Chapter 1

AI-based 5G RAN slicing

Francesco Agostini
Nicola Gastaldello

1.1 Introduction

Mobile networks, in the near future, will face several technical challenges due to the ever-increasing service requirements both from the ISP point of view (such as operational optimizations, i.e., cost and energy efficiency) and the users perspective (such as throughput, reliability, availability, latency)[1]. This increment of network demands derives from the simultaneous increase of mobile terminals, the increasing diversity of requested services and the rapid evolution of new application scenarios, which range from the usual telecommunications sector to vertical industries, including smart factories, autonomous driving, the health sector and so on (Fig. 1.1)[10]. In particular, all these application areas has been grouped in four categories by the Third Generation Partnership Project (3GPP), as follows[2]:

- **Enhanced mobile broadband (eMBB)**, requiring very high data traffic and bit rate
- **Critical communications (CriC)**, for low latency, high density distribution and ultra-high reliability communications
- **Massive Internet of Things (mIoT)**, characterized by large numbers of connected users in high user-density environments
- **Vehicle-to-X (V2X) communications**, requiring low latency, high speed and reliability, and position accuracy.

Thus, it follows that 5G mobile networks must meet the following goals: the achievement of the aforementioned service requests together with the adoption of strategies not providing a singular mobile network architectural model, in order to provide cost and energy-efficient solutions and a design which aims to be as future-proof as possible. For these reasons, the upcoming network designs need to be scalable and full-flexible, two fundamental requirements which play a key-role in the insurance of the adaptability of the mobile network to the specific scenario that the various use cases offer.

Currently, the key enabler to meet the aforementioned functional and operational diversity is Network Slicing, defined by the Next Generation Mobile Networks (NGMN) Alliance as the concept of running multiple logical networks as independent flows upon a common physical infrastructure. More in details, a network slice is a self-contained, virtualized and independent end-to-end network that allows operators to execute different deployments in parallel, each one based on its own architecture[3]; furthermore, a network slice is capable to offer customized functionalities, including those in the UE, and delivers services to each device using network function chains and their corresponding computing, networking and storage resources.

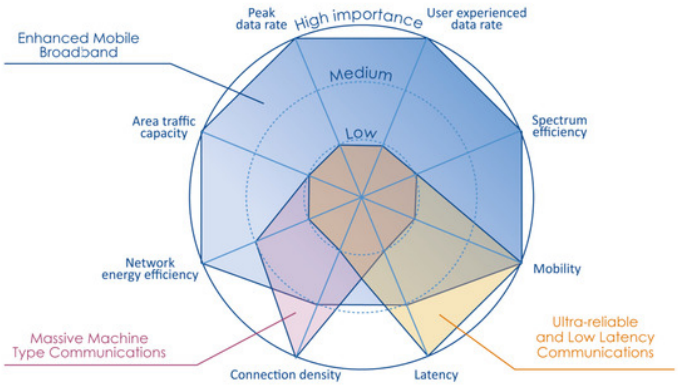

Figure 1.1: 5G use cases and requirements: the further the distance of a requirement from the center, the higher is its importance to the corresponding use case.

However, the implementation of such concept in 5G networks gives birth to a variety of challenges; one of the biggest adversities lies in the management of the repartition and virtualization of the network into different

slices of the RAN[1]. Furthermore, considering that network slicing requires innovations at each layer of the protocol stack, in both the CN and RAN, it's important to highlight the fact that that most the lately research studies for 5G communication systems have given priority cn-slicing, focussing on the network infrastructure and its virtualization; hence, it derives the importance of slicing the RAN, in order to achieve successful 5G cellular network deployments.

Besides Network Slicing, there are other technologies which can be considered as key enablers for 5G mobile cellular networks. One of them consists on the so called mmWave communications, that have been raising interest in the scientific community. MmWave communications exploit the currently unused and large portion of the radio spectrum lying in the bands between 30 and 300 GHz and promises to achieve the throughput requirements of future mobile networks[4], since the possibility, due to the small wavelength of mmWave signals, of integrate large numbers of antennas in relatively small areas (i.e., cover of a cellphone or a chip), enabling the implementation of concepts like massive mimo[4]. Another technique that is deemed to be promising is Carrier Aggregation (CA), which enables multi-connectivity at the MAC layer by exploiting multiple links (called cc), possibly at very different carrier frequencies. This concept has been already widely used in lte-Advanced networks[5][7], in order to aggregate more bandwidth, and thus achieving higher data-rates, while, at the same time, exploiting the frequency diversity as a mean to overcome the interference phenomena.

Furthermore, given the ever-increasing complexity of the networks, and, as mentioned before, the emergence of novel applications and uses cases, a 5G network cannot be fully operative without the adoption of machine learning (ML) routines. Hence, the support of artificial intelligence (AI), and in particular neural networks, is expected to play a key role in making the 5G vision conceivable.

In this article we provide a potential way to combine the use of Recurrent Neural Networks (RNNs) applied to the concept of Network Slicing. The content of this paper is organized as follows: in Section 2 we provide a brief Network Slicing overview, explaining the functioning that underlies this concept; in Section 3, we introduce RNNs and discuss on how they can be applied to the 5G field. Section 4 provides our proposal of NN-based Network Slicing, including some simulation-based performance analysis and finally we conclude this work and highlight potential future improvements with Section 5.

1.2 Network slicing overview

As mentioned in the section above, Network Slicing, which spans several domains, such as Core Networks and Radio Access Networks, represents the key solution that addresses the diverse requirements of 5G mobile networks, providing the necessary flexibilty and scalability[15]. The concept of Network Slicing provides for partitioning the physical network into many virtual netwoks, each one customizable and optimizable for every specific type of application[10]. Then, through virtulization technologies, the physical resources can be dynamically scheduled to logical network slices, according to user requests, allowing the virtualized networks to adapt to changing requirements. More specifically, according to[2], network slicing consists of the three following layers:

- **Service Instance Layer**, the end user services that can be supported. Each service is represented by a service instance.
- **Network Slice Instance Layer**, represents the network slice instances that can be provided, including the network features required bu the service instance.
- **Resource Layer**, provides the necessary virtual and physical resources and network functions to create a network slice instance.

In the following, related key enabling technologies, such as network function modularization and virtualization, dynamic service chaining and management are discussed.

1.2.1 Virtualization and Modularization

Virtualization technologies, which represent an important foundation for Network Slicing, allows to remove dependencies on dedicated hardware, thus enabling a flexlible slice creation on a shared physical infrastructure.
This type of service-based architecture introduces more granularity and decoupling on network functionalities[6]: in this way, since each service is self-contained and realized by a specific functionality, the 5G network is capable to comprise highly cohesive services. In addition, each of these services can be independently updated, translating in a less impact on the other services. This modularity allows to flexibly chain and connect the network function components (NFCs) to end-to-end network slices on demand.

1.2.2 SDN and Service Chaining

Network programmability and centralized network intelligence are achieved through Software-Defined-Networking (SDN), decoupling the data plane from the control plane. In this way, new network services can be created dynamically, based on user requirements and network conditions[8]. The key feature of SDN is the efficient support of multiple client istances on a common infrastructure, due to the capability of the SDN controller to dynamically allocate resources for network slices belonging to the same context, through programmable inrefaces.
Service chaining, instead, introduces an application-driven-networking approach, making flexible to chain both control and data plane functions: the traffic generated by specific users or applications only traverses a particular set of functions, and each class of service is addressed to necessary network functions, according to service chaining policies. In this way, dynamic service chaining allows network operators to create, remove, scale and modify network functions of a network slice, according to users demands and network information.

1.2.3 Management

Network slicing introduces more complexity on the management of the network, especially when large number of network slices are supported. For this reason, automated managements solutions are critical to find, especially because there is not yet a unified vision about the strategy of network slice management[fonte]. More in general, however, we can logically divide the life cycle of a network slice into four main phases[9]. The first is the design phase, in which a catalog of NFCs is created, in order to generate related network slice templates, based on the available NFCs and on the performance requirements from users.
The second phase is the activation phase: a specific network slice template, based on service requirements from users, is selected. In other words, a mapping between service requirements and slice templates is performed, while the NFCs of the specific network slice are instantiated and chained together via virtual connections[15].
The third phase is is the Run-TIme Assurance phase, in which the performance indicators such as resource utilization, network function load and QoS are constantly monitored. In this phase, machine learning tools could be adopted to analyze all the performance information and to reconfigure network slice instances for meeting the service requirements.

The last phase coincides with the Decommissioning phase: based on the business strategy or the service requirements, active network slice instances are deactivated, and their related resources are finally released.

1.3 Recurrent Neural Network (RNN) overview

As a discipline, 5G systems have been evolving, like artificial intelligence (AI) and its sub-categories. To make 5G wireless mobile networks to be predictive and proactive, the use of AI is needed. At the state-of-art, there are a lot of attempt to use AI to improve 5G. The main application are developed in problems of optimization, for example the dynamic frequency and bandwidth allocation and path loss prediction model for urban environments, and in data analysis (i.e., anomaly detection in mobile wireless networks)[11].

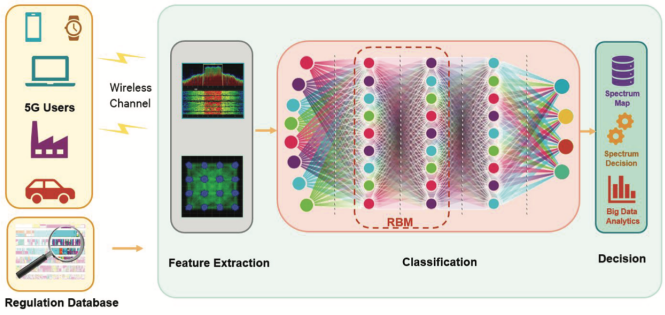

Figure 1.2: Structure example of an integration between 5G-based applications and AI-tecnique .

AI is a branch of computer science whose goal is the simulation of the human intelligence over machines. The most developed application of AI developed in recent times is machine learning (ML) and it aims to create algorithms which can automatically learn and improve from the past experience, without explicit programming. There are three main subclasses of ML[12]:

- **Supervised learning:** is the task of learning a function, based on input-output pairs, which maps an input to an output. Therefore, given training data, which are composed by the samples and the labels

(correct answer), the algorithm will try to create the map function between the samples and the label.

- **Unsupervised learning:** is the task of leaning as more as possible from the data, since we have the input data but not the corresponding output; thus, the model doesn't need to be supervised.

- **Reinforcement learning:** is the task of finding a tradeoff between exploration and exploitation, namely between the unknown knowledge and the current knowledge. Therefore, there will not be input/output pairs, but the algorithm will take decision based on a target to reach, in order to maximize reward in a special situation.

An important class of supervised learning is deep learning (DL), whose main feature is to use artificial neural networks (ANN) in order to extract high level features from the input. ANN are computing systems inspired by human brain, namely an interconnected group of nodes. The nodes are artificial neurons and they are arranged in several layers; the artificial neurons positioned in differents layers are linked between them.

In this work we use Recurrent Neural Networks (RNN), a class of ANN, where connections between neurons form a directed cycle along a temporal sequence. Therefore, with this particular type of ANN, we can analyze temporal data inputs and generate sequential data output[14]. In particular, we use the so-called LSTM RNN (Long Short-Term Memory), that uses a special combination of hidden neurons to implement the memory cells. These cells are needed to preserve information for periods of time without modification. For these reasons, we decided to use this type of architecture, as it allows to predict the future behavior at time $t+1$, based on the past[16].

1.4 Our solution

This section offers our proposal on how to combine AI with 5G mobile cellular networks. More specifically, we first simulated the environment of a 5G cell, implementing a static band allocation for each class of applications; in other words, in this scenario, the BS allocates fixed portion of the available bandwidth to each type of slice, independently from traffic demands. Then, with the support of RNN, we tried to achieve a dynamic and more efficient bandwidth usage and allocation, based on the users requests, in order to improve the system performance. The rest of this section is organized as follows: after a brief description of the scenario and the technical parameters,

a simulation-based performance analysis will be provided. For complete information, we want to highlight that the simulation was entirely written in Python.

1.4.1 Scenario and parameters

With the goal of achieving a more efficient band allocation with the support of RNN, as mentioned in the previous sections, we built a dynamic scenario which simulates as well the environment of a cell of a 5G cellular system. We adopted a dynamic scenario in order to have the possibility of studying different situations, e.g., urban or suburbs environments, by modulating the population density and the population types. From what is explained above, the main components of our scenario are the following: the base station (BS), located at the centre of the cell, and the clients, users connected to the base station, generating and consuming traffic (Fig. 1.3).

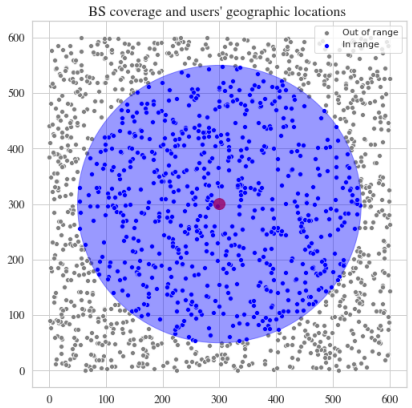

Figure 1.3: Simulation scenario: BS, coverage area and clients.

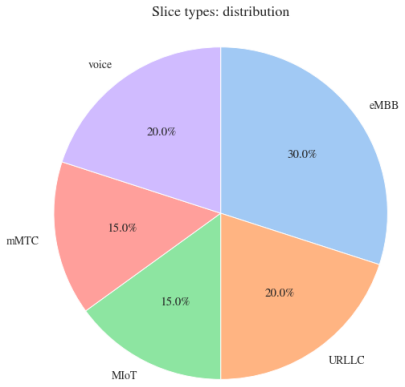

Figure 1.4: Slice distribution percentage over the users.

The base station parameters have been chosen in according to the 5G specifications:

- **Capacity:** maximum amount of traffic supported ($20Gbps$);
- **Coverage:** the coverage area of the BS is represented by a circle of radius = $250m$;

- **Ratios:** portion of the total capacity to be assigned to a specific type of slice;
- **Position:** geographic coordinates in the map.

The *ratios* parameter shows that the the base station is configured to serve 5 different types of traffic: eMBB, URLLC, mMTC, MIoT and voice. The traffic parameters of each class of slices are the following:

- **Maximum bandwidth:** maximum bandwidth per user that the basestation can give to this type of slice;
- **Client weight:** percentage to distribute traffic in the cell, respecting the real distribution;
- **Bandwidth guaranteed:** minimum bandwidth threshold for which the base station can accept or reject the connection requests, in order to guarantee the QoS (i.e., the average throughput per user must be above the bandwidth guaranteed threshold);
- **Range of request:** minimum and maximum bandwidth (bps) that each specific application can request.

In Table 1.1, we reported the chosen parameters for each type of slice, in order to simulate an accurate and reliable 5G scenario.

	Bandwitdh max (Mbps)	**Client weight**	**Bandwidth guaranteed (Mbps)**	**Range of request (Mbps)**
eMBB	100	0.3	0	(4, 500)
URLLC	10	0.2	5	(5, 8)
mMTC	10	0.15	1	(1, 8)
MIoT	10	0.15	1	(1, 8)
voice	1	0.2	0.5	(4, 8)

Table 1.1: Slice types: traffic parameters configuration.

Finally, the second component of our simulation model are clients: a specific type of slice has been assigned to each client, as well as a random

mobility pattern, selected from the following list: car, walk user, stationary user, slack user and public transport. All the users are in constant motion over the area, according to the above mentioned mobility pattern. When a user leaves the base station coverage area, the connection to the current BS is disabled, as the model foresees that it will be served by the nearest adjacent station.

1.4.2 Simulations

To reach our goal, we developed two different environments, in order to simulate the two bandwidth allocation approaches (i.e., static and dynamic). Nevertheless, the two environments work similar. First, we populate the scenario above described with a fixed number of clients, assigning to each of them a slice type and a mobility pattern. Once the population is completed, only a portion of the users in the coverage area are activated, following a random distribution. Now, for each active clients, we compute the following steps:

- **Connection:** In this phase the active client generates a request based on the assigned slice type. The base station accepts the user only if the bandwidth guaranteed is respected, otherwise the connection will be refused. When the client connection is accepted, it will start to consume, otherwise the request will remain in memory and will be satisfied as soon as possible;
- **Consume:** The client can consume in two different ways: the request is immediately satisfied, or the request is above the maximum bandwidth and so the clients will receive the amount allowed according to the slice type until the usage remaining is below this threshold, and the request will be finally satisfied. Once a request is totally served, the user will release resources to the basestation and disconnect.
- **Movement:** All clients move with different velocity based on the assigned mobility pattern. In each iteration, the base station checks which are the users in the coverage area, discarding those who have gone out.

The above steps works equal in both the simulations. The main difference between them is the dynamic band allocation that we introduce with our proposal. More specifically, at every cycle we compute a new dynamic and adaptive bandwidth allocation, based on the past requests. This is

achieved through a prediction of our pre-trained model of an RNN, namely, at each iteration, based on the total request of the clients of the previous cycle, we aim to predict the behaviour of the next total amount of request per slice. By doing this, we can redistribute the base station ratios, preparing it to undergo the new traffic flow.

1.4.3 Performance analysis

As previously introduced, the performance analysis has been carried using the means of simulation, in particular using Python libraries. Our approach (dynamic band allocation) has been benchmarked against the static band allocation one, in order to highlight the performance improvement introduced by the use of a RNN, combined with a smarter and more efficient spectrum usage policy. Regarding the performance measurements, we focused on the average throughput per user and the average number of connected users metrics, as well as a spectrum efficiency one.
As we can see from Fig. 1.5, which shows the average eMBB througput per user, the flexibility of our solution enables a more effective explotation of the radio resources: the eMBB throughput is kept almost constant (around 85 Mbps), while, at the same time, the number of connected users significantly grows (from 62 to 77 on average, +24%, Fig. 1.5).

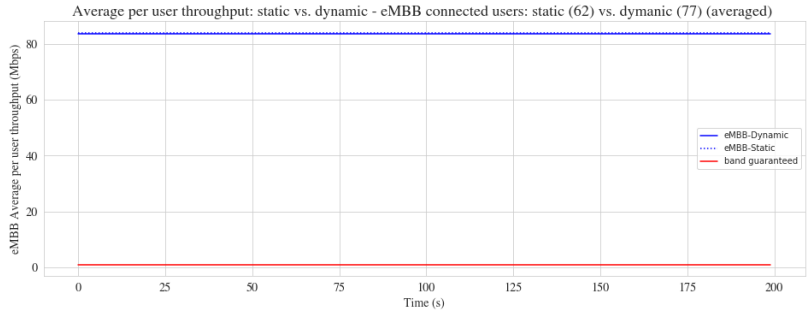

Figure 1.5: eMBB average per user throughput.

Regarding URLLC traffic, Fig. 1.6 shows that our system is capable to sustain 40% more of consuming users, but the price to pay is a slight reduction of the average throughput per user. In any case, however, the throughput is constantly kept above the minimum threshold of 5 Mbps, en-

suring in this way the QoS. It is important to highlight the importance of serving a larger number of users while keeping the QoS, specially in the case of URLLC traffic, whose requirements are ultra reliability and low latency. In this way, in fact, the waiting times of each URLLC user are significantly reduced, and our solution allows to improve the overall performance in terms of latency and reliability, meeting the strict technical requirements of this specific type of slice.

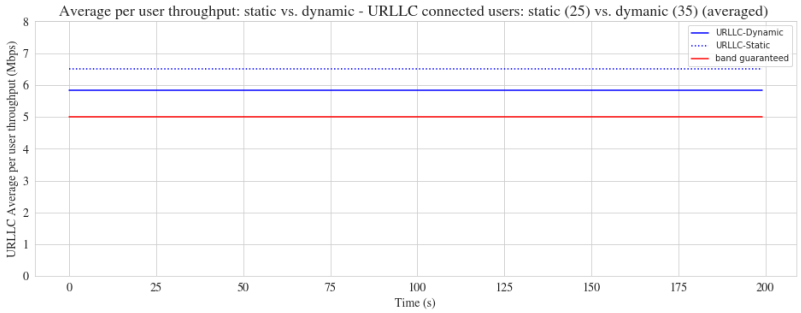

Figure 1.6: URLLC average per user throughput.

In Fig 1.7, 1.8, 1.9 are shown the trends of MIoT, mMTC and voice throughput respectively. As we can see, in all the cases the number of connected users significantly increases, while the bit rate is slightly lower, but always well above the minimum threshold guaranteed. The only exception is represented by the mMTC slice type, which doesn't show significant performance increases, neither in terms of served users, nor in terms of average throughput. This phenomena is probably due to the randomness of the parameters introduced by the simulations (more specifically, the randomness of the user request and the number of active clients) could negatively affect the results.

The effectiveness of our proposed solution is finally summarized in Fig 1.10 and 1.11. The flexibility introduced by the dynamic band allocation strategy, combined with the support of the RNN, allows an adaptive and more efficient radio resources allocation. In particular, we achieved a 10% overall increase in the total available bandwidth usage, while increasing the number of total connected users by 20%, ensuring the QoS and widely meeting the technical requirements of each specific class of traffic (in particular, the bandwidth guaranteed).

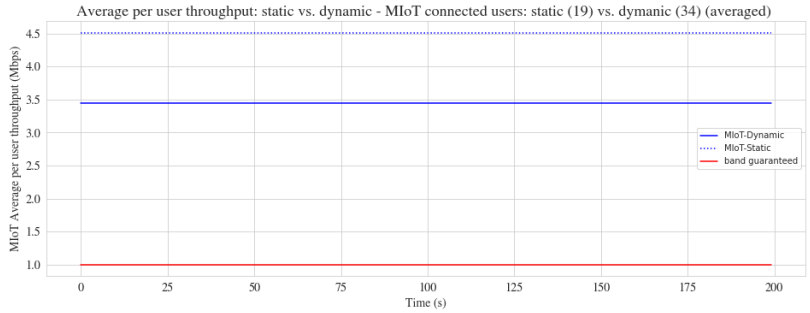

Figure 1.7: MIoT average per user throughput.

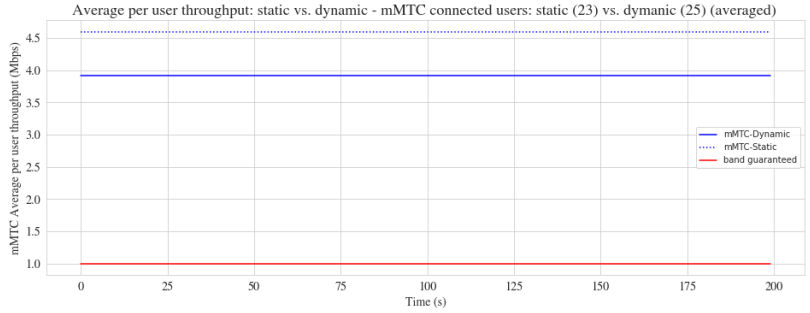

Figure 1.8: mMTC average per user throughput.

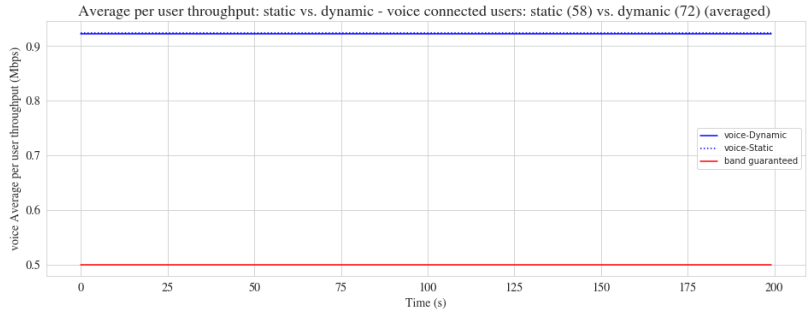

Figure 1.9: voice average per user throughput.

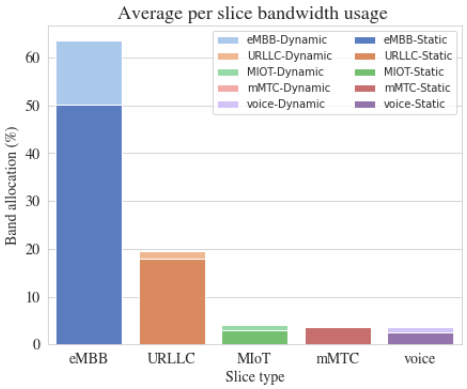

Figure 1.10: Summarization of per slice bandwidth usage, static vs. dynamic.

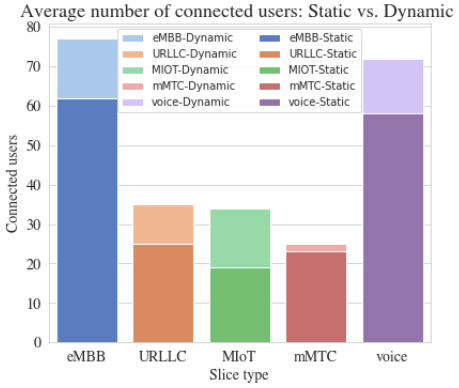

Figure 1.11: Summarization of per slice average connected users, static vs. dynamic.

1.5 Conclusions and future works

Network slicing is a fundamental concept that will allow to meet the performance requirements that future mobile networks will demand. To support the development of 5G networks, in addition to network slicing, several other techniques will play a key role, such as Carrier Aggregation and Massive MIMO technologies, as well as the exploitation of previously unexplored portions of the spectrum, the so called mmWave frequencies. In addition, AI-based framework will represent a fundamental tool in order to enable the imminent demands on 5G. These AI-based models, applied to the 5G world, could be very effective for enabling network evolution.
In this work, we proposed a solution which applies a DL algorithm to the bandwidth allocation strategy. As we can see from the section above, we managed to achieve more flexibility and a more efficient usage of radio resources. Moreover, through the smart and adaptive band allocation strategy, we achieved a 17% increase in the total available bandwitdh usage, increasing at the same time the number of total connected users by 30%, guaranteeing the QoS for each specific class of traffic.
Regarding the future works of our application, a future useful improvements would be to adapt the simulation to a more realistic 5G scenario, including more base stations and cells, and increasing the density and the number of clients. In addition, with greater computational power available, it would

be possible to train the RNN with more samples, in order to obtain best results. Furthermore, the training dataset could be extended with more features, in order to make the response of the scheduling policy of the base stations more dynamic and efficient.

Bibliography

[1] Peter Rost, Christian Mannweiler, Diomidis S. Michalopoulos, Cinzia Sartori, Vincenzo Sciancalepore, Nishanth Sastry, Oliver Holland, Shreya Tayade, Bin Han, Dario Bega, Danish Aziz, and Hajo Bakker, *A Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks*, IEEE, 2017.

[2] NGMN, *NGMN 5G White Paper*, NGMN Alliance, 2015.

[3] NGMN, *Description of network slicing concept*, NGMN Alliance, 2016.

[4] Sundeep Rangan, Senior Member IEEE, Theodore S. Rappaport, Fellow IEEE, and Elza Erkip, Fellow IEEE, *Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges*, Proceeding of the IEEE, 2014.

[5] Jeanette Wannstrom, *LTE-Advanced*, 3GPP, 2013.

[6] NGMN, *Service-Based Architecture in 5G*, NGMN Alliance, 2018.

[7] Klaus Ingemann Pedersen, Frank Frederiksen, and Claudio Rosa, Nokia Siemens Networks Hung Nguyen, Luis Guilherme Uzeda Garcia, and Yuanye Wang, Aalborg University, *Carrier Aggregation for LTE-Advanced: Functionality and Performance Aspects*, IEEE, 2011.

[8] Salvatore D'Oro, Student Member, IEEE, Laura Galluccio, Member, IEEE, Sergio Palazzo, Senior Member, IEEE, and Giovanni Schembra, *Exploiting Congestion Games to Achieve Distributed Service Chaining in NFV Networks*, IEEE Journal, 2017.

[9] Xenofon Foukas, Georgios Patounas, Ahmed Elmokashfi, and Mahesh K. Marina, *Network Slicing in 5G: Survey and Challenges*, IEEE, 2017.

[10] Alexandros Kaloxylos, *A Survey and an Analysis of Network Slicing in 5G Networks*, IEEE, 2018.

[11] Manuel Eugenio, Morocho Cayamcela and Wansu Lim, *Artificial Intelligence in 5G Technology: A Survey*, IEEE, 2018.

[12] Manuel Eugenio, Morocho-Cayamcela, Haeyoung Lee and Wansu Lim, *Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions*, IEEE, 2019.

[13] Miao Yao, Munawwar Sohul, Vuk Marojevic and Jeffrey H. Reed, *Artificial Intelligence Defined 5G Radio Access Networks*, IEEE, 2019.

[14] Robert Nisbet, Gary Miner and Ken Yale, *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier, 2018.

[15] Shunliang Zhang, *An Overview of Network Slicing for 5G*, IEEE, 2019.

[16] Ian H. Witten, Eibe Frank, Mark A. Hall and Christopher J. Pal, *Data Mining*, Elsevier, 2017.