Edge to Cloud Network Function Offloading in the ADAPTO Framework

Alessio Botta[1], Roberto Canonico[1], Annalisa Navarro[1], Giovanni Stanco[1], Giorgio Ventre[1], Antonio Buonocunto[2], Antonio Fresa[2], Enzo Gentile[2] Leonardo Scommegna[3], and Enrico Vicario[3]

[1] Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Naples, Italy
{a.botta, roberto.canonico,annalisa.navarro, giovanni.stanco,giorgio.ventre}@unina.it
[2] Ericsson Telecomunicazioni S.p.A., Pagani, Italy
{antonio.buonocunto,antonio.fresa,enzo.gentile}@ericsson.com
[3] Department of Information Engineering, University of Florence, Florence, Italy
{leonardo.scommegna,enrico.vicario}@unifi.it

Abstract. As telcos increasingly adopt cloud-native solutions, classic resource management problems within cloud environments have surfaced. While considerable attention has been directed toward the conventional challenges of dynamically scaling resources to adapt to variable workloads, the 5G promises of Ultra-Reliable Low Latency Communication (URLLC) remain far from being realized. To address this challenge, the current trend leans toward relocating network functions closer to the edge, following the paradigm of Mobile Edge Computing (MEC), or exploring hybrid approaches. The adoption of a hybrid cloud architecture emerges as a solution to alleviate the problem of the lack of resources at the edge by offloading network functions and workload from the Edge Cloud (EC) to the Central Cloud (CC) when edge resources reach their capacity limits. This paper focuses on the dynamic task offloading of network functions from ECs to CCs within cloud architectures in the ADAPTO framework.

Keywords: Resource management · Virtual Network Function · Edge-to-cloud offloading.

1 Introduction

Telco Networks have always been subject to a continuous architectural evolution to satisfy the demanding requirements associated with each network generation. One of the most important evolutions introduced with the 5G networks is the massive adoption of the cloud-native principles that provide a flexible way to design and orchestrate Network Functions. A cloud native network function is designed to better sustain all mutable network characteristics thanks to a finer granularity achieved by decomposing a monolithic SW component in several microservices. The advantages gained in breaking a Network Function (NF) come

with the risk of increasing resource utilization if the NF is operated without a clear and automated optimization strategy to keep the application dimension in line with the network demands.

The advent of 5G Mobile Edge Computing (MEC) represented a significant stride towards materializing the promise of Ultra-Reliable Low Latency Communication (URLLC) within the domain of 5G networks. MEC extends the cloud-computing capabilities near mobile users, providing computational and data processing capabilities and enabling time-sensitive services such as driverless vehicles, augmented reality, robotics, and immersive media [1]. However, the scarcity of resources in the Edge Clouds (ECs) limits the delivery of URLLC, Enhanced Mobile Broadband (eMBB), and Massive Machine Type Communication (mMTC) services. Consequently, hybrid methodologies have emerged, proposing the integration of Edge and the Central Cloud (CC) to provide the proximity advantages offered by edge computing and the extensive computational resources housed within the central cloud, catering to the multifaceted and stringent demands of URLLC, eMBB, and mMTC use cases in 5G networks.

In this context and with these objectives, the ADAPTO project was established. ADAPTO is part of the larger RESTART initiative, funded by the Italian government in the context of the *National Recovery and Resilience Plan* (NRRP) as part of the Next Generation EU (NGEU) programme. Leveraging recent advancements in Software-Defined Networking (SDN) and Network Function Virtualization (NFV), ADAPTO seeks to craft a framework to orchestrate the utilization of computational and network resources across distributed cloud infrastructures to adapt to variable workloads and network conditions. The overarching goals of the ADAPTO framework are threefold: first, the autonomous scaling and orchestration of Virtual Network Functions (VNFs) based on the operational load; second, the intelligent offloading of the VNFs between edges and the central cloud, based on 5G backhaul conditions and edge resource status; and finally, the selection of the most suitable 5G backhaul between the edge and the central cloud to ensure a seamless handoff and guarantee URLLC requirements.

This paper specifically focuses on a critical aspect within the framework - i.e. the seamless handoff mechanism - presenting the examination of the role of an intelligent agent employing Reinforcement Learning (RL) positioned at the network's edge. This agent is tasked with decisions on the offloading of the workload between EC and CC, as well as the optimal selection of the most suitable 5G backhaul for data and signaling transfer between EC and CC. Through continuous assimilation of real-time data encompassing network conditions, edge resource utilization, and workload variations, this RL-based agent acquires the capacity to make informed decisions.

The rest of the paper is organized as follows: the technological background of the application context investigated by the ADAPTO project is provided in Section 2. In Section 3, we describe in detail how the ADAPTO project can help telco operators fulfill 5G promises, describing the design goals and the architectural view (Section 3). Finally, we summarize our paper and draw conclusions in Section 4.

2 Technological context

In this section, we briefly highlight the technological context in which the ADAPTO project provides its research contributions.

2.1 5G architecture

The 5G system, standardized by 3GPP [2], comprises three primary components: User Equipments (UEs) comprising mobile devices such as smartphones or tablets, the Radio Access Network (RAN), that connects UEs to the fixed infrastructure through wireless connections, and the Core Network (CN), that performs essential functionalities such as authentication, session management, and data routing.

The 5G Core Network (depicted in Figure 1) consists of various 5G Core network functions engineered as cloud-native SW components that, with their modular architecture, can better adapt to the high and various demands of an evolved and dynamic telco network. The Access and Mobility Function (AMF) handles communication between the UEs and the RAN, controlling access, registration, and mobility. The Session Management Function (SMF) manages the User Plane Functions (UPFs) for routing decisions and Quality of Service (QoS) settings. The Policy and Charging Function (PCF) is used for enforcing subscriber policies and charging the users. A centralized state is kept in the Unified Data Management (UDM), which synchronizes the states in the different network functions. All of the 5G Core components have been designed according to the Service-based Architecture (SBA) framework, implementing an HTTP/2-based Service Bus Interface (SBI). The main characteristic of this architecture is a decomposition of function in multiple independent entities (microservices) that interact via Application programming Interface (API) to deliver network functionalities.

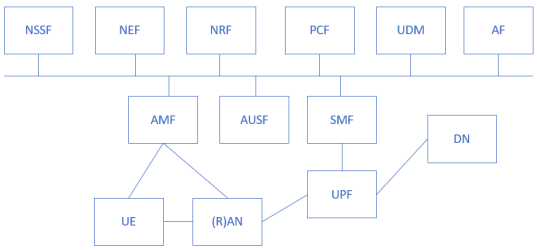

Fig. 1. 5G network functions

2.2 Network Function Virtualization Scaling and Placement

Network Function Virtualization (NFV) technology stands as a cornerstone in the evolution of 5G and beyond networks by detaching the traditional network functions from specialized hardware and transforming them into versatile, cross-platform, Virtualized Network Functions (VNFs).

VNF Scaling NFV enables flexible Network Function (NF) deployments, in which they can scale dynamically according to the rapid fluctuations in mobile data traffic demands and can be placed at different locations, e.g. at the edge or at a central position [3]. Two options are available for VNF scaling: Vertical and Horizontal. *Vertical scaling* involves resizing existing VNFs by adjusting computational, memory, or storage resources as needed. Conversely, *Horizontal scaling* creates additional instances of the same VNF or terminates redundant instances. While horizontal scaling enhances scalability and service reliability, it intensifies resource consumption and poses challenges related to state migration. Vertical scaling optimizes resource utilization but falls short in scalability and lacks flexibility in changing VNF hosts, impacting practical implementation.

VNF Placement The promises of 5G to provide low-latency and ultra-reliable connectivity have led to the need for reconsidering the fully centralized architecture typical of 4G networks, where all core functions were confined to a single central location connected to the RAN through high-capacity backhaul.
Reenvisioning cloud computing, edge computing moves resources from centralized data centers to the network's edge, bringing them closer to users and application-produced data. Edge computing stands out as one of the vital elements essential for meeting the rigorous Key Performance Indicators (KPIs) of 5G, driving the development of *5G Mobile Edge Computing (MEC)* [4]. MEC offers the flexibility to position both the essential 5G core VNFs and application VNFs at the mobile network's edge. Notably, certain 5G core VNFs, such as UPF, maintain close ties with application VNFs. Consequently, it becomes imperative to locate the UPF, and not only services, at MEC servers closer to users [3]. Strategically placing VNFs in both edge and core cloud servers enables the accommodation of different possible demands. Edge servers cater to applications requiring immediate responsiveness, while core servers accommodate VNFs needing more computational power or serving larger user bases. This dual placement ensures each service receives tailored computational capabilities and responsiveness [5].

Several possibilities have been so far explored for the so-called *Edge-Core split Option* [6]: the Local Offload Split, where only the UPF is at the edge, causing significant signaling through the backhaul; the Locally Administered Edge Split, involving the AMF and SMF at the edge, reducing backhaul signaling but potentially leading to handovers due to mobility or overload; and the Autonomous Edge Split, which operates autonomously at the edge, handling all

functions locally but limiting its scope to a local area and eliminating handover possibilities in case of overload or mobility. Two specific connectivity setups have been examined experimentally [7]: direct connectivity, where the UPF is positioned within the central core, and a local offloading scenario, in which the UPF is placed near the gNodeB. This analysis revealed that deploying a local UPF notably reduces latency. This effect was especially evident when utilizing a satellite backhaul between the edge and the central core.

2.3 SD-WAN for 5G backhauls

The 5G backhaul between the edges and the core network, as well as between the edges and a possible central cloud, is critical for meeting strict latency and throughput requirements. Backhaul options are usually based on fiber or millimeter-wave wireless connections. However, employing these technologies alone could not be feasible, especially in cases when there is the need to cover very distant areas. [8]. For this reason, the edge and the central core can be interconnected through a joint combination of backhaul networks based on different technologies, such as fibers, wireless, satellite, each one bringing its own benefits in terms of guaranteed throughput or latency. [6].

The advancements in the Software Defined Networking (SDN) paradigm have led to extending its applicability across time and scope, towards increasingly broader areas and with a finer temporal granularity. SD-WAN (Software Defined Wide Area Network) comes into play when catering to the real-time and adaptable demands of new 5G services flexibly and cost-effectively [9]. SD-WAN orchestrates traffic routing between multiple distant geographical sites. The distributed sites keep a connection with the Internet and with each other through one or more Edge Routers (ERs) leveraging a combination of different transport technologies (e.g., mobile, fiber). Each ER is linked to an SD-WAN controller that enforces network policies by sending rules to the ERs, exerting specific actions. These actions enable leveraging the multiple transport networks available in order to perform different kinds of optimization such through bandwidth aggregation, duplicating packets across different transports to ensure successful delivery or error correction of packets, or adaptively switching the used transport if the current one fails to guarantee desired performance or becomes unavailable. SD-WAN has also been successfully employed to enable communications between Service VNFs spread across the globe to form SFCs [10].

The application of SD-WAN to monitor and control traffic traversing the 5G backhauls involves placing providers' ERs at the edges and at the central site (e.g. between the access network and the core network [11]). The ER function in the network serves as an SD-WAN router, making decisions on available backhauls and directing data traffic based on specific characteristics of these backhauls.

3 The ADAPTO project

Virtualization of network functions, along with service function chaining, allows network infrastructure providers to achieve both agility and cost reductions by

replacing traditional dedicated hardware devices with flexible software modules running in virtualized components. To satisfy the expectations of operators, however, sophisticated algorithms are required to properly map virtualized network functions onto the physical infrastructure that is composed of distributed datacenters located both in the core and at the edge of the network. This requires management and control plane functions that are able to take optimal resource allocation decisions within a fair time constraint and that properly take into account the different end-to-end requirements of diverse classes of applications. The ADAPTO project aims to define a management framework for 5G and beyond networks that allows infrastructure providers to properly manage the computational resources required to accommodate the dynamic instantiation of network functions by combining QoS requirements with at-scale energy-saving objectives.

In a telco network, an optimization function requires a comprehensive view of all entities involved, from network-level metrics to the characteristics of individual microservices composing a network function. Once a proper data strategy is in place, it becomes possible to monitor the behavior of each functionality active in the network by also tracking the resources consumed to deliver it.

In the following, we give a brief overview of the design goals of the ADAPTO project and an overview of the edge agent's architecture.

- **VNFs autoscaling**: The ADAPTO framework dynamically adjusts the allocated resources for Virtual Network Functions (VNFs) based on operational load. The primary aim is to increase resources when the workload is high, avoid overburdening available resources, and reduce them during low load periods, optimizing the overall energy consumption. This requires monitoring of VNF resource utilization and the development of intelligent approaches seeking to adapt the allocated resource to the variable workload dynamically. The ADAPTO framework investigates the two possible VNF scaling possibilities mentioned earlier in this paper: vertical and horizontal.
- **VNFs placement** The placement and scaling of functions and applications in edge environments are constrained by limited computational resources. Consequently, apart from deploying these functions and applications at the edge, their deployment in a central cloud with an assumed pool of nearly limitless resources is also anticipated. An offloading agent is envisioned to detect when the edge resources are inadequate for accommodating the actual workload and utilize resources present in the central cloud.
- **Seameless Offloading** Another objective is to mask the geographical dispersion of resources, aiming to access central cloud resources as if they were local, ensuring low latency and consistently meeting the Quality of Service (QoS) requirements for various applications. This involves employing multiple backhaul connections between the edge and the central cloud and utilizing SD-WAN Edge Routers to select the backhaul that facilitates achieving the required low latency.

Figure 2 illustrates the reference architecture. The UEs establish connections to the edge cloud via the RAN. Within the edge deployment, the UPF manages

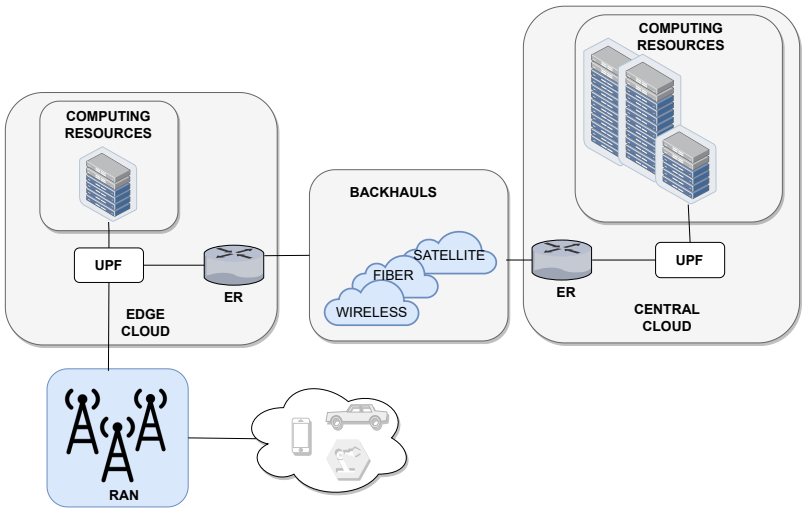

Fig. 2. Reference Architecture

traffic direction, either directing it to the edge cloud or the central cloud through the ER. Different ERs are employed - one situated at the edge and another at the central cloud - as in the SD-WAN paradigm. These ERs hold the responsibility of selecting the optimal backhaul among available options (such as Satellite, Fiber, or Wireless) to connect the edge and central cloud, aiming for the lowest latency possible. In Figure 3, the architecture of the ADAPTO agent devised for the Edge Cloud is shown. In the following, the main components are described:

- **NFV Controller**: The NFV controller is tasked with continuous monitoring of resources at the edge, specifically tracking the percentage usage of memory, CPU, and storage across various virtual network functions (VNFs). It also oversees the overall utilization of computing resources to assess available computation capacity.
- **NFV Scaling Decision Engine**: This component holds the responsibility of determining whether a particular instance of an NFV should scale in or out based on the potential overloading of edge resources. Additionally, it assesses and decides to offload services to the central cloud in case the edge resources become insufficient.
- **SD-WAN Controller**: The SD-WAN controller's primary function is to monitor latency by employing probing mechanisms on the available 5G backhauls. It configures the ERs to select the 5G backhaul that aligns with the Quality of Service (QoS) requirements.
- **Backhaul Selection Decision Engine**: Activated when the decision is made to offload services to the central cloud, this engine determines the backhaul to utilize for reaching the central cloud based on backhaul monitoring. It employs an intelligent Reinforcement Learning (RL) agent, which will be further discussed in the subsequent section.

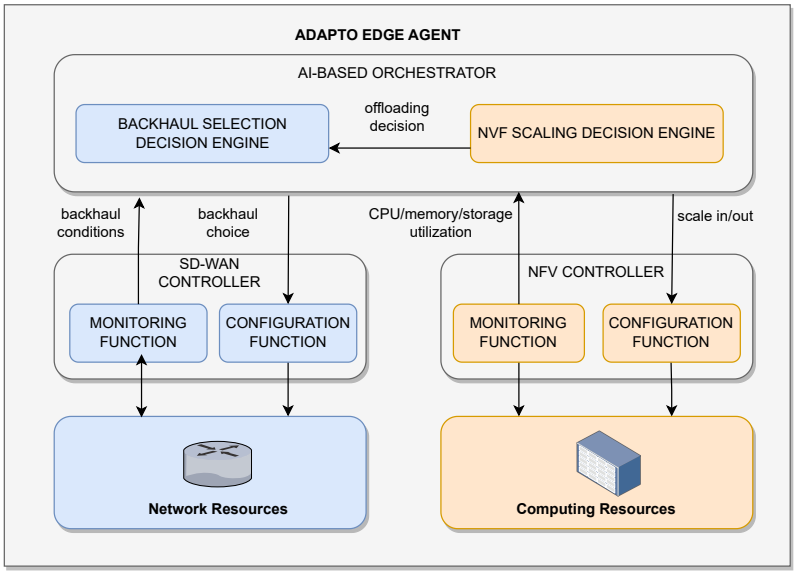

Fig. 3. Architecture of the ADAPTO agent deployed at the Edge.

3.1 Seamless Offloading based on Reinforcement Learning

We propose an Edge-Cloud offloading mechanism that considers the possible overload of resources at the edge and the near real-time edge-to-cloud latency. It is assumed that cloud resources are unlimited, yet the edge-to-cloud latency varies and is generally higher than the time required to complete a task at the edge, in case the edge resources are not overloaded [12]. The proposed mechanism entails real-time monitoring of edge resources (CPU or memory) and end-to-end latency between the cloud and the edge. Specifically, an agent situated at the edge collects network device status and computational resource data from edge servers (e.g., CPU, memory, and storage utilization). These collected data, together with the WAN latency, serve as inputs for intelligent algorithms to determine whether a service should be processed at the edge or in the cloud.

It is crucial that when offloading the task to the cloud, the WAN transport chosen ensures the QoS required by the application. Specifically, our focus is on reducing edge-to-cloud latency to guarantee a seamless handoff between edge services and services in the central cloud. The solution proposed is based on RL, employing an agent and an environment as depicted in Figure 4. The agent interacts with the environment, altering its state based on actions and receiving rewards or penalties for its behavior. This continuous trial-and-error process allows the agent to learn the best strategies for achieving objectives without prior knowledge of the network's details. The environment comprises the edge and central site that needs to communicate, the 5G backhauls, and the ERs. Reward functions are designed to reflect adherence or violation of network policies, where

positive rewards are given to the agent when the policy is met and negative rewards are given when the policy is violated. Policies enable establishing a maximum allowed latency for the 5G backhaul connection to provide certain QoS levels for 5G applications. The control action involves instructing ERs to select the 5G backhaul suitable for forwarding traffic between the edge and the cloud. Thanks to the reward, the agent can learn which backhaul to use based on the current backhaul conditions, providing, in this way, a method for enhancing the QoS for 5G applications and enabling a seamless handoff.

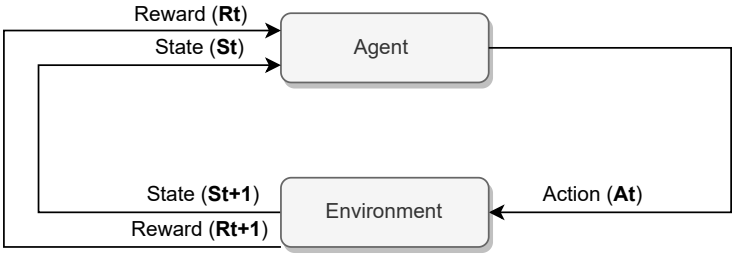

Fig. 4. Reinforcement Learning Loop

4 Conclusion

The design goals of the ADAPTO project revolve around the realization of 5G use cases such as URLLC, eMBB, and MTC while emphasizing energy efficiency. By synergizing the strengths of 5G Core Network SBA, advancements in NFV, 5G MEC, and SD-WAN, this project aims to establish a network framework tailored for telcos reliant on cloud-native architectures and next-generation networks. This paper has provided an overview of the foundational principles and technologies underpinning the ADAPTO framework, namely NFVs, MEC, SBA, and 5G MEC. Additionally, it offers a high-level insight into the ADAPTO EDGE agent, focusing on the control and orchestration of computing resources and network elements to facilitate intelligent scaling, edge-to-cloud transitions, and optimal selection of 5G backhauls.

Furthermore, a specific aspect of the ADAPTO framework regards the strategic integration of a Reinforcement Learning (RL) agent at the edge. This intelligent agent assumes a pivotal role in enabling seamless edge-to-cloud transitions by proficiently selecting the most suitable 5G backhaul. Leveraging its decision-making capabilities, the RL agent can ensure the desired Quality of Service (QoS) for applications, thereby contributing to the realization of 5G promises for URLLC.

Acknowledgment

This work was partially supported by the European Union through the ADAPTO project, part of the RESTART program, NextGenerationEU PNRR, CUP E63C2 2002040007, CP PE0000001.

References

1. Quoc-Viet Pham, Fang Fang, Vu Nguyen Ha, Md. Jalil Piran, Mai Le, Long Bao Le, Won-Joo Hwang, and Zhiguo Ding. A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art. *IEEE Access*, 8:116974–117017, 2020.
2. 3GPP. *System architecture for the 5G System (5GS)*, 2022. v16.12.0.
3. Davit Harutyunyan, Rasoul Behravesh, and Nina Slamnik-Kriještorac. Cost-efficient placement and scaling of 5G core network and MEC-enabled application VNFs. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 241–249. IEEE, 2021.
4. Sami Kekki, Walter Featherstone, Yonggang Fang, Pekka Kuure, Alice Li, Anurag Ranjan, Debashish Purkayastha, Feng Jiangping, Danny Frydman, Gianluca Verin, et al. MEC in 5G networks. *ETSI white paper*, 28(2018):1–28, 2018.
5. Qixia Zhang, Fangming Liu, and Chaobing Zeng. Adaptive interference-aware VNF placement for service-customized 5G network slices. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2449–2457. IEEE, 2019.
6. Marius Corici, Pousali Chakraborty, and Thomas Magedanz. A Study of 5G Edge-Central Core Network Split Options. *Network*, 1(3):354–368, 2021.
7. Pheobe Agbo-Adelowo and Petra Weitkemper. Analysis of Different MEC Offloading Scenarios with LEO Satellite in 5G Networks. In *2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6. IEEE, 2023.
8. Md Maruf Ahamed and Saleh Faruque. 5G backhaul: requirements, challenges, and emerging technologies. *Broadband Communications Networks: Recent Advances and Lessons from Practice*, 43:2018, 2018.
9. Min-Han Hung, Che-Chun Teng, Chin-Ping Chuang, Chi-Sheng Hsu, Jai-Wei Gong, and Mei-Chung Chen. A SDN Controller Monitoring Architecture for 5G Backhaul Networks. In *2022 23rd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pages 1–4. IEEE, 2022.
10. Aris Leivadeas, Nikolai Pitaev, and Matthias Falkner. Analyzing the performance of sd-wan enabled service function chains across the globe with aws. In *Proceedings of the 2023 ACM/SPEC International Conference on Performance Engineering*, pages 125–135, 2023.
11. Marius Corici, Ilie Gheorghe-Pop, Eleonora Cau, Konstantinos Liolis, Christos Politis, Alexander Geurtz, F Burkhardt, S Covaci, J Koernicke, F Völk, et al. SATis5 solution: A comprehensive practical validation of the satellite use cases in 5G. In *Proceedings of the 24th Ka and Broadband Communications Conference, Niagara Falls, ON, Canada*, pages 15–18, 2018.
12. Yi Zhang, Changqiao Xu, and Gabriel-Miro Muntean. Revenue-Oriented Service Offloading through Fog-Cloud Collaboration in SD-WAN. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 5753–5758. IEEE, 2022.