses on joint energy efficient and QoS-aware service function chaining, while considering constraints on maximum tolerable end-to-end delay and physical resource consumptions. J. Pei et al. [8] proposed a Binary Integer Programming (BIP) optimization model to address SFC embedding problem with dynamic VNF placement in geo-distributed cloud systems. Their proposed approach provides higher performance in terms of SFC request acceptance rate, network throughput and mean VNF utilization rate. Y. Li et al. [9] did a study on cost- and QoS-based NFV service function chain mapping mechanism. They take resource constraints, end-to-end delay requirement, and reliability requirement into account in their optimization model.

B. Reliability Oriented

P. K. Thiruvasagam et al. [10] investigated on SFC embedding problem with respect to the reliability, delay, and resource efficient in softwarized 5G networks. If reliability requirement is not met after their subchaining method, they add backup to VNFs to meet the reliability requirement. They formulated the SFC placement problem as an ILP problem. They also studied [11] reliable placement of SFCs and virtual monitoring functions with minimal const in softwarized 5G. networks. They enhanced the reliability of SFCs from VNF failures by exploring latency-aware and reliable SFC placement to meet the requirement of users. They formulated their optimization model as an ILP problem. Y. Wang et al. [12] did a study on reliability-oriented and resource-oriented SFC construction and backup. They considered SFC construction phase. In the construction phase of SFCs, if the deployed SFC cannot meet the reliability requirements, a backup VNF instance will improve the reliability of the SFC. Their proposed backup method reduces the bandwidth consumption and backup resources. An ILP mapping algorithm is used in their study. L. Qu et al. [13] investigated on reliability-aware VNF chain placement and flow routing optimization. They determined the number of required VNF backups to ensure the required reliability. Their proposed resource-sharing-based VNF placement scheme and formulated it as an ILP optimization model. J. Zhou et al. [14] studied network function parallelism architecture to improve the reliability and reduce the delay. They proposed a learning based SFC deployment strategy to improve the service reliability while reducing the end-to-end delay. They improved the service reliability by deploying back-up VNFs.

Different from the above mentioned studies, we propose a novel SFC embedding algorithm with respect to the priority of each SFC requests aiming to minimize the latency and increase the reliability of ULLRA without a backup method at the first phase. Then in the next phase if necessary, adding an efficient backup method to fulfill the requirements. We formulate our optimization model as an ILP optimization model. We also provide a heuristic algorithm to achieve near optimal solutions to reduce the execution time for large networks.

III. PROBLEM STATEMENT

The ultra low-latency and reliable applications in 5G, 6G, IoT, and fixed line networks require not only extremely low latency, but also high reliability. Network providers should have an optimal plan for embedding SFC requests. Each SFC request needs to traverse several VNF instances in a predefined order and this routing length influences the latency, reliability, and physical resources consumption. In addition, the physical network resources and VNF instances are limited in ISP's network. Achieving higher reliability may lead to higher latency and place redundant network elements.

Our goal in the first place in this study is to improve the latency and the reliability of ULLRA in the SFC embedding phase with minimal side effects on other applications and without adding any redundant elements. In the next step we plan to add an efficient backup method to increase the reliability further with redundancy and minimal side effects on latency and physical resource consumption. In this regard, we consider constraints on bandwidth, memory, CPU, reliability, and maximum tolerable end-to-end delay.

IV. SYSTEM MODEL AND CURRENT STATUS

In this study, we represent the physical network as our substrate network with a matrix. The Gridnet network topology [15] as depicted in Fig. 4 is used as our substrate network, which consists of 8 nodes and 18 links. We consider two types of nodes in our simulation. One type is Core-Data-Center (CDC) nodes, which can host the VNF types. The other type is switching nodes, which just forward traffic to the next nodes. In Fig. 4, the CDC nodes are represented with green squares and switching nodes are represented with red dots. We generate different numbers of SFC requests with random source and destination nodes. Each SFC request requires a random amount of physical resources in a predefined range. In our optimization model we prioritize ULLRA as high priority SFC requests and other applications as low priority SFC requests. Then, we reserve an amount of physical resources (bandwidth, memory, CPU) for embedding only high priority SFC requests (see Fig. 3). Therefore, high priority SFC requests obtain more optimal deployment paths by using the reserved physical resources, and they can obtain lower latency and higher reliability without adding redundant elements. To minimize side effects on other applications, we consider the maximum tolerable end-to-end delay and reliability requirements of other applications (low priority SFC requests) in our SFC embedding algorithm. In this study, we also study the impact of changing the amount of reserved physical resources for high priority SFC requests on latency and reliability.

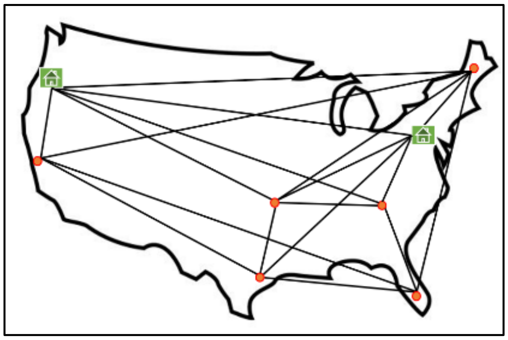

Fig. 4. Gridnet network topology [15]

As stated before, we formulate our optimization problem in the form of an ILP optimization model. We name our proposed SFC embedding algorithm *Optimal Application-Aware and Reliable SFC* (OARS) embedding algorithm and examine it with Gridnet network topology as our substrate network. We define a priority coefficient factor presented as (μ) to reserve an amount of physical resources (bandwidth, memory, CPU) for path selection of high priority SFC requests. We define constraints on physical resource consumption, reliability, maximum tolerable end-to-end delay, flow control, and VNFs order control. We generate different numbers of SFC requests and create different test scenarios to evaluate the OARS embedding algorithm by changing the number and the length of SFC requests, the proportion of low-latency traffic flows to the total traffic flows, and the proportion of reserved physical resources for high priority SFC requests. We evaluate our OARS embedding algorithm with respect to latency, reliability, bandwidth consumption, and path length.

We compare our algorithm to the Nearest Service Function First (NSF) algorithm used in [7]. The initial results show that OARS embedding algorithm effectively improves the latency and reliability of ULLRA by applying the SFC prioritization without any major side effects on other applications (low priority SFC requests). The initial optimal results show around 17 percent and 14 percent improvement in terms of latency and bandwidth consumption of ULLRA without any negative side effects on other applications. We use the same simulation parameter settings when comparing the algorithms for fairness. We implemented the simulation in Python using PuLP library and CBC MILP Solver.

At this stage of our study, we are optimizing our heuristic algorithm named *Near Optimal Application-Aware and Reliable SFC* (NOARS) embedding algorithm to reduce the execution time and obtain near optimal results in an acceptable time frame to make it usable for real-world networks. In the next step, we are going to add an efficient backup method to increase the reliability further with a minimal negative impact on latency and physical resources consumption.

V. CHALLENGES

There are some challenges that we need to overcome in our study, which we are working on. First, solving the SFC embedding problem is an NP-hard problem as proven in [7]. As stated before, since obtaining the exact numerical solution via OARS for large networks is very complex and time consuming, one challenge is to optimize NOARS regarding its execution time and optimality-gap to make it usable for real-world networks.

The next challenge is to improve the reliability further. So far, we have improved the reliability of ULLRA by prioritization without using a backup method and redundancy. Adding a backup method and redundancy may lead to higher latency and physical resource consumption. How to minimize this impact is another challenge that we need to take care of.

VI. Conclusion

In this study, we studied application- and reliability-aware service function chaining to support ultra low latency and reliable applications in an NFV-enabled network. We investigated on a novel and efficient SFC embedding algorithm aiming to minimize the latency and improve the reliability of ULLRA with minimal side effects on other network services, first without a backup method and then with a backup method and redundancy. We formulate our optimization model named OARS in the form of an ILP optimization model to obtain the optimal solutions, while taking constraints on maximum tolerable end-to-end delay, physical resources consumption (bandwidth, memory, CPU), and reliability into account. We also study a heuristic algorithm named NOARS with respect to the priority of each SFC requests to obtain near optimal solutions with a minimum execution time and optimality gap to make it usable for real-world networks. In the next step, we will add an efficient backup method to improve the reliability further with minimal side effects on latency and resource consumption. As future work, we plan to study the proposed algorithms in an online allocation scenario with flow arrivals and departures over time.

References

[1] J. de Jesus Gil Herrera and J. Felipe Botero Vega, "Network Functions Virtualization: A Survey," *IEEE Latin America Transactions,* vol. 14, no. 2, pp. 983-997, 2016.

[2] K. Kaur, V. Mangat and K. Kumar, "A comprehensive survey of service function chain provisioning approaches in SDN and NFV architecture," *Computer Science Review,* vol. 38, 2020.

[3] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Communications Surveys & Tutorials,* vol. 18, no. 1, pp. 236-262, 2016.

[4] G. Sun, Z. Xu, H. Yu, X. Chen, V. Chang and A. V. Vasilakos, "Low-Latency and Resource-Efficient Service Function Chaining Orchestration in Network Function Virtualization," *IEEE Internet of Things Journal,* vol. 7, no. 7, 2020.

[5] G. Sun, G. Zhu, D. Liao, H. Yu, X. Du and M. Guizani, "Cost-Efficient Service Function Chain Orchestration for Low-Latency Applications in NFV Network," *IEEE Systems Journal,* vol. 13, 2019.

[6] G. Sun, R. Zhou, J. Sun, H. Yu and A. V. Vasilakos, "Energy-Efficient Provisioning for Service Function Chains to support Delay-Sensitive Applications in Network Function Virtualization," *IEEE Internet of Things Journal,* vol. 7, 2020.

[7] M. M. Tajiki, S. Salsano, L. Chiaraviglio, M. Shojafar and B. Akbari, "Joint Energy Efficient and QoS-Aware Path Allocation and VNF Placement for Service Function Chaining.," *IEEE Transactions on Netwrok and Service Management,* vol. 16, no. 1, 2019.

[8] J. Pei, P. Hong, K. Xue and D. Li, "Efficiently Embedding Service Function Chains with Dynamic Virtual Network Function Placement in Geo-Distributed Cloud System.," *IEEE Transactions on Parallel and Distributed Systems,* vol. 30, no. 10, 2019.

[9] Y. Li, L. Gao, S. Xu, Q. Qu, X. Yuan, F. Qi, S. Guo and X. Qiu, "Cost- and QoS-based NFV Service Function Chain Mapping Mechanism," *IEEE/IFIP Network Operations and Management Symposium,* 2020.

[10] P. K. Thiruvasagam, V. J. Kotagi and C. S. R. Murthy, "A Reliability-Aware, Delay Guaranteed, and Resource Efficient Placement of Service Function Chains in Softwarized 5G Networks.," *IEEE Transactions on Cloud Computing,* 2020.

[11] P. Kaliyammal Thiruvasagam, A. Chakraborty , A. Mathew and C. S. Ram Murthy, "Reliabalie Placement of Service Function Chains and Virtual Monitoring Functons With Minimal Cost in Softwarized 5G Networks.," *IEEE Transactions on Network and Service Management,* vol. 18, no. 2, 2021.

[12] Y. Wang, L. Zhang , P. Yu, K. Chen, X. Qiu, L. Meng, M. Kadoch and M. Cheriet, "Reliability-oriented and Resource-efficient Service Function Chain Constrution and Backup," *IEEE Transactions on Network and Service Management,* 2020.

[13] L. Qu, M. Khabbaz and C. Assi, "Reliability-Aware Service Chaining In Carrier-Grade Softwarized Networks," *IEEE Journal on Selected Areas in Communications,* vol. 36, no. 3, 2018.

[14] J. Zhou, G. Feng and Y. Gao , "Network Function Parallelization for High Reliability and Low Latency Services," *IEEE Access,* 2020.

[15] "The Internet Topology Zoo," The University of Adelaide, [Online]. Available: http://www.topology-zoo.org/. [Accessed 03 2021].