Joint Power Allocation and Network Slicing in an Open RAN System

Mojdeh Karbalaee Motalleb
School of ECE, College of Engineering, University of Tehran, Iran
Email: {mojdeh.karbalaee}@ut.ac.ir,

***Abstract*—**

***Index Terms*—**

I. INTRODUCTION

Network slicing, where several logical networks share a single physical network, has been considered as one of the key enablers for 5G.

Network slicing is a promising technology for 5G to provide a network as a service (NaaS) for a wide range of services that run on different virtual networks deployed on a shared network infrastructure.

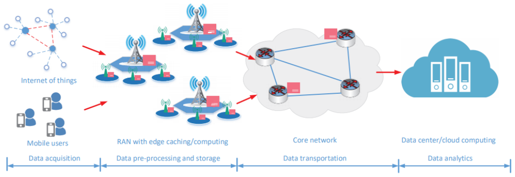

Fig. 1: Big data driven networking

A critical component of network slicing is resource allocation, which needs to ensure that slices receive the resources needed to support their mobiles/services while optimizing network efficiency.

In a sliced network, network operators realize the network slices instances to provide the required network characteristics to different services, pertaining to third-party tenants. Therefore, there will be Service Level Agreements (SLAs) between network operator and the tenants to declare the requirements of a particular service and the operator should fulfill these SLAs via instantiating appropriate network slices. Requirements of the service instances are specified in terms of Key Performance Indicators (KPIs), such as throughput, latency, availability, coverage, etc.

An important aspect of network slicing is to guarantee that the slices operate independently, i.e., the performance, congestion, failure, etc., in one slice will not negatively influence the performance of other slices which are sharing the resources. This means that providing protection from other slices is one to enable Internet of things (IoT) applications, including high data rate, numerous devices connection and low service latency. Network slicing and fog computing have been envisioned as promising solutions in service-oriented 5G architecture.

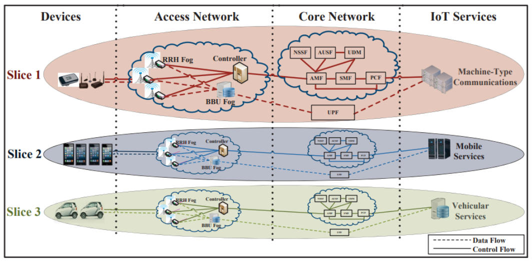

Fig. 2: Service-oriented Architecture of Fog-assisted 5G Network with Network Slicing

Assume a system model with different types of user equipment (UE). There exists two different types of UEs with different requirements: low-latency users (LUs) and high-data rate Tactile users (TUs).

The first group is very sensitive to the latency. The second group needs high data rate. Each slice sustains a queue for the incoming slice traffic, which has a backlog. The network model is assumed to be composed of a set of M base stations (BSs). We have three types of slice resource requirements.

- A fixed slice requests dedicated radio resources along time and frequency domains. Therefore, it is isolated from other slices and does not multiplex its resources with others. One example is the bandwidth parts (BWPs) defined in 5G (see 3GPP TS 38.211) that operate on disjoint parts of the spectrum with a given numerology.
- A dynamic slice requests a share of resources in terms of aggregate throughput. It can be mapped to the eMBB usage scenario, and thus a precise radio resource allocation is less important than the efficient use of available resources for a guaranteed throughput.
- An on-demand slice exhibits more stringent requirements in terms of latency, and hence it can be mapped to the URLLC scenario. Therefore, such slice should be assigned radio resources with short delay, in comparison with the dynamic slice [1].

A. Introduction to Network slicing

Network slicing is an important network architecture innovation in 5G that is also expected to be inherited in the next generation. Network slicing enables the coexistence of multiple isolated and independent virtual (logical) networks, i.e., slices, on the same physical network infrastructure. The advantages of network slicing are multifold. First, through the multiplexing of the virtual networks, network slicing supports multi-tenancy, i.e., multiple virtual network operators (VNOs) sharing the same physical network infrastructure. This reduces capital expense in network deployment and operation. Second, network slicing provides the potential to create customized slices for different service types with various Quality of Service (QoS) requirements, which can achieve service differentiation and guarantee service level agreement (SLA) for each service type. Third, as slices can be created on-demand and modified or annulled as needed, network slicing increases the flexibility and adaptability in network management.

A network slice is a virtual network which is implemented on top of a physical network in a way that creates the illusion to the slice tenant of operating its own dedicated physical network. Network Slicing offers a number of significant advantages that are particularly useful in the design of next generation wireless networks, namely:

- Slice isolation: The complete isolation of slices allows for a simpler and more efficient design of each slice with the goal of meeting the requirements of the particular vertical applications and services offered by the slice tenant. In addition, network failure, overload, or security attacks in one slice will not affect the operation of other slices in the network.
- Simplified service chains: In contrast to traditional cellular communications in which all services consist of the same functions, in network slicing each service may rely on a different subset of functions.
- Flexible VNF placement: NFV introduces an additional degree of freedom regarding the placement of these functions on the network. Intelligent placement may improve network performance and reduce operating costs.
- Transparent slice management: Subsets of the physical network resources might belong to different network domains (or even operators). Network Slicing provides an abstraction of the physical resources and makes slice management transparent to the slice tenant.

We have core and ran slicing in network slicing architecture. RAN slicing faces the following unique challenges:

- Resource interplay Since a service may consume multiple network resources, there exists an inherent tradeoff among the network resources. For example, in computing offloading services, the service latency consists of two elements: task transmission latency and task processing latency. If a user associates with a remote MEC server having abundant computing resources for task processing, a high task transmission latency will incur. On the other hand, if a user associates with a nearby MEC server having insufficient computing resources, it takes a longer time for task processing. In such a manner, the allocation of computing and communication resources is coupled with each other in the exemplary computing offloading services. Similarly, the allocation of multiple network resources is intertwined, which complicates the RAN slicing. **A joint multiple network resource allocation scheme should be judiciously designed to maximize network welfare.**
- Strict QoS requirements Compared with traditional 4G networks, 5G networks and beyond have stricter QoS requirements, including a higher throughput and a lower latency. Especially, the typical URLLC service in 5G requires ultra-high reliability (e.g., 99.999), which is much stricter than that of other services. In addition, the payload of data packets in URLLC services is usually small, such as 32 bytes. The transmission performance of short-length packets cannot be characterized by the traditional Shannon theory which is suitable for long-length packet transmission due to a large transmission overhead. Instead, the finite block length channel coding theory should be applied to characterize the achievable rate for short-length packets. Traditional QoS provisioning is unsuitable for short-length packet URLLC services with ultra-high reliability. Thus, **an accurate QoS provisioning for URLLC services is desired in the RAN slicing framework.**
- User mobility Due to the high network density, users may frequently move out the coverage of its associated network infrastructure, which results in a dynamic network topology. For example, high-mobility vehicle users can trigger handover frequently. The dynamic network topology changes the service traffic distribution, rendering previously optimal slice allocation suboptimal over time, degrading network performance, and may even violate users QoS requirements. When the network performance degrades to a threshold, adjusting existing slices or creating new slices will be triggered, which incurs slice reconfiguration overhead. Therefore, **dynamic yet efficient RAN slicing to accommodate user mobility remains a challenging issue** [2].

II. System Model and Main Problem

In the network slicing architecture, each service is admitted by specific slice. Also each service has a special QoS that must be reached. QoS parameters include:

- Delay
- Throughput
- Packet loss
- Out of order delivery

Different services have one or more of these QoS. AI-based methods become promising techniques to provide potential benefits to address the difficulties and complexities of slicing. We can use AI-based methods to provide **accurate service-specific traffic prediction**. Only with such accurately predicted service-specific traffic, RAN slicing can effectively facilitate network resource allocation to accommodate service demands in the near future. Recent studies show that AI-based methods, such as deep neural network (DNN) and long short-term memory (LSTM), are capable of accurately forecasting service-specific traffic load. AI-based methods can facilitate efficient resource allocation in RAN slicing. An online AI-based resource allocation decision process has the potential to achieve a

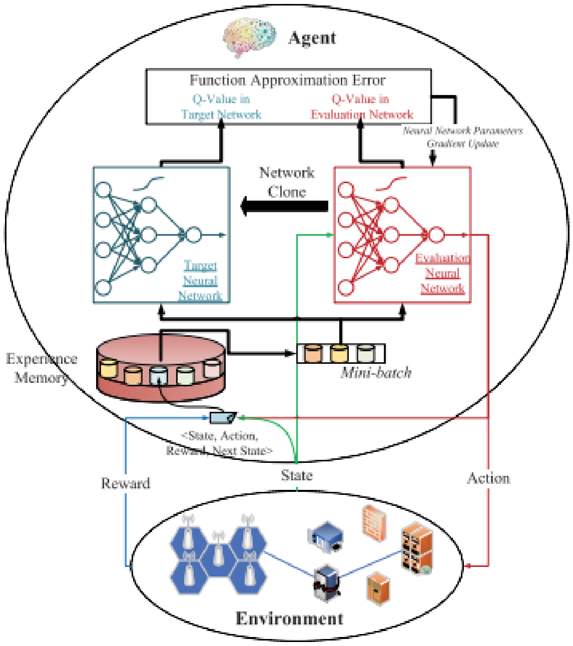
Agent
Function Approximation Error
Q-Value in Target Network
Q-Value in Evaluation Network
Neural Network Parameters Gradient Update
Network Clone
Target Neural Network
Evaluation Neural Network
Experience Memory
Mini-batch
<State, Action, Reward, Next State>
Reward
State
Action
Environment

Fig. 3: Deep Q learning structure

low complexity after an offline training procedure, which addresses the high computational complexity challenge in the conventional model-based optimization methods. Note that traditional RL methods, such as Q-learning, suffer from the curse of dimensionality, which are only suitable for RAN slicing problems in small-scale networks. Deep RL methods incorporate deep learning networks in the RL framework can effectively address the complexity issues in large-scale networks.

Cooperative Game with Distributed Learning while existing applications of distributed learning in this field generally consider non-cooperative games where the Nash equilibriums are achieved, there is a great potential to adopt the concept of cooperative game, where tenants/slices can learn to make decisions in an organized and cooperative way, in order to maximize the global social welfare instead of their own interests. In this way, a Pareto optimum can be expected instead of the Nash equilibrium.

A. system model

In this part, deep reinforcement learning (DRL), which focuses on how to interact with the environment by trying alternative actions and reinforcing the tendency actions producing more rewarding consequences, is assumed to be a promising solution for network slicing [3]. Assume we have V services and S slices. Each service needs physical resource blocks (sharing the aggregated bandwidth W), RRHs (radio unit head) and CU (control unit) in RAN and VNFs in core (set of requirements $\mathbf{q} = \{q_1, ..., q_{n_v}\}$) and requests specific QoS includes delay, throughput, packetloss, ... So each service v needs specific demands $\mathbf{d} = \{d_1, ..., d_{n_v}\}$. The target is to maximize long-term reward expectation $E\{R(\mathbf{q}, \mathbf{d})\}$.

$$\text{argmax}_{\mathbf{d}} E\{R(\mathbf{q}, \mathbf{d})\} \tag{1a}$$

$$\text{subject to} \quad \mathbf{q}, \mathbf{d} \tag{1b}$$

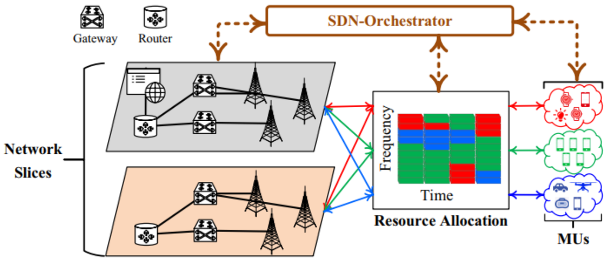

Fig. 4: Network slicing structure [4]

Set of QoSes $\mathbf{d} = \{d_1, ..., d_{n_v}\}$ includes M/M/1 delays, throughput and packetloss.

1) Delay: Assume the packet arrival of UEs follows a Poisson process with arrival rate λ_v for the UEs of the v^{th} service. Therefore, the mean arrival data rate of UEs is

$$D_v = \frac{1}{\mu - \lambda_v}. \tag{2}$$

where $1/\mu$ is the mean service time and λ is the rate of arrival packets.

2) Throughput: Let ρ_v be the mean downlink rate of user i served by service v, which is, for simplicity, defined by Shannon theory as follows

$$\rho_v = w_v log(1 + \frac{P_v}{N_v}) \tag{3}$$

where w_v is the bandwidth of service v.

References

[1] R. Schmidt, C.-Y. Chang, and N. Nikaein, "Slice scheduling with qos-guarantee towards 5g," pp. 1–7, 2019.

[2] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "Ai-assisted network-slicing based next-generation wireless networks," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 45–66, 2020.

[3] R. Li, Z. Zhao, Q. Sun, I. Chih-Lin, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74 429–74 441, 2018.

[4] Y. Hua, R. Li, Z. Zhao, H. Zhang, and X. Chen, "Gan-based deep distributional reinforcement learning for resource management in network slicing," *arXiv preprint arXiv:1905.03929*, 2019.