ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
AUTOMOTIVE PER UNA MOBILITA INTELLIGENTE

CICLO XXXIV
SETTORE CONCORSUALE 09/F2
SETTORE SCIENTIFICO DISCIPLINARE ING-INF/03

MODELING AND OPTIMIZATION TECHNIQUES FOR ADVANCED VEHICULAR NETWORKING

Presentata da
ELISABETTA AMATO

Supervisore
Prof. CARLA RAFFAELLI

Co-supervisor
Prof. DANIELE TARCHI

Coordinatore Dottorato
Prof. NICOLÒ CAVINA

ESAME FINALE ANNO 2022

"There are two primary choices in life: to accept conditions as they exist, or accept the responsibility for changing them."
(cit. Denis Waitley)

Contents

Abstract

This thesis deals with optimization techniques and modeling of vehicular networks.
Thanks to the models realized with the integer linear programming (ILP) and the heuristic ones, it was possible to study the performances in 5G networks for the vehicular. Thanks to Software-defined networking (SDN) and Network functions virtualization (NFV) paradigms it was possible to study the performances of different classes of service, such as the Ultra Reliable Low Latency Communications (URLLC) class and enhanced Mobile BroadBand (eMBB) class, and how the functional split can have positive effects on network resource management. Two different protection techniques have been studied: Shared Path Protection (SPP) and Dedicated Path Protection (DPP). Thanks to these different protections, it is possible to achieve different network reliability requirements, according to the needs of the end user.
Finally, thanks to a simulator developed in Python, it was possible to study the dynamic allocation of resources in a 5G metro network. Through different provisioning algorithms and different dynamic resource management techniques, useful results have been obtained for understanding the needs in the vehicular networks that will exploit 5G. Finally, two models are shown for reconfiguring backup resources when using shared resource protection.

Chapter 1

Introduction

Smart mobility was born with the aim of making the roads safer, exploiting a set of technologies in wide development, thanks to the attraction by companies and international research. To make roads and cities intelligent, a telecommunications network is needed that allows vehicles to communicate with the infrastructure, with the vehicles themselves and with all the smart devices present in the network. This network must be reliable and offer a continuous connection to the devices so that it can be secure for the users. Over the years, various protocols have followed one another with the aim of achieving the performance, from 802.11 P to 4G. The evolution of 4G for vehicles (C-V2X) [B1] has made it possible to approach the requisites necessary to make vehicles safe, but it is with 5G that service classes have been defined that perfectly enclose the requirements of reliability and vehicular safety [B2] [B3] . For this reason, companies and telephone companies are aiming at 5G mobile networks, which makes it possible to achieve a latency lower than 10 milliseconds, a very high number of connections available at the same time and edge computing, which allows much faster processing with local management. In order to achieve these key performance indicators (KPI), 3 service classes have been defined: enhanced Mobile BroadBand (eMBB), massive Machine Type Communications mMTC and Ultra Reliable Low Latency Communications (URLLC). With the eMBB class, the aim is to allow high effective data transmission speed for the user and greater network capacity, which will be able to support the transmission of video streams to a greater number of active users simultaneously, even when

they are concentrated in limited areas. The mMTC class of services is basically the revisiting of the low speed and low consumption connection modes already standardized in the most recent LTE networks, such as NB-IoT (NarrowBand IoT). The URLLC class of services, on the other hand, will represent the true technological innovation of 5G, opening the way to potential new applications, for example in the automotive and industrial sector, where real-time response times are needed with orders of magnitude less than possible with today's LTE networks.
In order to guarantee the desired performance, these classes of service must be at the same time optimized and well connected network. With the increase of the antennas in the area, functions distributed in the network must therefore be associated which will allow the use of the same physical network, with the achievement of the different KPIs required by the different services [B4]. This new architecture is based on a functional separation of the processing capacity of mobile radio sites (baseband): the so-called functional split. This technology separates the real-time and non-real-time functions: the first most critical are managed on the site where the antennas closest to the user are present, the second are centralized and virtualized at higher network levels. The physical network can therefore be divided into several sub-parts, which must be optimized according to use. The connection between the two parts of the split takes place through an efficient ethernet fronthauling interface. The network chain is then divided into several parts: Radio Unit (RU), Distributed Unit (DU), Centralized Unit (CU), Core / Cloud. In this way, the baseband is able to control thousands of cells, increasing network efficiency through dynamic management of resources to respond to rapid traffic changes. In addition, baseband virtualization through paradigms as Network Functions Virtualization (NFV) and Software-Defined Networking (SDN) introduce agility, flexibility, reliability, and security[B5]. The application of these paradigms is called network slicing, which allows the coexistence of multiple classes of service with different KPIs within the same physical network. Protection schemes need to be deployed in support of different slices on the same transport infrastructure and related resources need to be minimized accordingly, to possibly limit cost and energy consumption.

The objective of this PhD thesis is to study methods for optimizing and modeling the vehicular network, allowing the achievement of the necessary requirements to make the network reliable and secure.

Chapter 2

Network solutions for vehicular networks

The automotive sector offers different service, as autonomous driving, connection between vehicles, on-board signaling, communication with other devices. To provide complete and continuous connectivity among vehicles and with the network that surrounds them is necessary a communication network infrastructure. Over the years, various technologies have followed one another (from 4G to 802.11p), which however have not allowed to achieve the necessary requirements to make the vehicular network reliable and secure. On the evolution of 4G the first experiments involving the use of Cellular Vehicle-to-everything (C-V2X) are taking shape. This technology, based on the use of release 14 of the 3GPP [B1], and then evolved with the subsequent specifications, allows both direct interaction between devices without using the network, and connection with the infrastructure and the network itself.
However, to make the most of it, the network needs to be reliable and with high-level of service continuity. These are precisely the key points of 5G and beyond objectives, which is why the automotive industry, telecommunications companies and telephone companies push for the fifth generation mobile networks to become the de facto standard for connected or self-driving vehicles. The ability to have a latency lower than 10 milliseconds, the very high number of connections available at the same time and edge computing, which allows much faster processing with local management, are fundamental for vehicular communication.

In this chapter, solutions for the vehicular network will be shown, starting from the description of the C-V2X to get to 5G, going to describe the innovations compared to the previous generation that allow the achievement of the network requirements. Finally, to complete the description of the solutions, the C-RAN and VEC architectures will be described.

2.1 Paradigms C-V2X

The importance of Cellular V2X has grown rapidly, and inherits the results of previous standardization work obtained by various standardization bodies. In 3GPP documents the term V2X refers to a communication between different entities. In particular, it is possible to define the following classification:

- **V2V**: are the set of systems and functionalities that allow cars to "communicate" with the other cars nearby, sending signals, information and warnings to drivers in case of danger: information in a circle between vehicle and the other include position, direction and speed and elements processed by the on-board systems to predict potential collisions

- **V2I**: it rappresents any type of communication between vehicles and infrastructure. Through the protocols, it is possible to exchange security information with the infrastructure, which will in turn be responsible for managing the data and transmitting relevant information to the machines. The idea is to replace road signs (traffic lights and signs), with intravehicular information, to optimize traffic management and promptly report hazards in the route. In addition to replacing traffic signs, communications with the infrastructure will also serve to warn vehicles of possible road hazards (such as adverse weather conditions or road works in progress and lane restrictions). Communications with the infrastructure, together with vehicle communications, will ensure an adequate knowledge of traffic and effectively prevent accidents, and effective disposal of road traffic.

- **V2N**: it will be necessary to provide real-time information to the

various vehicles. In addition to the realization of a V2I-I2V communication, the vehicles will be connected in real time to the network, thus having the possibility to have real-time traffic information (and plan route changes), as well as having the Cloud services available.

- **V2P**: it is a type of connection between vehicles and people. Pedestrian detection systems can be implemented in vehicles, infrastructure or pedestrians themselves to provide alerts to drivers, pedestrians or both.

V2X allows to cover a large number of use cases, which directly depend on the resources and their purpose ([B6]). These categories are grouped into:

- **Safety**, to reduce the frequency and severity of vehicle collisions.
- **Advanced driving assistance**, for operations relating to autonomous vehicles.
- **VRU (Vulnerable Road User)**, to make safety the interaction between vehicles and non-vehicle users.
- **Convenience**, to offer diagnostic services and software updates for the vehicle.

Table 2.1: Requirements of V2X autonomous driving use cases.

Application	Main communication mode	Latency (ms)	Reliability	Data rate (Mbps)
Vehicles platooning	V2V, V2I	10-20	90-99.999 %	0.012-65
Advanced driving	V2V, V2I	3-100	90-99.999 %	0.096-53
Extended sensors	V2V, V2I, V2P	3-100	90-99.999 %	10-1000
Remote driving	V2N	5	99.999 %	25 (Uplink) 1 (Downlink)

These use cases require the most restrictive performance and have captured the interest of the 3GPP, which has further classified them into 4 groups and defined the required requirements (2.1). Vehicles platooning forming a group of vehicles moving in the same direction for short

distances. Advanced driving allows vehicles to share data and information needed to coordinate maneuvers and trajectories. Extended sensors are used for exchanging raw/processed sensor data or live video. Remote driving allows to remotely drive or remotely control a private or public vehicle ([B7]). These requirements cannot be supported by Radio Access Technologies (RATs), neither IEEE 802.11 variants nor LTE and C-V2X Releases 14 and 15. This has led to the need to create a more performing type of communication, which includes not only improvements in network performance, but which includes a global end-to-end approach, in which the network and use case requirements can be customized to make the communication experience dependent on the requests of the end user.

2.2 C-RAN and VEC

A mobile network is generally composed of a radio unit, a transport segment and a core. The RAN access network allows data to be exchanged with users via Base stations (BS). Each BS performs the radio functions and is divided into two blocks: a remote radio unit (RRU) and a baseband unit (BBU). The RRU contains components for the frequency up / down conversion, power amplification and filtering of the radio signals. Each RRU is connected directly to the antennas and BBUs. BBUs perform physical and upper layer functionality and directly interface with the transport segment, commonly called backhaul. This transport segment therefore has the task of transporting data from the BBU to the core and viceversa. Initially, the BBUs were located in close proximity to the RRUs, and this high density of BBUs in overpopulated areas led to problems such as interference. In order to overcome this problem, it was decided to centralize the processing functions in the baseband, by inserting another transport segment, the fronthaul ([B8]). The centralized radio access network (C-RAN) is a network architecture that leads to the centralization of baseband processing functions in a few stations, called BBU hotels. This centralization allows to decrease the latency and improve the efficiency and maintenance of the BBUs, which will be found in selected areas and no longer close to the RRUs.
VEC is the integration of the MEC with vehicular networks[B9]. The

goal is to bring communication, processing and storage resources closer to vehicular users. The VEC therefore plays a potentially fundamental role in addressing the exponentially growing needs of low delay and high reliability devices. To diversify the VEC from the MEC are the complicated communication characteristics, due to the channel environment that varies rapidly over time, and the rapid mobility of vehicular users, which leads to frequent and more dynamic topology changes. Using the VEC architecture, the user can request the necessary content from the caching nodes, without having to access the main network. Doing so reduces end-to-end latency, increasing the efficiency of network bandwidth usage.
The VEC architecture is divided into 3 levels: users level, edge level and cloud level. In the first level it is possible to find the terminals, represented by the vehicles that can exchange information with each other or with road infrastructure through different protocols, such as the ones reported in references [B10, B11]. At the Edge level there is the Road-Side Units (RSU), responsible for receiving the information sent by the vehicles, processing this information received and uploading this information to the cloud. Compared to the edge level, the cloud level has components with greater processing and storage capacity, and allows to cover a larger area. The cloud paradigm can provide global management and centralized control.The emergence of VEC allows for different types of applications, such as road safety, traffic control, the possibility of having a real-time navigation system, low latency services and high processing capacity ([B12]).
By combining the concept of C-RAN with a VEC infrastructure, it is possible to bring services closer to vehicles and support applications that require low latency and high reliability.

2.3 5G cellular networks

5G represents the fifth generation of cellular technology. It is designed to increase speed, reduce latency and improve the flexibility of wireless services. While previous generations of cellular technology (such as 4G LTE) focused on ensuring connectivity, 5G takes connectivity to the next

level by delivering connection experiences that range from the cloud to customers. 5G networks are virtualized and software-based, and take advantage of cloud technologies.
In order to meet the new requirements, 5G takes advantage of new solutions such as functional split and network slicing. The first allows the more efficient distribution of the functionalities between the baseband and radio frequency module, trying on the one hand to reduce the bandwidth requirements (bringing them to values close to the transported capacity), and latency (reaching values of the order of milliseconds), on the other hand trying to keep LTE Advanced performances similar to those obtainable in traditional Cloud RAN architectures with CPRI ([B13]) as much as possible. The second allows to virtualize the network, exploiting the same physical network and managing the network resources in order to offer a "slice" of the network based on the characteristics required by the use case.

2.3.1 Network slicing

Network slicing was born as tool for managing a network, designed keeping in mind the idea of satisfying the requests of multiple categories of users at the same time. On top of a shared physical architecture, a slice-based network is modeled as a series of logical networks. In order to implement it, the combined use of the Software Defined Networking (SDN) and Network Functions Virtualization (NFV) paradigms will be required. SDN contributes to the control of the various devices involved by disassociating packet forwarding and routing processes, centralizing the intelligence of a network, made up of various controllers, on a detached plane. The NFV instead offers the possibility to virtualize network services, such as routers or firewalls, normally performed on hardware.

The network is divided into multiple networks, providing an end-to-end connection whose particularity is the possibility of being adapted to accommodate a wide range of functions. Based on the requirements, three macro categories have been defined to identify the use cases. The following service classes are thus defined:

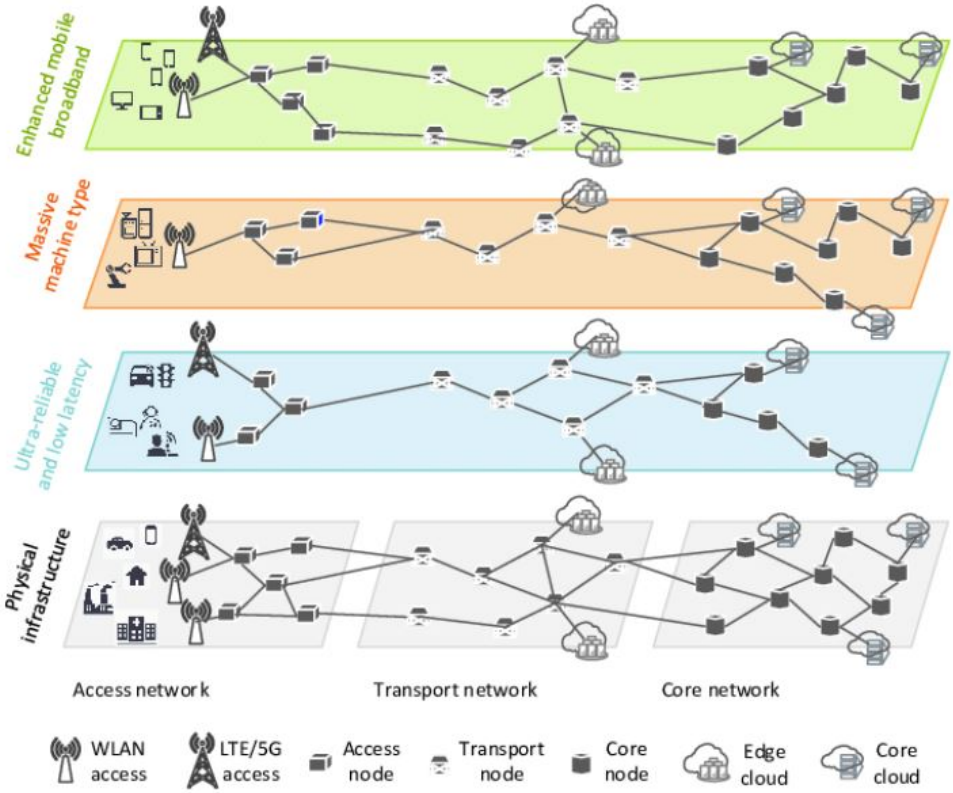


Figure 2.1: Network slicing concept.

- **enhanced Mobile BroadBand (eMBB)**: the aim is to allow high effective data transmission speed for the user and greater network capacity, which will be able to support the transmission of video streams to a greater number of active users simultaneously, even when they are concentrated in limited areas.
- **Ultra Reliable Low Latency Communications(URLLC)**: this class of service will represent the true technological innovation of 5G, opening the way to potential new applications, for example in the automotive and industrial sector, where real-time response times are needed with orders of magnitude less than possible with today's LTE networks.
- **massive Machine Type Communications (mMTC)**: it is basically the revisiting of the low speed and low consumption connection modes already standardized in the most recent LTE networks,

such as NB-IoT (NarrowBand IoT).

In order to guarantee the desired performance, these classes of service must be at the same time optimized and well connected network. With the increase of the antennas in the area, functions distributed in the network must therefore be associated which will allow the use of the same physical network, with the achievement of the different KPIs required by the different services (fig. 2.1). This new architecture is based on a functional separation of the processing capacity of mobile radio sites (baseband): the so-called functional split.

2.3.2 Functional split

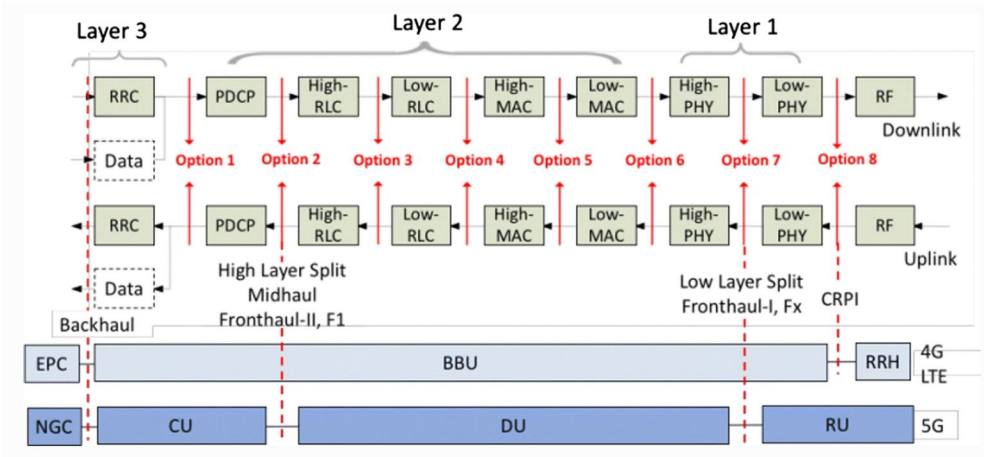


Figure 2.2: Split options.

The 5G network is a disaggregated network topology, where it is possible to separate the real-time and non-real-time functions: the first most critical are managed on the site where the antennas closest to the user are present, the second are centralized and virtualized at higher network levels. The physical network can therefore be divided into several sub-parts, which must be optimized according to use. The connection between the two parts of the split takes place through an efficient ethernet fronthauling interface. The network chain is then divided into several parts: Radio Unit (RU), Distributed Unit (DU), Centralized Unit (CU), Core / Cloud.

With the functional split it is possible to choose the functions that each block will have to perform. By associating the concept of split with virtualization, it is easy to imagine networks that are based on the same physical network but, thanks to a different subdivision of roles and a different location of the blocks, they allow to reach different requirements, according to the user's needs.

Table 2.2: 3GPP splits bitrate requirements in Gbps for a sample antenna configuration, options from 1 to 5

Option	**1**	**2**	**3**	**4**	**5**
DL (Gbps)	4	4	4	4	4
UL (Gbps)	3	3	3	3	3
Latency (ms)	10	1.5-1	1.5	0.1	0.1-0.2

Table 2.3: 3GPP splits bitrate requirements in Gbps for a sample antenna configuration, options from 6 to 8

Option	**6**	**7a**	**7b**	**7c**	**8**
DL (Gbps)	5.6	13.6-21.6	53.8-86.1	53.8-86.1	157.3
UL (Gbps)	5.6	13.6-21.6	53.8-86.1	53.8-86.1	157.3
Latency (ms)	0.25	0.25	0.25	0.25	0.25

As shown in the figure 2.2, each option corresponds to different functionalities, which can be distributed in the RU, DU, CU. Each block corresponds to a level: physical level, data level and network level. Depending on the option chosen, the bandwidth and latency requirements from the network are shown in tabs 2.2 and 2.3. Each split requires different requirements at the network, and offers centralization solutions that can vary according to the needs of the network.

Chapter 3

Reliability support for vehicular networks based on functional split

As described in the chapter 2, classes of service are provided in the 5G network that allow to support use cases that place specific latency and bandwidth requirements in the different segments of the end-to-end transport network. These use cases certainly include those dedicated to the automotive sector: low latency is an indispensable requirement to be able to guarantee safety in the vehicular world, and among the use cases we can find interactive applications for automotive safety, control of unmanned aerial vehicles and intelligent health emergencies. The class of service foreseen to reach these requirements is the Ultra Reliable Low Latency communications (URLLc) service, a class that allows to reach a maximum latency of up to 1 ms and guaranteeing high reliability.
The 5G network design will make full use of Network Function Virtualization (NFV) combined with the Software Defined Network (SDN) paradigm, to ensure unprecedented network flexibility and reconfigurability. To improve the efficiency of the network and the flexibility of the optical aggregation network segment, fronthaul and midhaul sections have been created in the optical aggregation network. This approach allows for dynamic resource utilization based on information multiplexing based on statistical packets, resulting in greater scalability and more relaxed centralization, which allows for more relaxed bandwidth and la-

tency constraints than C-RAN.
The main problem is how to divide the required functionalities into nodes in relation to the bandwidth available on the links and the processing capacity in the nodes themselves, with the aim of obtaining complete coverage of the packet metro network. Each node is associated with a virtual function hotel that performs the associated functions in relation to the subdivision option chosen.

In the specific case of URLLc, it is necessary to meet extremely stringent requirements in terms of latency and reliability, optimally associating functions to nodes. For this purpose it is necessary to provide additional backup resources ready to use in the event of a connection or failure of the hotel, capable of satisfying the same quality of service as the primary ones. As a result, it is possible to perform a fast connectivity swap within the slice. This approach is known in the literature as asset protection and, according to previous classifications, different protection schemes can be applied for the resilience of the fast slice [B14]. Protection schemes can be dedicated paths (DPP) or shared paths (SPP). In DPP, backup resources are dedicated and therefore cannot be shared with any other protection path. Conversely, SPP allows the sharing of backup resources between protection locations. In order to allow the sharing of resources, it is necessary to respect some rules, based on the level of protection you want. In the case of a single node or link failure, in order to share network resources it is necessary that the main paths (or primary paths) are totally separate. If the primary paths share part of the path, part of the backup path resources can be shared. SPP mechanisms are expected to require fewer additional resources than DPP, but are usually more complicated to use.
In this chapter, a network design optimization methodology to implement URLLC service in a metro area will be shown. A novel optimization algorithms based on Integer Linear Programming (ILP) are defined to solve dedicated and shared path protection problems, in the presence of single failure with the aim to minimize the number of active nodes, by adopting the functional splitting options defined by 3GPP.

3.1 Problem formulation: latency constrained slicing

The functions composing the layers of the mobile network protocol stack can be split into multiple nodes and performed in sequence, forming a chain of functions. Function chaining has been investigated before, where service end points are known and only end to end bandwidth and latency constraints are applied [B15]. However, when providing a service, the requirements of the baseband processing must be satisfied along with the one of the end to end service, usually performed in the cloud. The bandwidth requirement to carry different functions of the protocol stack are shown in Table 3.1 for a sample antenna configuration [B16]. An example of a possible end to end URLLC service slice is presented in Fig. 3.1. The reliability required by this class of services implies the allocation of primary and backup path resources for each chain in the slice.

Table 3.1: Link capacity requirements for different 3GPP split options [B16].

Layers	Split option	Bandwidth [Gbps]
L1	Opt.8	2.4
L2	Opt.6	0.152
L3	Opt.2	0.151
Core	Opt.1	0.150
Cloud	-	0.150

The formulation of the BBU hotel location problem with resiliency is as follows:

- **Given**: a set of nodes and related resources, which are candidates as hotels to host baseband, core and cloud functions, properly connected through a set of links.
- **To find**: a suitable functions placement, such that the number of

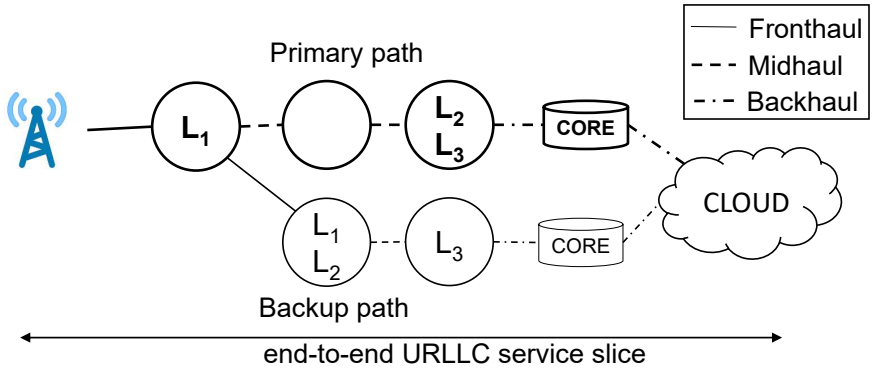


Figure 3.1: Example of functional split for URLLC service slice.

active nodes (i.e., nodes hosting any function) is reduced to a minimum while reliability against single link or hotel failure is provided.

– **To ensure**: that each antenna is connected to URLLC service through a set of properly ordered functions forming a chain running in active nodes, so that the maximum allowed distance between the antenna and the cloud is not exceeded, the maximum bandwidth available on each link is not exceeded, and the available computational resources in each node are not exceeded.

Let us consider a network characterized by a set of nodes N, each with capacity ρ_i, interconnected by links, captured by matrix $\gamma_{i,j}$, with bandwidth $\lambda_{i,j}$ and introducing a delay $\tau_{i,j}$, with $i, j \in N$. Each node is considered to be a source $s \in S$, with a set of antennas to be connected, through a chain of functions, to the service in the cloud. Let us consider an ordered set $T = \{t_1, .., t_q, .., t_k\}$ of k transport segments, with cardinality $|T| = k$ equal to the number of functions to be executed. Each transport segment corresponds to a couple of VNFs, one originating the transport flow and one terminating it. For instance, a generic transport segment t_q is originated by one VNF (v_{q-1}) and requires a VNF (v_q) performing the functions required to elaborate its traffic, and originates the traffic towards the next transport segment t_{q+1}. A binary variable $x^n_{i,t_q,s}$ is introduced to model the assignment of sources to nodes performing related functions. When $x^n_{i,t_q,s}$ is equal to 1, v_q (VNF function terminating transport segment t_q) is performed at node i for source s, and requires the activation of node i, modeled by the binary variable z_i, for primary

and backup paths $n \in P = \{p, b\}$. Each transport segment t_q produces a primary (p) and backup (b) flow of data for each source s through the links, captured by the binary variable $w^n_{i,j,t_q,s}$, with a certain bitrate β_{t_q} and subject to latency requirements δ_{t_q}. Each VNF related to a transport segment t_q needs computational resources μ_{t_q}. The notation used in the two strategies is reported in Tables 3.2 and 3.3 .

Table 3.2: List of parameters of the ILP and corresponding definitions.

Parameter	**Definition**
T	set of transport segments.
S	set of source nodes.
N	set of network nodes, candidates to host virtual functions.
P	set of paths. P = {p,b} (p for primary, b for backup).
$\alpha_z, \alpha_c, \alpha_f$	tuning parameters for SPP objective function.
β_{t_q}	bandwidth requirement for transport segment $t_q \in T$.
δ_{t_q}	latency requirement for transport segment $t_q \in T$.
$\gamma_{i,j}$	1 if exists a link between nodes $i \in N$ and $j \in N$ in the physical network; 0 otherwise.
μ_{t_q}	capacity required to execute virtual function terminating transport segment $t_q \in T$.
ρ_i	computational resources available at node $i \in N$.
$\lambda_{i,j}$	available bandwidth over the link connecting nodes $i \in N$ and $j \in N$.
$\tau_{i,j}$	delay introduced by the link connecting nodes $i \in N$ and $j \in N$.
M	a large number.

3.1.1 Dedicated Path Protection model

The model for dedicated protection is as follows:

Table 3.3: List of variables of the ILP and corresponding definitions.

Parameter	**Definition**
$x^n_{i,t_q,s}$	1 if node $i \in N$ is performing VNF terminating transport segment $t_q \in T$ for source $s \in S$ for path $n \in P$; 0 otherwise.
$w^n_{i,j,t_q,s}$	1 if link connecting nodes $i \in N$ and $j \in N$ is carrying transport traffic $t_q \in T$ originated at source $s \in S$ for path $n \in P$; 0 otherwise.
z_i	1 if node $i \in N$ is selected to host at least one VNF; 0 otherwise.
c_i	computational capacity required at node $i \in N$ for backup purposes.
$f_{i,j}$	required bandwidth over the link $(i,j) \in N$ for backup purposes.
$y_{i,j,t_q,s}$	1 if the nodes $i \in N$ and $j \in N$ are performing VNF terminating transport segment $t_q \in T$ as primary and backup, respectively, for the source $s \in S$; 0 otherwise.
$l_{i,j,k,m,t_q,s}$	1 if the links $(i,j) \in N$ and $(k,m) \in N$ are used to transport data of segment t_q for primary and backup paths, respectively, for the source $s \in S$; 0 otherwise.
$d_{i,k,m,t_q,s}$	1 if the link $(k,m) \in N$ is used to transport data of segment t_q for backup and the primary path passes through node $i \in N$ for the source $s \in S$; 0 otherwise.

Objective function:

$$Minimize \sum_{i \in N} z_i \tag{3.1}$$

Constraints:

$$\sum_{i \in N} x^n_{i,t_q,s} = 1, \quad \forall t_q \in T, s \in S, n \in P \tag{3.2}$$

$$\sum_{n \in P} \sum_{t_q \in T} \sum_{s \in S} x^n_{i,t_q,s} \leq M \cdot z_i, \quad \forall i \in N \tag{3.3}$$

$$\sum_{n\in P}\sum_{t_q\in T}\sum_{s\in S} x^n_{i,t_q,s}\cdot\mu_{t_q}\leq\rho_i,\quad\forall i\in N \tag{3.4}$$

$$\sum_{n\in P}\sum_{t_q\in T}\sum_{s\in S} w^n_{i,j,t_q,s}\cdot\beta_{t_q}\leq\lambda_{i,j},\quad\forall i,j\in N \tag{3.5}$$

$$\sum_{t_c=1}^{t_q}\sum_{i\in N}\sum_{j\in N} w^n_{i,j,t_c,s}\cdot\tau_{i,j}\leq\delta_{t_q},\quad\forall t_q,t_c\in T,s\in S,n\in P \tag{3.6}$$

$$w^p_{i,j,t_q,s}+w^b_{i,j,t_m,s}\leq\gamma_{i,j},\quad\forall t_q,t_m\in T,s\in S,i,j\in N \tag{3.7}$$

$$\sum_{j\in N} w^n_{i,j,t_q,s}\leq 1,\quad\forall i\in N,t_q\in T,s\in S,n\in P \tag{3.8}$$

$$w^n_{i,i,t_q,s}\leq x^n_{i,t_q,s},\quad\forall i\in N,t_q\in T,s\in S,n\in P \tag{3.9}$$

$$x^p_{i,t_m,s}+x^b_{i,t_q,s}\leq 1,\quad\forall i\in N,t_q,t_m\in T,s\in S \tag{3.10}$$

If $t_q=t_1$:

$$\sum_{j\in N} w^n_{s,j,t_q,s}=1,\quad\forall s\in S,n\in P \tag{3.11}$$

$$\begin{gathered}\sum_{j\in N} w^n_{j,i,t_q,s}-\sum_{j\in N} w^n_{i,j,t_q,s}=x^n_{i,t_q,s},\\ \forall i\in N,i\neq s,t_q\in T,s\in S,n\in P\end{gathered} \tag{3.12}$$

If $t_q\neq t_1$:

$$\sum_{j\in N} w^n_{i,j,t_q,s}\geq x^n_{i,t_{q-1},s},\quad\forall i\in N,s\in S,n\in P \tag{3.13}$$

$$\begin{gathered}\sum_{j\in N} w^n_{j,i,t_q,s}-\sum_{j\in N} w^n_{i,j,t_q,s}+x^n_{i,t_{q-1},s}=x^n_{i,t_q,s},\\ \forall i\in N,t_q\in T,s\in S,n\in P\end{gathered} \tag{3.14}$$

Constraint (3.2) ensures that only one node is active for each VNF and source along the whole path. Constraint (3.3) selects the active nodes (i.e., nodes that host at least one VNF). Constraint (3.4) ensures that

computational resources required at node i to perform all VNFs from all sources and all the path are not exceeded. Constraint (3.5) guarantees that the bandwidth required over each link does not exceed the maximum link capacity for that link. Constraint (3.6) limits the delay of each path transport segment. Constraint (3.7) allows routing only over link of the physical topology. Also ensures that the links used for different path are different.

Function chaining is modeled as follows. Constraint (3.8) limits the sum of outgoing paths in each node, for each source and transport link. Constraint (3.9) forbids unnecessary loops. Constraint (3.10) ensures that the nodes used for primary and backup are different, for all the transport segment.

For the first transport segment (t_1), constraint (3.11) ensures that there is one outgoing flow for each source s while constraint (3.12) represents the flow conservation towards the ending VNF for transport segment t_1. For the subsequent transport segments ($\{t_2, .., t_k\}$), constraint (3.13) ensures that there is a transport flow starting from the node (i) performing the previous transport function ($x_{i,t_{q-1},s}$) for each source. Constraint (3.14) represents the flow conservation of each transport segment t_q.

3.1.2 Shared Path Protection model

In the SPP model, four additional variables have been introduced: c_i and $f_{i,j}$ which allow the reduction of computational resources and bandwidth reserved for backup, and $y_{i,j,t_q,s}$ and $l_{i,j,k,m,t_q,s}$ to find the pairs of nodes and links of the primary and backup. A new objective function is also introduced.

Objective function:

$$Minimize\ \alpha_z \cdot \sum_{i \in N} z_i + \alpha_c \cdot \sum_{i \in N} c_i + \alpha_f \cdot \sum_{i \in N} \sum_{j \in N} f_{i,j} \tag{3.15}$$

The constraints (3.4) and (3.5) have been replaced by:
Additional constraints:

$$y_{i,j,t_q,s} \geq x^p_{i,t_q,s} + x^b_{j,t_q,s} - 1 \quad \forall i,j \in N, s \in S, t_q \in T \tag{3.16}$$

$$c_j \geq \sum_{t_q \in T} \sum_{s \in S} y_{i,j,t_q,s} \cdot \mu_{t_q} \quad \forall i,j \in N \tag{3.17}$$

$$\sum_{t_q \in T} \sum_{s \in S} x^p_{i,t_q,s} \cdot \mu_{t_q} + c_i \leq \rho_i, \quad \forall i \in N \tag{3.18}$$

$$\begin{gathered} l_{i,j,k,m,t_q,s} \geq w^p_{i,j,t_q,s} + w^b_{k,m,t_q,s} - 1 \\ \forall i,j,k,m \in N, s \in S, t_q \in T \end{gathered} \tag{3.19}$$

$$f_{k,m} \geq \sum_{t_q \in T} \sum_{s \in S} l_{i,j,k,m,t_q,s} \cdot \beta_{t_q} \quad \forall i,j,k,m \in N \tag{3.20}$$

$$\begin{gathered} d_{i,k,m,t_q,s} \geq w^b_{k,m,t_q,s} + \frac{\sum_{t_q \in T} x^p_{i,t_q,s}}{M} - 1, \\ \forall i,k,m \in N, s \in S, t_q \in T \end{gathered} \tag{3.21}$$

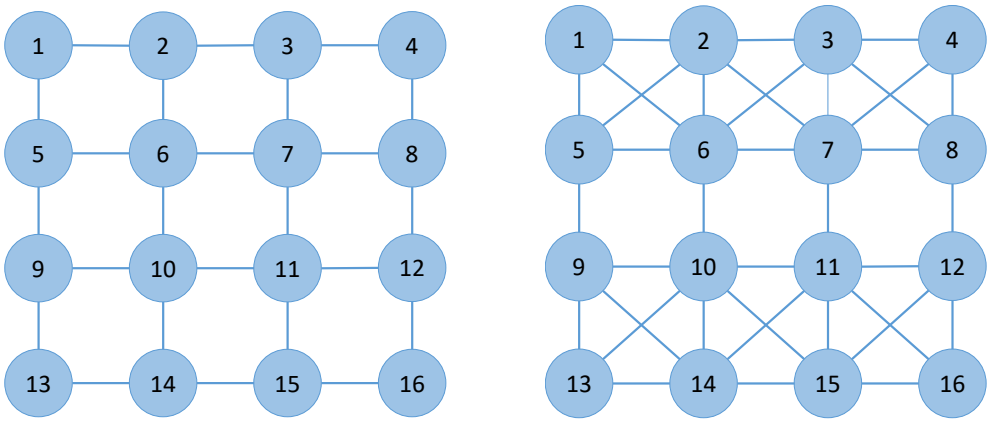
$$f_{k,m} \geq \sum_{t_q \in T} \sum_{s \in S} d_{i,k,m,t_q,s} \cdot \beta_{t_q}, \quad \forall i,k,m \in N \tag{3.22}$$

$$\sum_{t_q \in T} \sum_{s \in S} w^p_{i,j,t_q,s} \cdot \beta_{t_q} + f_{i,j} \leq \lambda_{i,j}, \quad \forall i,j \in N \tag{3.23}$$

Constraint (3.16) finds the primary (i) and backup (j) nodes performing the different VNFs for each source. Constraint (3.17) ensures that the capacity reserved for the backup node j is greater than or equal to the capacity required to perform primary functions at node i, to ensure reliability against single primary hotel failure. Constraint (3.18) ensures that computational resources required at node i to perform all VNFs from all sources and all the paths are not exceeded. Constraint (3.19) finds the primary link (i,j) and the backup link (k,m) carrying traffic of each transport for each source. Constraint (3.20) ensures that the bandwidth reserved for the backup path is greater than or equal to the bandwidth required in case of a single primary link failure. Constraint (3.21) finds the sources affected by a BBU hotel failure in i that are sharing the backup link (k,m) while constraint (3.22) counts the bandwidth required over link k,m in the case of hotel i failure. Constraint (3.23)

guarantees that the bandwidth required over each link does not exceed the maximum link capacity for that link.

3.2 Comparison of protection schemes and numerical results



(a) Network A. (b) Network B.

Figure 3.2: Reference 16 node networks A and B with different connectivity.

In this section, the results obtained by running the two algorithms using the CPLEX commercial tool [B17] are reported. The DPP strategy is firstly evaluated referring to the 16 node networks represented in Fig. 3.2. Each network node is connected to 10 antennas, collecting traffic from the radio section. The available bandwidth on each link is set to 40 Gbps (in each direction). In addition, each node is equipped with processing units (PUs) according to traffic generated at each layer of the functional splitting as shown in Table 3.1. In particular, 0.5, 0.3, 0.2, 0.1, 0.1 PUs are assumed as requirements for L1, L2, L3, core and cloud virtual functions, respectively [B20, B18, B19]. The length of each link or, equivalently, each hop is assumed to be 1 km, which results in a delay $\tau = 5\mu s$. It should be noted that, given the limited size of the scenario, the latency constraints of each transport link are always satisfied. However, to satisfy the tight service requirements imposed by URLLC applications, all the nodes are allowed to host edge core and

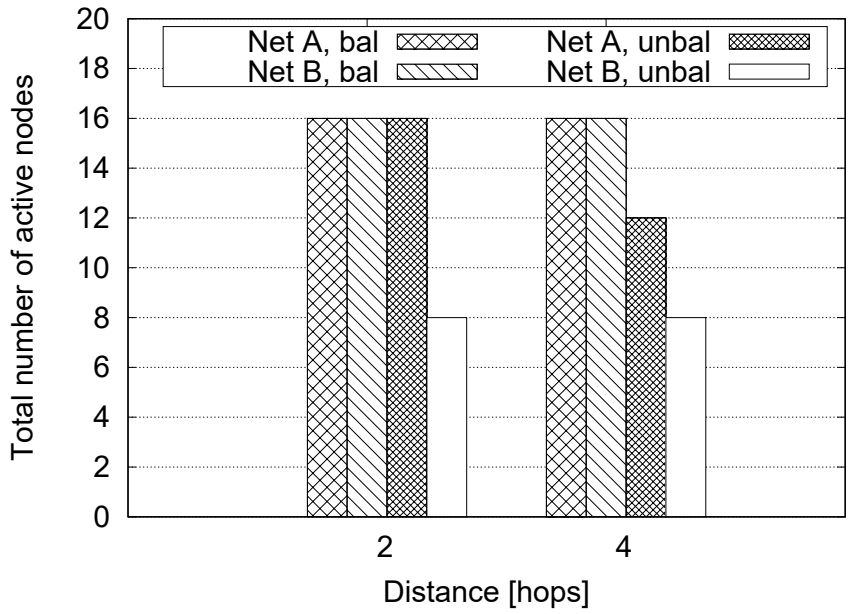


Figure 3.3: Active nodes for the networks A and B in the balanced and unbalanced cases with 2 and 4 hops as distance constraint.

cloud functions. The SPP approach is evaluated then, assuming a 6 node network, not to incur in out of memory exception due to the higher complexity of the SPP algorithm.

3.2.1 Dedicated Path Protection evaluation

The two networks A and B used to evaluate the DPP algorithm are presented in Fig. 3.2.

The two networks differ for the connectivity represented by a different number of links. Two cases have been considered in the following. The case in which all the nodes candidate to host an hotel are equal to 25 PUs is referred to as balanced (bal). The unbalanced (unbal) case, instead, has nodes 6,7,10,11 with infinite capacity in terms of PUs, thus emulating centralized data centers, while all the other nodes are equipped with 10 PUs. The number of active nodes (i.e., nodes hosting baseband, core or cloud functions) obtained with the DPP model in the networks A and B in the balanced and unbalanced case under different hop constraints is reported in Fig. 3.3. The balanced case always requires the activation of all the nodes, as a consequence of the limitation of node resources,

regardless the number of hops and network connectivity. Conversely, the unbalanced case shows a reduction in the number of active nodes. The effects of an increased network connectivity are evident, with only 8 active nodes required for both 2 and 4 hops. The network A allows a reduction of 4 nodes when moving from 2 to 4 hops, thanks to high capacity nodes that perform multiple functions in few nodes, while in network B no additional node reduction is possible due to the limited resources over links connecting high capacity nodes.

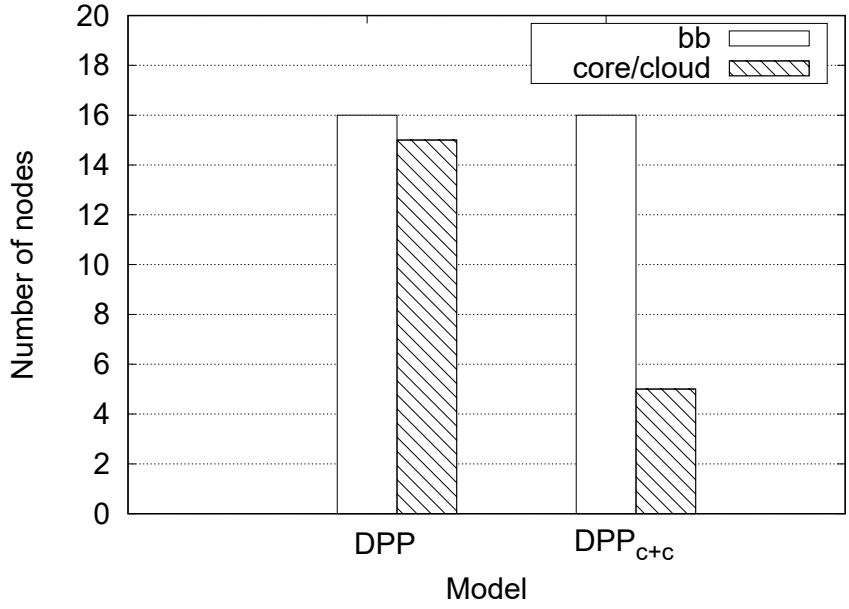


Figure 3.4: Number of nodes performing at least one baseband (bb) processing (either L1, L2 or L3), core and cloud functions for the balanced case with 2 hops constraints for DPP and modified DPP_{c+c} models in network A.

To facilitate the deployment of edge core and cloud functions, minimization of active nodes performing those functions can be added to the objective function of the DPP model. A new term is added to (3.1) with a lower priority, so that the primal objective remains the same (i.e., the minimization of the total active nodes). This case is referred to as DPP_{c+c}. Figure 3.4 shows the number of nodes performing at least one baseband processing (either L1, L2 or L3) and core/cloud functions for the balanced case with 2 hops constraints in the traditional DPP and modified DPP_{c+c} formulation. While the number of nodes performing

baseband processing is the same for the two cases, the nodes performing core and cloud functions in DPP_{c+c} is considerably lower than the one of DPP, thus simplifying the deployment of these functions from a network operator and/or cloud provider point of view.

Figure 3.5 depicts the link usage of DPP in the network A, under 2 and 4 hop constraints, for both balanced and unbalanced cases. In the figure the links are sorted in increasing order of usage for each curve. Depending on the specific curve, links show different usage, which indicates potential statistical multiplexing gain when multiple slices are embedded on the same network. Some links exhibit a very low usage or even no usage, especially with 2 hop constraint. Many links needs 24 Gbps or slightly higher due to the capacity required for option 8 with 10 antennas (see Table 3.1).

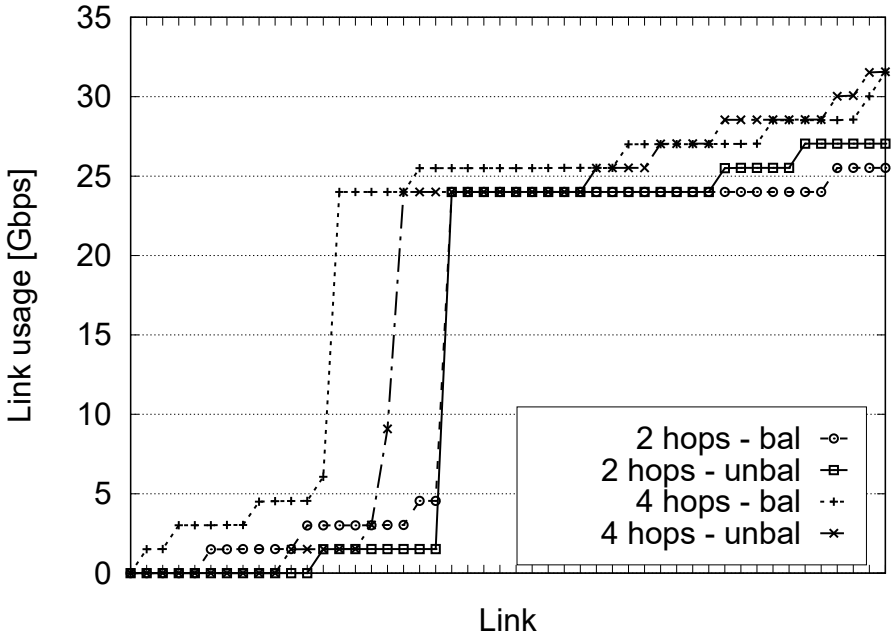


Figure 3.5: Link usage of DPP in the network A, balanced and unbalanced, under different hop constraints. Links are sorted from the lowest to the highest usage one.

3.2.2 Shared Path Protection evaluation

A 6 node network is considered to compare DPP and SPP (Figs. 3.6 and 3.7). All nodes are connected to 10 antennas. In the unbalanced case, nodes 2 and 5 have unlimited resources, while all the other node capabilities are limited to 10 PUs. The tuning parameters of the objective function are set as $\alpha_z >> \alpha_c = \alpha_f$ to prioritize the minimization of the overall active nodes.

Table 3.4 reports the number of active nodes, capacity and node savings for the 6 node networks in the balanced and unbalanced cases under 2 and 3 hop constraints. The SPP is capable of reducing the number of active nodes in the balanced case by 33.3%. In addition, the SPP approach allows to share backup node resources among antennas assigned to different primary paths, leading up to 66.6% and 27.8% node capacity savings in the balanced and unbalanced cases, respectively.

Table 3.4: Active nodes, capacity and node savings for 6 node networks in the balanced and unbalanced cases under 2 and 3 hop constraints.

	Active nodes		Saved	
Network	DPP	SPP	Capacity	Nodes
2 hops - bal	6	4	66.6%	33.3%
3 hops - bal	6	4	66.6%	33.3%
2 hops - unbal	4	4	23.6%	0%
3 hops - unbal	4	4	27.8%	0%

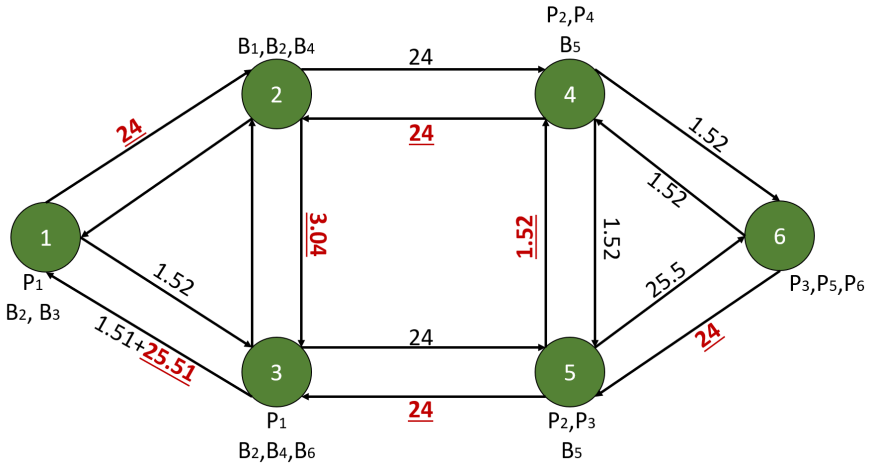


Figure 3.6: Outcome of the DPP model in the balanced case for 6 node network and 2 hops as distance constraint. The bandwidth values are in Gbps, red and underlined for the backup path. Dark green color for active nodes.

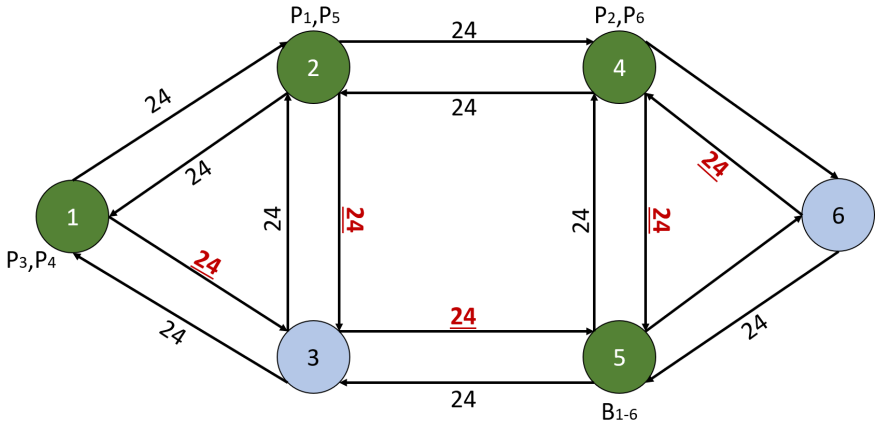


Figure 3.7: Outcome of the SPP model in the balanced case for 6 node network and 2 hops as distance constraint. The bandwidth values are in Gbps, red and underlined for the backup path. Dark green color for active nodes.

Figures 3.6 and 3.7 depict the outcome of DPP and SPP models, respectively, in the balanced case for 6 node network and 2 hops as distance constraint. The figures report the active nodes, that are the nodes

performing primary and/or backup functions, eventually split in multiple nodes. The figures also report the link usage for both primary and backup (in underlined, red). In the case of SPP, the number of active nodes is lower than in the DPP case, thanks to the sharing of the backup paths. For instance, in the SPP case nodes 2, and 3 are able to reach node 5, for backup purposes, by sharing link 3-5. In the DPP case instead, they cannot reach node 5 due to the dedicated resource allocation and limited bandwidth over the links, thus requiring to activate additional nodes.

Chapter 4

Multiple service class 5G vehicular network

In the previous chapter, we highlighted the network performance in the case where the network is composed of only one type of slice, with the same requirements for each slice. It has been highlighted how different protection schemes can act in the network. In this chapter we will analyze the problem of multiple slices. In particular, the concept of differentiated path protection will be applied.

4.1 Problem formulation: latency constrained and high capacity slicing

The differentiated path protection concept is a means to apply protection schemes in relation to service needs. This approach has been considered in the past to choose the proper protection scheme in relation to connection needs [B21]. Similarly, this approach can be considered in the design of reliable 5G network slices with eMBB or URLLC services when critical contexts need to be deployed [B22]. As far as the URLLC class a dedicated path protection scheme is adopted to ensure high availability and fast provisioning of alternate resources in case of failure. As far as eMBB, a shared path protection approach is adopted to save resources in this highly demanding service class, while preserving acceptable data service continuity. To optimize the resource assignment for the two classes

of service, joint and disjoint assignment approaches can be adopted. The joint approach consists in assigning resources to the different classes in a target area with a unique optimization procedure. This approach is expected to achieve the best results but it is potentially not scalable. For the sake of scalability, the assignment can be performed sequentially for the different classes.

The differentiated slice protection optimization problem finds the nodes which host needs to be activated with related processing, link and bandwidth resources. So the definition of the problem is as follows:

- **Given**: a set of nodes with given available computational resources (PU), properly connected through a set of links with given available bandwidth.
- **To find**: the best position of network functionality according to [B23], minimizing the number of active nodes and links, providing protection in the presence of link or node failure.
- **To ensure**: that each node must guarantee the services URLLC and eMBB, so that the latency is less than required, the bandwidth used is less than that available in the links, and the computational resources of the nodes are not exceed.

Differently from previous chapter, two different classes are here considered, URLLC and eMBB, which require to modify objectives and constraints to guarantee the KPIs of both classes. In particular, the minimization of the active links is introduced which avoids the need to minimize the bandwidth and shared computational resources, so that better scalability can be achieved.

4.2 ILP model

Let us consider a network characterized by a set of nodes N, each with capacity ρ_i, interconnected by links, captured by matrix $\gamma_{i,j}$, with bandwidth $\lambda_{i,j}$ and introducing a delay $\tau_{i,j}$, with $i, j \in N$. Each node is considered to be a source $s \in S$, with a set of antennas dedicated to a each class of service cs to be connected, through a chain of functions

Table 4.1: List of parameters of the ILP and corresponding definitions.

Parameter	**Definition**
T	set of transport segments.
S	set of source nodes.
N	set of network nodes, candidates to host virtual functions.
P	set of paths. P = {p,b} (p for primary, b for backup).
CS	set of classes of service. CS = {1,2} (1 for SPP protection, 2 for DPP protection)
α_z, α_a	tuning parameters for the objective function.
$\beta_{t_q,cs}$	bandwidth requirement for transport segment $t_q \in T$ for different classes of service $cs \in CS$.
$\delta_{t_q,cs}$	latency requirement for transport segment $t_q \in T$ for different classes of service $cs \in CS$.
$\gamma_{i,j}$	1 if exists a link between nodes $i \in N$ and $j \in N$ in the physical network; 0 otherwise.
$\mu_{t_q,cs}$	capacity required to execute virtual function terminating transport segment $t_q \in T$ for different classes of service $cs \in CS$.
ρ_i	computational resources available at node $i \in N$.
$\lambda_{i,j}$	available bandwidth over the link connecting nodes $i \in N$ and $j \in N$.
$\tau_{i,j}$	delay introduced by the link connecting nodes $i \in N$ and $j \in N$.
M	a large number.

(baseband and core), to the service in the cloud. Let us consider an ordered set $T = \{t_1, .., t_q, .., t_k\}$ of k transport segments, with cardinality $|T| = k$ equal to the number of functions to be executed. Each transport segment corresponds to a couple of VNFs, one originating the transport flow and one terminating it. For instance, a generic transport segment t_q is originated by one VNF (v_{q-1}) and requires a VNF (v_q) performing the functions required to elaborate its traffic, and originates the traffic towards the next transport segment t_{q+1}. Each VNF related to a trans-

Table 4.2: List of variables of the ILP and corresponding definitions.

Variable	**Definition**
$x^n_{i,t_q,s,cs}$	1 if node $i \in N$ is performing VNF terminating transport segment $t_q \in T$ for source $s \in S$ for path $n \in P$ for classes of service $cs \in CS$; 0 otherwise.
$w^n_{i,j,t_q,s,cs}$	1 if link connecting nodes $i \in N$ and $j \in N$ is carrying transport traffic $t_q \in T$ originated at source $s \in S$ for path $n \in P$ for classes of service $cs \in CS$; 0 otherwise.
z_i	1 if node $i \in N$ is selected to host at least one VNF; 0 otherwise.
c_i	computational capacity required at node $i \in N$ for backup purposes.
$f_{i,j}$	required bandwidth over the link $(i,j) \in N$ for backup purposes.
$y_{i,j,t_q,s,1}$	1 if the nodes $i \in N$ and $j \in N$ are performing VNF terminating transport segment $t_q \in T$ as primary and backup, respectively, for the source $s \in S$ for class of service $cs = 1 \in CS$; 0 otherwise.
$l_{i,j,k,m,t_q,s,1}$	1 if the links $(i,j) \in N$ and $(k,m) \in N$ are used to transport data of segment t_q for primary and backup paths, respectively, for the source $s \in S$ for class of service $cs = 1 \in CS$; 0 otherwise.
$d_{i,k,m,t_q,s,1}$	1 if the link $(k,m) \in N$ is used to transport data of segment t_q for backup and the primary path passes through node $i \in N$ for the source $s \in S$ for class of service $cs = 1 \in CS$; 0 otherwise.

port segment t_q needs computational resources $\mu_{t_q,cs}$. A binary variable $x^n_{i,t_q,s,cs}$ is introduced to model the assignment of sources to nodes performing related functions for each protection required by the class of service cs. When $x^n_{i,t_q,s,cs}$ is equal to 1, v_q (VNF function terminating transport segment t_q) is performed at node i for source s and protection of the classes cs, and requires the activation of node i, modeled by the binary variable z_i, for primary and backup paths $n \in P = \{p, b\}$. Each

transport segment t_q produces a primary (p) and backup (b) flow of data for each source s and for each service cs through the links, captured by the binary variable $w^n_{i,j,t_q,s,cs}$ with a given bit rate $\beta_{t_q,cs}$ and subjected to latency requirements $\delta_{t_q,cs}$, and requires the activation of link (i,j) modeled by the binary variable $a_(i,j)$. Both protection schemes have a primary and a backup path, to work in the event of a link or node failure. This means that primary and backup path cannot share the same nodes, the same links and the same path per source. While Dedicated Path Protection (DPP) reserves the same bandwidth and PU resources for backup as the primary, in the Shared Path Protection (SPP) scheme it is possible to share bandwidth and PU resources with other backup paths in the following cases:

- **PU sharing**: nodes must not use the same node for any primary virtual functionality;
- **Bandwidth sharing**: the links must not use the same link for any primary virtual functionality and the different backup links must not have the same primary path.

The following model allows the resolution of the problem:
Objective function:

$$Minimize\ \alpha_z \cdot \sum_{i \in N} z_i + \alpha_a \cdot \sum_{i \in N} \sum_{j \in N} a_{i,j} \tag{4.1}$$

Constraints:

$$\sum_{i \in N} x^n_{i,t_q,s,cs} = 1, \quad \forall t_q \in T, s \in S, n \in P, cs \in CS \tag{4.2}$$

$$\sum_{n \in P} \sum_{t_q \in T} \sum_{s \in S} \sum_{cs \in CS} x^n_{i,t_q,s,cs} \leq M \cdot z_i, \quad \forall i \in N \tag{4.3}$$

$$\sum_{n \in P} \sum_{t_q \in T} \sum_{s \in S} \sum_{cs \in CS} w^n_{i,j,t_q,s,cs} \leq M \cdot a_{i,j}, \quad \forall i,j \in N \tag{4.4}$$

$$\sum_{t_q=1}^{t_q} \sum_{i \in N} \sum_{j \in N} w^n_{i,j,t_e,s,cs} \cdot \tau_{i,j} \leq \delta_{t_q,cs}, \quad \forall t_q, t_e \in T, s \in S, n \in P, cs \in CS \tag{4.5}$$

$$w^p_{i,j,t_q,s,cs} + w^b_{i,j,t_e,cs} \leq \gamma_{i,j}, \quad \forall i,j \in N, t_q, t_e \in T, s \in S, cs \in CS \tag{4.6}$$

$$\sum_{j \in N} w^n_{i,j,t_q,s,cs} \leq 1, \quad \forall i \in N, t_q \in T, s \in S, cs \in CS, n \in P \tag{4.7}$$

$$w^n_{i,i,t_q,s,cs} \leq x^n_{i,t_q,s,cs}, \quad \forall i \in N, t_q \in T, s \in S, cs \in CSn \in P \tag{4.8}$$

$$x^b_{i,t_q,s,cs} + x^p_{i,t_e,s,cs} \leq 1, \quad \forall i \in N, t_q, t_e \in T, s \in S, cs \in CS \tag{4.9}$$

$$\sum_{n \in P} \sum_{t_q \in T} \sum_{s \in S} w^n_{i,j,t_q,s,2} \cdot \beta_{t_q,2} + + \sum_{t_q \in T} \sum_{s \in S} w^p_{i,j,t_q,s,1} \cdot \beta_{t_q,1} + f_{i,j} \leq \lambda_{i,j}, \quad \forall i,j \in N \tag{4.10}$$

$$\sum_{n \in P} \sum_{t_q \in T} \sum_{s \in S} x^n_{i,t_q,s,2} \cdot \mu_{t_q,2} + + \sum_{t_q \in T} \sum_{s \in S} x^p_{i,t_q,s,1} \cdot \mu_{t_q,1} + c_i \leq \rho_i, \quad \forall i \in N \tag{4.11}$$

$$y_{i,j,t_q,s,1} \geq x^p_{i,t_q,s,1} + x^b_{j,t_q,s,1} - 1 \quad \forall i,j \in N, s \in S, t_q \in T \tag{4.12}$$

$$c_j \geq \sum_{t_q \in T} \sum_{s \in S} y_{i,j,t_q,s,1} \cdot \mu_{t_q,1} \quad \forall i,j \in N \tag{4.13}$$

$$l_{i,j,k,m,t_q,s,1} \geq w^p_{i,j,t_q,s,1} + w^b_{k,m,t_q,s,1} - 1 \quad \forall i,j,k,m \in N, s \in S, t_q \in T \tag{4.14}$$

$$d_{i,k,m,t_q,s,1} \geq w^b_{k,m,t_q,s,1} + \frac{\sum_{t_q \in T} x^p_{i,t_q,s,1}}{M} - 1, \quad \forall i,k,m \in N, s \in S, t_q \in T \tag{4.15}$$

$$f_{k,m} \geq \sum_{t_q \in T} \sum_{s \in S} l_{i,j,k,m,t_q,s,1} \cdot \beta_{t_q,1}, \quad \forall i,j,k,m \in N \tag{4.16}$$

$$f_{k,m} \geq \sum_{t_q \in T} \sum_{s \in S} d_{i,k,m,t_q,s,1} \cdot \beta_{t_q,1}, \quad \forall i,k,m \in N \tag{4.17}$$

If $t_q = t_1$::

$$\sum_{j \in N} w^n_{s,j,t_q,s,cs} = 1, \quad \forall s \in S, cs \in CS, n \in P \tag{4.18}$$

$$\begin{gathered} \sum_{j \in N} w^n_{j,i,t_q,s,cs} - \sum_{j \in N} w^n_{i,j,t_q,s,cs} = x^n_{i,t_q,s,cs}, \\ \forall i \in N, i \neq s, t_q \in T, s \in S, cs \in CS, n \in P \end{gathered} \tag{4.19}$$

If $t_q \neq t_1$:

$$\sum_{j \in N} w^n_{i,j,t_q,s,cs} \geq x^n_{i,t_{q-1},s,cs}, \quad \forall i \in N, s \in S, cs \in CS, n \in P \tag{4.20}$$

$$\begin{gathered} \sum_{j \in N} w^n_{j,i,t_q,s,cs} - \sum_{j \in N} w^n_{i,j,t_q,s,cs} + x^n_{i,t_{q-1},s,cs} = x^n_{i,t_q,s,cs}, \\ \forall i \in N, t_q \in T, s \in S, cs \in CS, n \in P \end{gathered} \tag{4.21}$$

The objective function (3.1) minimizes the number of active nodes and active links in the network.
Constraint (4.2) ensures that only one node is active for each VNF and source along the whole path.
Constraint (4.3) selects the active nodes (i.e., nodes that host at least one VNF).
Constraint (4.4) selects the active links (i.e., links that host at least one VNF).
Constraint (4.5) limits the delay of each path transport segment.
Constraint (4.6) allows routing only over links of the physical topology.

Also ensures that the links used for different paths are different.
Constraint (4.7) limits the sum of outgoing paths in each node, for each source and transport link.
Constraint (4.8) forbids unnecessary loops.
Constraint (4.9) ensures that the nodes used for primary and backup are different, for all the transport segment.
Constraint (4.10) guarantees that the bandwidth required over each link does not exceed the maximum link capacity for that link.
Constraint (4.11) ensures that computational resources required at node i to perform all VNFs from all sources and all the paths are not exceeded.
Constraint (4.12) finds the primary (i) and backup (j) nodes performing the different VNFs for each source.
Constraint (4.13) ensures that the capacity reserved for the backup node j is greater than or equal to the capacity required to perform primary functions at node i, to ensure reliability against single primary hotel failure.
Constraint (4.14) finds the primary link (i, j) and the backup link (k, m) carrying traffic of each transport for each source.
Constraint (4.15) finds the sources affected by a BBU hotel failure in i that are sharing the backup link (k, m).
Constraint (4.16) ensures that the bandwidth reserved for the backup path is greater than or equal to the bandwidth required in case of a single primary link failure.
Constraint (4.17) counts the bandwidth required over link k, m in the case of hotel i failure.

Function chaining is modeled as follows. For the first transport segment (t_1), constraint (4.18) ensures that there is one outgoing flow for each source s while constraint (4.19) represents the flow conservation towards the ending VNF for transport segment t_1. For the subsequent transport segments ($\{t_2, .., t_k\}$), constraint (4.20) ensures that there is a transport flow starting from the node (i) performing the previous transport function ($x_{i,t_{q-1},s}$) for each source. Constraint (4.21) represents the flow conservation of each transport segment t_q.

4.3 Numerical results

Different application methodologies can be investigated in relation to the above model:

- **Joint**: both service slices (URLLC and eMBB) are jointly optimized with related protection schemes (DPP and SPP, respectively);
- **DPP**: DPP protection scheme is applied for both service slices, which are jointly configured, and is used as a reference;
- **DPP-SPP**: the two service slices are sequentially optimized, DPP/URLLC first. Then the SPP/eMBB slice is added;
- **SPP-DPP**: the two service slices are sequentially optimized, SPP/eMBB first. Then the DPP/URLLC slice is added.

In the last two methodologies (DPP-SPP and SPP-DPP), the ILP was started twice, once for each type of service, and the following strategies are applied to the sequential methodologies for the second step where the second slice is configured:

- **Available resources (AV)**: the optimization algorithm is aware of the resources still available in the network, but not of which nodes and links are active;
- **Active resources (ACT)**: the optimization algorithm is aware of the nodes and links already active.

Table 4.3: Link capacity requirements for different 3GPP split options.

Layers	Split option	Bandwidth [Gbps]
L1	Opt.8	2.4
L2	Opt.6	0.152
L3	Opt.2	0.151
Core	Opt.1	0.150
Cloud	-	0.150

Table 4.4: Active nodes and links for 6 node network and 2 hop constraint on URLLC

Methodology	Nodes	Links
Joint	5	7
DPP-SPP AV	6	14
DPP-SPP ACT	6	12
SPP-DPP AV	6	14
SPP-DPP ACT	6	13

The scenario used for evaluations is the same as in the chapter 3, the 6 nodes network shown in figure 4.1: it consists of 6 nodes connected by 40 Gbps links. Each node is equipped with 25 processing units (PUs) and supports 10 antennas, 5 for the URLLC service and 5 for the eMBB service. The functional splits applied are are the same as in the 3 and shown in the following tab. 4.3, with related bandwidth requirements. The processing units used for virtual functions are 0.5, 0.3, 0.2, 0.1, 0.1 [B24]. The distance between the nodes or each hop is 1 km, which in terms of delay is equivalent to $\tau = 5\mu s$. Numerical results are obtained with by the CPLEX commercial tool [B17].

Table 4.4 shows active nodes and links for 2 hops constraints (only for the URLLC slice). With all methodologies except for the joint methodology, 6 active nodes are needed. This is due to the small size of the network that offers only few degrees of freedom. Differently, for the active links some improvement is present in the case of the ACT strategy when the further optimization is aware of the active nodes and links from the first optimization.

The comparison between Joint and DPP (fig 4.2) shows how the SPP scheme allows to optimize the computational resources (PU), reducing the PU used by 13%. The PU used for the primary and backup functionality of both services are highlighted. While the DPP protection scheme is used for the URLLC service, the SPP protection scheme is used for the eMBB service. It is therefore evident that it is precisely this latter service that allows to optimize the PU used.

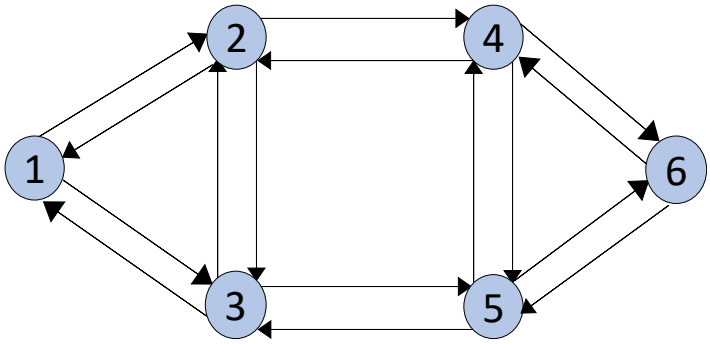


Figure 4.1: Reference 6 nodes network. Nodes are assigned functionalities reported in table 4.3 according to the optimization algorithm.

The reduction of the active links in the ACT case with respect to the AV in DPP-SPP case (tab 4.4) results in an increase in bandwidth usage. This is in accordance with the rules described for band sharing: since the number of active links is lower, the possibility of sharing resources for backup for shared is less, and the use of band resources present in the links increases.

The sequential methodology for slice configuration using DPP first (DPP-SPP) achieves a lower number of active links with respect to the SPP-DPP case (tab 4.4). This is due to the fact that by positioning the SPP slice first, the bandwidth left to DPP is harder to be minimized, either with ACT or AV strategies or even by increasing the number of hops.

Fig.4.3 shows that, with the same number of hops, the worst case is the DPP-SPP ACT case. This is because the links are saturated first with the DPP slices and consequently the slices that use an SPP protection cannot guarantee a greater bandwidth sharing.

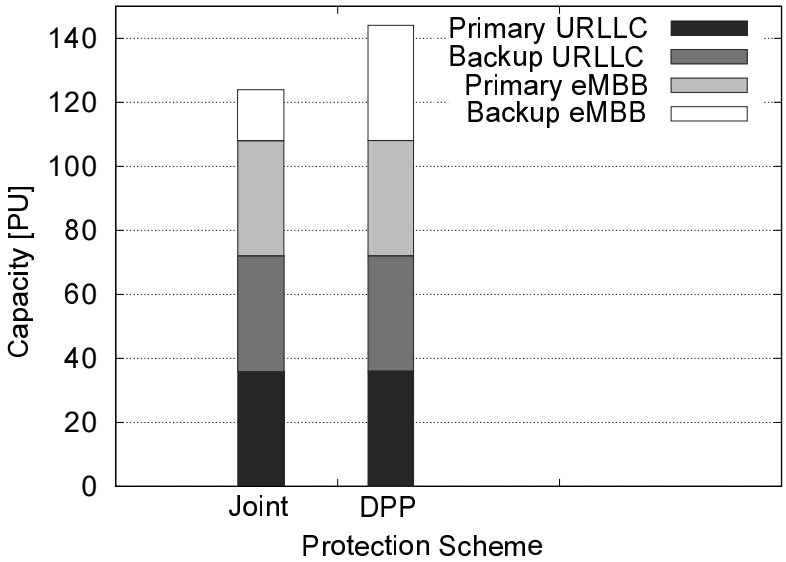


Figure 4.2: Processing resources (PU) for the differentiated joint scheme and the dedicated DPP scheme, evidencing the contributions of primary and backup resources for each class, URLLC and eMBB.

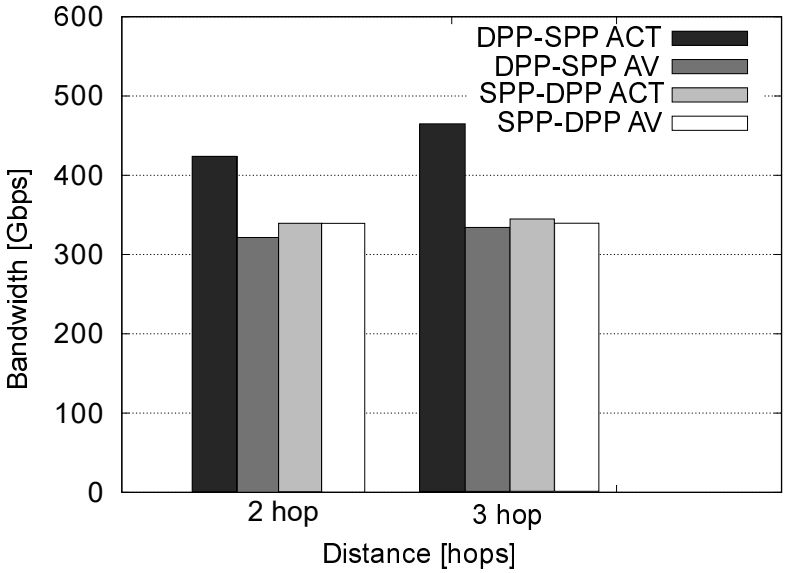


Figure 4.3: Required total bandwidth (Gbit/s) for different hop constraints, by applying ACT and AV strategies with sequential DPP-SPP and SPP-DPP.

Chapter 5

Deployment of scalable edge computing in 5G vehicular networks

Connected vehicles can provide a large set of services for smarter and safer mobility. As an example, a problem highly felt worldwide is road safety ([B25]) where vehicular networks can help in providing prompt information to drivers and alerting possible dangerous situations by allowing vehicles to communicate with each other. It is possible to distinguish between short-range direct communications and long-range network communications [B3, B26]. Different communications need different network requirements, such as low latency, high computational capacity, and high reliability, depending on the application [B27].

To provide the aforementioned services, 5G networks can be used to carry data to/from vehicles and road infrastructure. Centralized cloud-based Radio Access Networks (C-RANs) represent an effective solution to design high-capacity radio access in 5G networks and to support challenging use cases [B28], such as the ones of vehicular networks. C-RAN introduces unprecedented flexibility by efficient application of NFV [B29] jointly with SDN [B28, B30, B31]. To ensure timely network adaptation to user needs, SDN control and management must cope with a potentially high number of network elements and, consequently, the design of control algorithms calls for highly scalable approaches. Virtualized baseband functionalities are suitably located and centralized in BBU ho-

tels, nodes of the optical transport network implementing a C-RAN for enhanced functionality and cost-optimization purposes. This BBU hotels can be provided with an additional computational capacity to perform time-sensitive operations required by low-latency services, as per the MEC [B32]. MEC, by providing 5G with processing resources at the network edge, allows to achieve stringent application requirements. However, widespread deployment of these nodes may be costly; therefore, intelligent nodes hosting BBU hotels and edge computing resources need to be identified in relation to latency and processing constraints. Moreover, the problem of BBU hotel placement in C-RAN has been shown to be NP-hard [B33], requiring novel strategies to make optimal approaches more scalable.
To make the approach more scalable, a hybrid strategy was chosen, combining the potential of ILP and heuristics that will be shown in the following sections

5.1 Problem formulation for reliable service provisioning

The reference C-RAN architecture consists of a hierarchical SDN control plane with a lower layer split into as many controllers as the different kinds of network domains to control, namely the radio network, the optical transport network, and the cloud network. An example of this architectural solution applied to vehicular scenarios is shown in Figure 5.1. The radio domain is composed of antennas and RRUs located at cell sites, and baseband processing functions that are performed over general-purpose hardware in edge nodes. The radio controller is in charge of controlling radio and baseband resources that are remotized following the C-RAN design concept. The optical transport network consists of a set of intelligent nodes interconnected by Dense Wavelength Division Multiplexing (DWDM) optical links to support high-capacity fronthaul in C-RAN. For example, to support heavy and constant fronthaul traffic generated by the Common Public Radio Interface (CPRI) split [B13] (referred to as Option 8 in Reference [B34]), dedicated wavelengths are usually required. Nodes of the transport network, referred to here as

edge nodes, are equipped with processing capabilities to perform MEC functionalities and are managed by the cloud controller. Each controller interacts with the SDN orchestrator to provide information for interworking control and management functions through different domains. The orchestrator is in charge of accommodating new service requests by suitably allocating required resources across the different domains. The orchestrator applies suitable algorithms to properly select the nodes in which the BBU functionalities and services are executed, depending on service and physical network constraints.

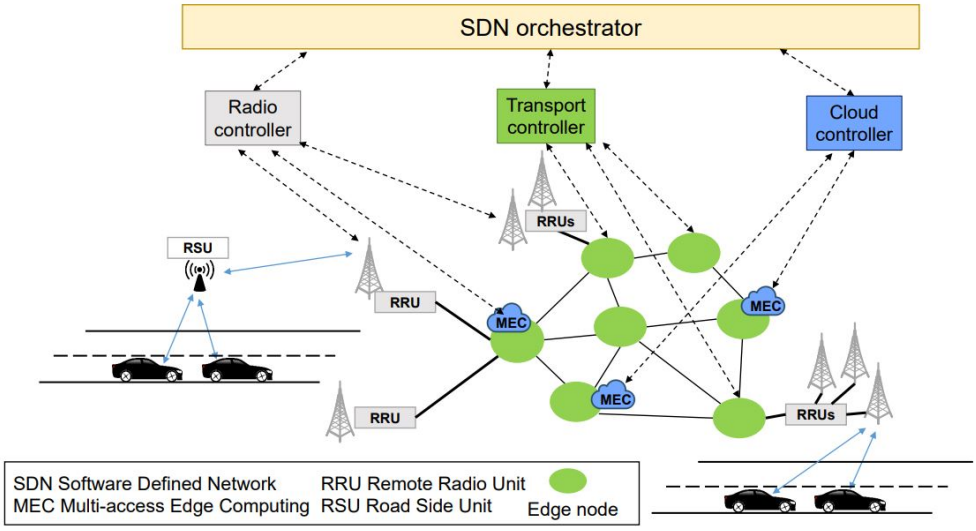


Figure 5.1: Softward Defined Networking (SDN)-controlled Cloud Radio Access Network (C-RAN) architecture for vehicular communications.

C-RAN architecture can be used as an enabler for vehicular communications providing network assistance and commercial services, as depicted in Figure 5.1. Vehicles communicate directly with the mobile network or with Road Side Units (RSUs), that send collected data through the mobile network. Data concerning low-latency applications can be elaborated directly in the edge nodes, thanks to the computational resources offered by the MEC. Computational resources in edge nodes can be used for (i) virtual baseband processing; (ii) virtual mobile core network functions; and (iii) edge application services [B35]. Non-time-sensitive data can be delivered to applications performed in remote locations (not reported in the figure). The traffic destined to remote cloud resources is

user dependent and requires lower bandwidth with respect to fronthaul requirements [B8] and is out of the scope of this paper. In this work, we propose to co-locate, within the same edge node, cloud and BBU processing functions. An edge node is considered to be active when it hosts physical or virtual functions, either for BBU processing or edge core/cloud services.

To provide a reliable C-RAN against single node failures, a 1 + 1 protection solution is desirable to avoid temporary service outages due to resource restoration. Primary and backup path resources must be allocated to provide resiliency against hardware failures. This work considers single active edge node failures (i.e., a failure of all servers placed in an active edge node). The formulation of the joint BBU hotel and edge cloud processing location problem with resiliency is as follows:

- **Given** a set of RRUs to be connected to active edge nodes, a set of edge nodes (candidates to host BBU and edge processing resources), and a set of links connecting edge nodes.
- **Find** active edge nodes and suitable optical resource assignment such that (i) the number of active nodes and (ii) total wavelengths are minimized.
- **Ensure** that each RRU is connected to two active edge nodes (one for primary and one for backup purposes) and that the maximum available wavelengths per link and maximum allowed distance to provide target service are not exceeded.

5.2 ILP and Hybrid strategy

This section proposes an Integer Linear Program (ILP) to solve the joint deployment problem of baseband processing and edge computing with reliability against single-node failure in C-RAN. The main objective of this strategy is to minimize the nodes in which processing capabilities must be installed while ensuring latency and optical link (i.e., maximum wavelengths over fibers) constraints are not violated. To overcome the computational complexity of classical optimization approaches, a hybrid (based on both heuristic and ILP) deployment strategy is also proposed.

The algorithm performs a first phase in which the initial set of nodes candidate to host baseband and edge computing functions is reduced and a suboptimal solution is provided. Then, a second phase is executed for optimization purposes. The latter approach is shown to provide results close to optimal ones while considerably reducing computational time.

5.2.1 ILP model

This algorithm is expected to be executed by the orchestrator, which is assumed to have complete knowledge of the underlying network topology and available resources to provide the placement. The notation used in the algorithm is reported in Table 5.1. The set of nodes in the network, the candidate to host BBU and edge processing functions, is denoted as N, while the number of sources (RRUs) physically connected to node $s \in N$ is denoted as R_s. The connectivity among them is modeled by the C binary matrix. C has one row and one column for each node, and an element is equal to 1 if the two nodes are directly connected by a link, 0 otherwise. Binary variables p^H_{sd} and b^H_{sd} are equal to 1 if node $d \in N$ is the node processing data from RRUs located at node s for primary or backup, respectively. The binary variable h_d is equal to 1 if edge node d is active, i.e., if it acts as a primary or a backup for one or more RRUs. $h_d = 1$ also means that at least one between p^H_{sd} and b^H_{sd} is equal to 1. To connect each RRU to the nodes performing processing functions, one wavelength is reserved along the path, due to the high requirements of physical layer processing functions. This is captured by binary variables w^p_{sdij} and w^b_{sdij}. The maximum available wavelengths over each link and the maximum allowed distance between RRUs and BBUs are indicated with M^W and M^H, respectively. In this formulation, edge processing functions are co-located with BBU processing to reduce the delay to a minimum and to take advantage of the already active nodes, without requiring additional resources on fibers to reach farther facilities. For this reason, only M^H is considered, which is usually more stringent. If this is not the case, M^H could represent the service delay and be used as a more stringent delay requirement. In this work, all links are assumed to be equally long, so M^H is expressed in terms of hops.

The formulation is as follows.

Table 5.1: Notation for Integer Linear Program (ILP).

Parameter	**Definition**
N	set of edge nodes in the network, $\|N\| = n$.
R_s	number of sources (RRUs) directly connected to $s \in N$.
C	$n \times n$ matrix. $c_{ij} = 1$ if node i is directly connected to node j, 0 otherwise.
p^H_{sd}	binary variable, equal to 1 if edge node $d \in N$ acts as primary for RRUs at node (cell site) $s \in N$; 0 otherwise.
b^H_{sd}	binary variable, equal to 1 if edge node $d \in N$ acts as backup for RRUs at node (cell site) $s \in N$; 0 otherwise.
h_d	binary variable equal to 1 if edge node $d \in N$ is active, 0 otherwise.
w^p_{sdij}	binary variable, equal to 1 if the path to connect RRUs at node $s \in N$ and primary edge node $d \in N$ is using physical link $i - j$ $(i, j \in N)$; 0 otherwise.
w^b_{sdij}	binary variable, equal to 1 if the path to connect RRUs at node $s \in N$ and backup edge node $d \in N$ is using physical link $i - j$ $(i, j \in N)$; 0 otherwise.
M^W	max. available wavelengths in each link.
M^H	max. allowed distance between RRUs and edge nodes.
$\alpha, \beta \in \mathbb{N}$	tuning parameters for the objective function.
$L \in \mathbb{N}$	a large number (e.g., $10,000$).

Objective function:

$$Minimize\ F = \alpha \cdot \sum_{d \in N} h_d + \beta \cdot \sum_{s \in N} \sum_{d \in N} \sum_{i \in N} \sum_{j \in N} w^p_{sdij} + w^b_{sdij} \tag{5.1}$$

Constraints:

$$\sum_{d \in N} p^H_{sd} = 1, \quad \forall s \in N \tag{5.2}$$

$$\sum_{d \in N} b^H_{sd} = 1, \quad \forall s \in N \tag{5.3}$$

$$p^H_{sd} + b^H_{sd} \leq 1, \quad \forall s, d \in N \tag{5.4}$$

$$h_d \cdot L \geq \sum_{s \in N} p^H_{sd} + b^H_{sd}, \quad \forall d \in N \tag{5.5}$$

$$\sum_{s \in N} \sum_{d \in N} (w^p_{sdij} + w^b_{sdij} + w^p_{sdji} + w^b_{sdji}) \cdot R_s \leq M^W, \quad \forall i, j \in N \tag{5.6}$$

$$w^p_{sdij} \leq c_{ij}, \quad \forall s, d, i, j \in N \tag{5.7}$$

$$w^b_{sdij} \leq c_{ij}, \quad \forall s, d, i, j \in N \tag{5.8}$$

$$\sum_{i \in N} \sum_{j \in N} w^p_{sdij} \leq M^H, \quad \forall s, d \in N \tag{5.9}$$

$$\sum_{i \in N} \sum_{j \in N} w^b_{sdij} \leq M^H, \quad \forall s, d \in N \tag{5.10}$$

$$\sum_{i \in N} w^p_{sdij} - w^p_{sdji} = \begin{cases} p^H_{sd} & \text{if } j = s, s \neq d, \forall s, d, j \in N \\ -p^H_{sd} & \text{if } j = d, s \neq d, \forall s, d, j \in N \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

$$\sum_{i \in N} w^b_{sdij} - w^b_{sdji} = \begin{cases} b^H_{sd} & \text{if } j = s, s \neq d, \forall s, d, j \in N \\ -b^H_{sd} & \text{if } j = d, s \neq d, \forall s, d, j \in N \\ 0 & \text{otherwise} \end{cases} \tag{5.12}$$

The multi-objective function in equation (5.1) is composed of two members. The first term takes into account the activation cost of each node, while the second term accounts for the wavelengths required to connect RRUs to edge nodes, both primary and backup.
The constraints of Equations (5.2) and (5.3) ensure that there is only one primary and one backup edge node, respectively, for each RRU.
The constraint of Equation (5.4) guarantees that primary and backup nodes are disjoint.
The constraint of Equation (5.5) counts the number of active nodes (i.e., performing processing functions) in case they are acting either as a primary or a backup for any RRU.

The constraint of Equation (5.6) limits the number of wavelengths over each link for both primary and backup in both directions (i.e., from i to j and j to i together).
The constraints of Equations (5.7) and (5.8) ensure the feasibility of the connections so that a link between two nodes can be used if and only if it exists in the physical topology.
The constraints of Equations (5.9) and (5.10) limit the maximum distance between RRUs and BBUs to M^H for primary and backup paths, respectively.
Finally, Equations (5.11) and (5.12) are the flow conservation constraints for primary and backup paths, respectively. These constraints are needed to reserve the paths connecting RRUs to their primary and backup edge nodes. In this model, wavelength conversion is allowed in the network nodes.

5.2.2 Hybrid two-phases model

The hybrid approach proposed here is performed in two phases. In the first phase, a heuristic is proposed to provide a computationally simple but reliable C-RAN coverage by guaranteeing that each RRU has both a primary and a backup node and that minimum delay is achieved. The second phase is an optimization process, based on a modified version of the ILP proposed in Section 5.2.1, that aims at reducing the number of active nodes found in phase 1. The details of the hybrid algorithm are reported below.

Phase 1 is assumed to start from a C-RAN configuration where no edge node is active, i.e., BBU and edge functionalities have yet to be assigned to nodes. This has, anyway, no impact on the generality of the approach. In this phase, the edge node activation is performed within a 1-hop distance or, equivalently, RRUs can be connected only to the node itself or to a neighbor edge node. This implicitly assumes that there are enough resources on the links connecting neighbors and guarantees that delay constraints are always satisfied. It should be noted that, to solve the deployment problem, primary and backup nodes must be selected. Therefore, not satisfying the aforementioned condition on the link resources does not guarantee a solution to the problem.

In addition to the C matrix needed to model the physical links (see Table 5.1), two additional structures are introduced here:

- H **matrix**: This is an $n \times 2$ matrix, where each row represents a node of the network; the first column indicates which is the primary edge node chosen by the node on that row, while the second column indicates which is the backup node.

- W **matrix**: This is an $n \times n$ matrix which keeps track of the use of the links between nodes. In W, there is one row for each source edge node (where the RRUs are physically connected). W has one column for each edge node, that is, the possible locations for the edge server performing baseband and services for the specific RRUs. This matrix is needed to provide a feasible solution at the end of phase 1 but is not used in phase 2.

Algorithm 1 presents the pseudo-code of the algorithm executed by each node of the network during phase 1. In the beginning, the algorithm starts with empty H and W matrices (line 2). This algorithm executed in a sequence for each node until all nodes in the network have both primary and backup connections (condition in line 4). Then, node i checks some conditions for the primary and for the backup connection in order to find suitable edge nodes. If node i is already active (line 6), it can use itself as the primary edge node (line 7). Otherwise, node i must search among its neighbors to find an already active node (line 8) and, if it succeeds, makes the primary connection to the edge node j (line 9) and updates W matrix accordingly (line 10). The updating phase stores in the position i, j of the matrix the required wavelengths over link i–j. If no neighbor is active (line 11), node i activates itself and makes the primary connection to itself (lines 12 and 13).

After establishing the primary connection, node i executes a set of instructions to find the backup edge node. There are two possible situations. The first situation is when node i is already active and plays the primary role for the RRUs connected to itself or not active at all (line 16). In this case, node i either finds a directly connected neighbor node (j), which is already active and satisfies the distance restriction, and connects to it (lines 17–19) or chooses randomly one of the neigh-

bors as a backup, defines the backup connection, and updates W matrix accordingly (lines 20–23). The other situation happens when node i is active (line 25). Node i can take advantage of this situation and makes the backup connection to the local edge node (lines 26 and 27). Phase 1 stops when all nodes in the network have both connections to primary and backup nodes.

The objective of the second phase is to minimize the number of active nodes. This is achieved by reassigning the RRU connections and shutting down active nodes by further centralizing BBU and edge processing functions within the distance constraints (M^H). This is achieved by adding the following set of constraints to the ILP model presented in Section 5.2.1:

Equation (5.13) forces the node candidates to be 0 (non-active) for all the nodes excluded by phase 1 (i.e., for all the nodes that have no RRU assigned to them, either for primary or backup purposes). The ILP is then solved with a reduced set of candidate nodes that always ensures the feasibility of the solution.

$$h_d = \begin{cases} 0 & \text{if } H_{d0} + H_{d1} = 0, \quad \forall d \in N \\ \{0,1\} & \text{otherwise} \end{cases} \tag{5.13}$$

Algorithm 1 C-RAN reliable coverage (phase 1).

Initialization:
$H, W \leftarrow \emptyset$
Begin:
while exists node $i \in N$ s.t. $(H_{i0} = 0) \vee (H_{i1} = 0)$
//Primary connection assignment:
    if $h_i = 1$
        $H_{i0} = i$
    else if $\exists$ node j s.t. $c_{ij} = 1$ **and** $h_j = 1$
        $H_{i0} = j$
        update W
    else
        $h_i = 1$
        $H_{i0} = i$
    end if
//Backup connection assignment:
    if $(h_i = 1$ **and** $H_{i0} = i)$ **or** $(h_i = 0)$
        if $\exists$ node j s.t. $c_{ij} = 1$ **and** $h_j = 1$
            $H_{i1} = j$
            update W
        else
            activate random neighbor j $(h_j = 1)$
            $H_{i1} = j$
            update W
        end if
    else
        $h_i = 1$
        $H_{1i} = i$
    end if
end while
End

5.3 Numerical results

Numerical results are obtained in different networks to evaluate the effectiveness of the ILP and hybrid solutions in terms of active edge nodes and of the centralization gain, G_C, that is the advantage related to centraliz-

ing BBU and cloud functionalities, expressed by the following formula:

$$G_C = \frac{|N| - \sum_{d \in N} h_d}{|N|} \tag{5.14}$$

where $|N|$ and h_d have been defined in Table 5.1. Three sample networks, N_{38}, N_{20}, and N_{14}, consisting of 38, 20, and 14 nodes, respectively, are considered, as represented in Figure 5.2. Evaluations assume here that 10 RRUs are physically connected to each node to provide mobile network coverage and transmission capacity for vehicular network, and the adoption of CPRI (option 8 in Reference [B34]). The proposed algorithms and evaluations can be extended to different numbers of RRUs, possibly unbalanced among edge nodes and suitably adapted to different functional split, which is left for future works. The commercial tool CPLEX [B17] is used to run the ILP on a computer with 4 cores at 3.2 GHz and 8 GB of RAM. Tuning parameters α and β are set to a value of $\alpha >> \beta$ so that the minimization of active edge nodes is prioritized, while the maximum number of wavelengths over each link M^W is set to 80.

In Figures 5.3–5.5, comparisons are reported between the hybrid and the ILP approaches by plotting the results in terms of the number of active edge nodes as a function of the allowed distance, expressed in hops. The cost of the hybrid solution depends on the node from which the heuristic procedure starts: the maximum and minimum costs in terms of total number of active nodes obtained are both reported in the plots. In addition, the results at the end of phase 1 of the hybrid strategy are also shown, as lines and denoted as H, to outline the effect of the optimization phase. These lines are constant because they do not depend on the distance, as they provide a solution within 1 hop distance. The costs obtained with the hybrid and ILP approaches decrease with the distance in all networks. The minimum value that can be achieved is 2 because one primary and one backup node must be always present to cope with single edge node failure. In case of tight distance constraints (e.g., 1 or 2 hops), data cannot be transported far in the network; thus, many edge nodes must be activated. When the distance constraint increases, farther nodes in the network can be reached and, consequently, the number of total active nodes decreases. From the figures, it can be seen also the influence of the starting node, represented by the difference between the maximum

and the minimum costs. In the worst cases, only one additional node must be activated. In addition, the results of the hybrid are shown to be the same as the optimal ones in most of the cases. However, in very few cases, the hybrid approach cannot achieve optimal solutions due to the choices performed in phase 1, where some nodes are excluded by the pool of possible active nodes and cannot be activated in phase 2.

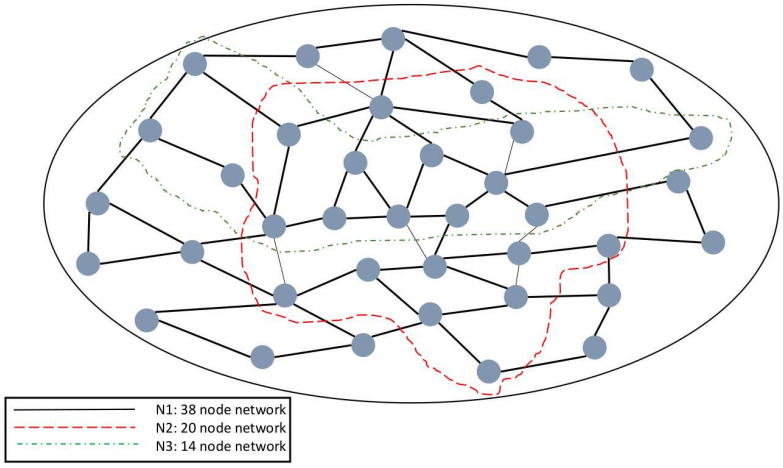


Figure 5.2: N_{38}, N_{20}, and N_{14} C-RAN topology for numerical evaluations.

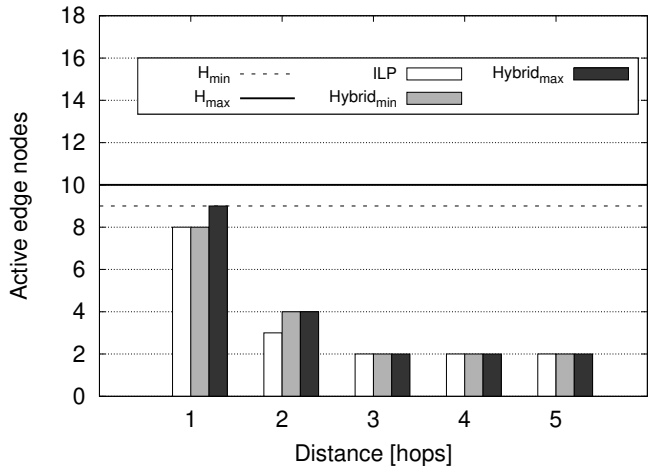


Figure 5.3: Total number of active edge nodes as a function of the allowed distance between RRUs and edge nodes for network N_{14}: Maximum and minimum costs of the hybrid results are reported after both phases.

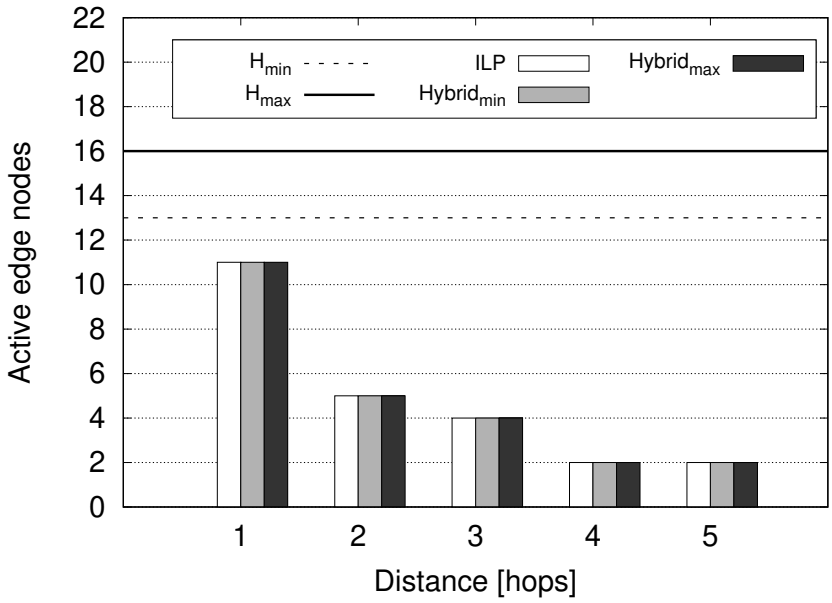


Figure 5.4: Total number of active edge nodes as a function of the allowed distance between RRUs and edge nodes for network N_{20}: Maximum and minimum costs of the hybrid results are reported after both phases.

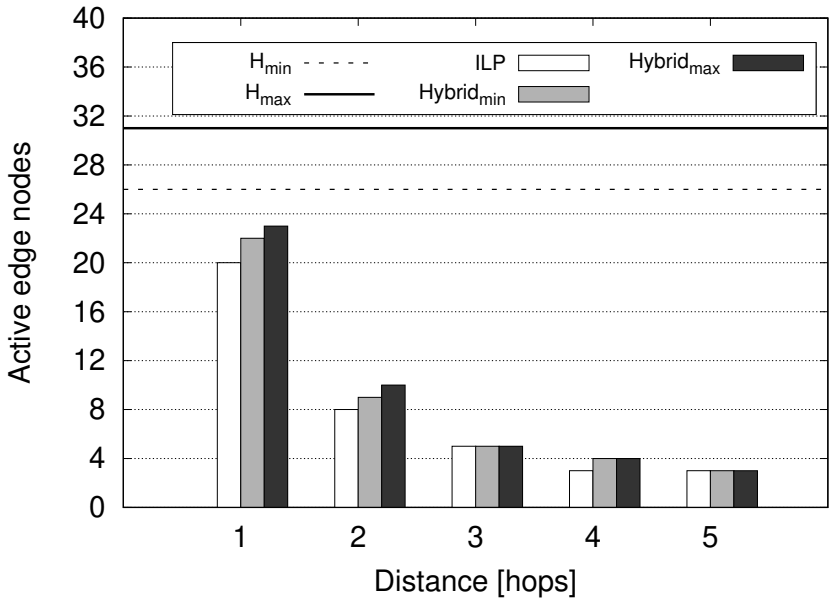


Figure 5.5: Total number of active edge nodes as a function of the allowed distance between RRUs and edge nodes for network N_{38}: Maximum and minimum costs of the hybrid results are reported after both phases.

In Figure 5.6, the gain of centralization of BBU and edge cloud functionalities is presented as a function of the allowed distance from RRUs by comparing the ILP results with the results of the hybrid approach at the end of phase 1 (denoted as H) and phase 2 in the maximum-cost case. This gain is relevant both for ILP and hybrid, with the hybrid being very close or coincident to the optimal solution. In the worst case (i.e., distance constraint equal to 1 hop), the hybrid provides only 8% gain reduction. As expected, phase 1 provides only suboptimal solutions. It is, therefore, evident the role of phase 2 of the hybrid approach in achieving a high centralization gain with respect to the plain coverage achieved in phase 1.

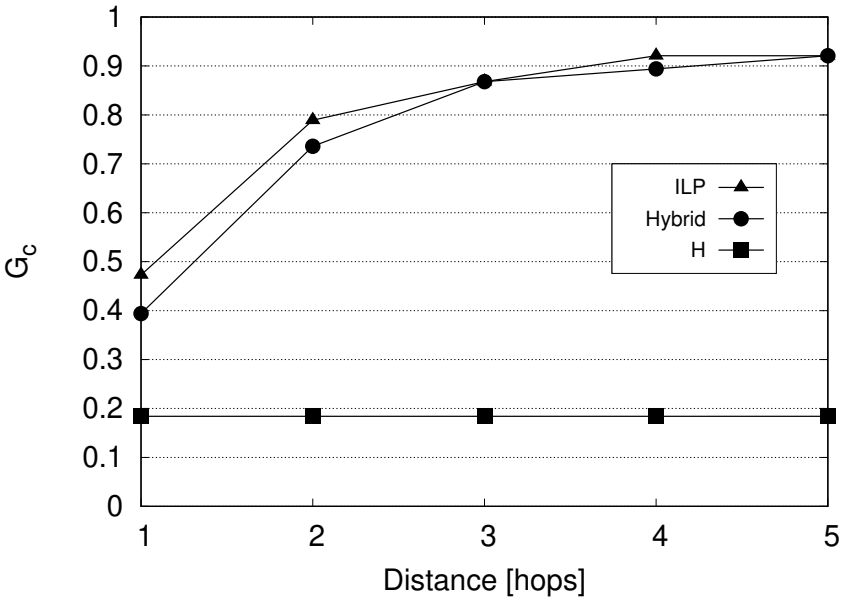


Figure 5.6: Centralization gain as a function of the allowed distance between RRUs and edge nodes for network N_{38}: Results are reported for the maximum cost for hybrid (phase 1 and phase 2), and ILP.

Table 5.2 reports the number of active links, wavelengths over the most used link, and overall wavelengths in network N_{38} for the two strategies. By comparing the strategies, it is possible to observe that the ILP requires a slightly higher number of wavelengths with respect to the hybrid approach when the number of active nodes is lower (distance constraints 1, 2, and 4). Nevertheless, because the activation cost of a

Table 5.2: Number of active links, wavelengths over the most used link, and total wavelengths for the hybrid and ILP for different distance constraints in network N_{38}.

Dist.	**Hybrid**			**ILP**		
[hops]	**Active**	**Max**	**Total**	**Active**	**Max**	**Total**
1	45	10	530	48	10	560
2	51	40	950	50	40	1040
3	49	70	1370	51	60	1350
4	52	70	1530	48	80	1830
5	51	80	1790	52	80	1780

node is much larger than the cost of a wavelength, the ILP solution always reaches a lower cost solution compared with the hybrid approach. When the ILP and hybrid require the same amount of active nodes (distance constraints 3 and 5) the ILP requires fewer wavelengths than the hybrid approach due to a wider set of choices. This happens for similar reasons also for the wavelengths required over the most used link. To solve the harder instances of the problem, the ILP takes 2.8 s, 22.75 s, and 10,010.17 s in the network N_{14}, N_{20}, and N_{38}, respectively, showing an increased computational complexity when the size of the problem increases. Solving the ILP with the hybrid approach instead allows to reduce the solving times to 2.2 s, 17.99 s, and 3647.88 s in the three networks due to the reduction of the solution space. It should be noted that, in order to see the differences between the two strategies, the evaluations proposed here are done for networks suitable to cover a small- or medium-sized city. In larger scenarios (i.e., networks with more edge nodes and links), it is not always possible to ensure a solution with the ILP approach. These scenarios can be instead tackled with the hybrid approach, which has been shown to provide results close to optimality.
Results show that the hybrid approach provides similar results to the ILP ones while considerably reducing the solving time.

Chapter 6

Dynamic slice simulation in a 5G vehicular network in the metro segment

Network slicing allows the provisioning of different services over the same 5G infrastructure, where virtual or physical resources are interconnected to form end-to-end logical networks (i.e., the slices) [B36]. Slicing allows service providers to offer 'network slices-as-a-service', tailored to different performance requirements [B37]. When a slice is admitted by a provider (i.e., it is deployed over its infrastructure), it needs to be assigned a proper set of resources (i.e., connectivity and compute) to meet the Service Level Agreement (SLA) stipulated with the client [B38]. In this respect, provisioning slices with very stringent reliability and latency requirements is an important challenge to tackle [B24].In the presence of failures, to avoid severe service interruptions while keeping the number of backup resources low, optimized provisioning of extra resources (dedicated or shared), is required. Compared to using dedicated backup resources, an approach leveraging on shared protection resources potentially leads to (*i*) fewer resources consumed by each slice, and, consequently, (*ii*) to a better slice admission ratio performance (or, equivalently, a lower blocking probability). These advantages come at the cost of a slightly longer recovery time (i.e., compared to using dedicated backup resources) due to the need to switch from the primary to the backup resources. In [B40, B39], the authors evaluate the impact

of different techniques for dedicated and shared backup protection on optical network resources. In [B42, B41] efficient shared and dedicated protection schemes for cloud and baseband resources in 5G access/metro networks are applied.

All the works mentioned so far consider only one technology domain (either transport or cloud), while [B24] presents static resource provisioning strategies for dedicated and shared backup resources, considering transport and cloud domains jointly, for a single URLLC service slice provisioning.

This chapter considers, on the other hand, the slice as a service paradigm. It proposes a heuristic that tackles the problem of dynamically provisioning slices with stringent latency and reliability requirements while minimizing the amount of transport and cloud resources assigned to each slice. The intuition behind the proposed approach is to encourage sharing of backup connectivity and cloud resources as much as possible. The performance of this shared protection scheme is compared against a conventional dedicated protection mechanism in terms of slice blocking probability and required processing resources considering a sample 5G metro network, where processing power is often limited and requires efficient resource allocation.

6.1 Problem formulation and methodology

In 5G networks, baseband functionalities can be virtualized over general-purpose hardware and centralized at different computing locations in the network, reaching different degrees of savings and performance targets [B36]. The transport network links interconnecting the different compute nodes must be dimensioned accordingly to meet bandwidth and latency requirements for baseband and service processing. This chapter considers a metro network comprising a set of source and target nodes connected by high-capacity optical transport links. The source nodes collect a set of antennas covering a given area and injecting traffic into the transport network. The target nodes are equipped with Processing Units (PU) to perform virtual baseband functions and services. The formulation of the dynamic and resilient slice allocation problem can be

Algorithm 2 Slice Admission

```
1: Algorithm:
2:     for all primary_i in Pair
3:         if primary_i meets slice_requirements
4:             for all backup_j associated with primary_i
5:                 if backup_j meets slice_requirements
6:                     compute cost_{i,j}
7:                     add primary_i and backup_j in List_pair
8:     if List_pair is Empty
9:         slice_rejected
10:    else ascending sort List_pair based on cost_{i,j}
11:        hold resources required by List_pair[0]
12:        slice_accepted
```

summarized as follows. **Given**: a network topology with available network resources (bandwidth and PU) and the requirements of a slice to be provisioned; **Find**: a suitable slice deployment, such that the allocated network resources are minimized; **To ensure**: reliability against single link or node failure (including the target node) while ensuring that the bandwidth and PU resources allocated at each link and node do not exceed the available resources, and that the maximum distance between a source node and a target node is enforced.

- **Given**: the network topology with available network resources (bandwidth and PUs) and slice requirements.
- **To find**: a suitable slice positioning, so that the allocated network resources are minimized.
- **To ensure**: reliability against single link or node failure while enforcing the maximum distance between a source node and a target node. Available bandwidth and PUs must not be exceeded.

To study the problem, an event-driven simulator written in Python was developed. Events consist of slice requests originating at random source nodes, which can be allocated or rejected in relation to the amount of resources available in the network. Each slice also has a lifetime after

which the allocated resources are released. A pre-processing phase generates a set (referred to as $Pair$) of primary-backup path pairs between each source and target node. Each primary-backup pair is path disjoint and terminates at different target nodes. Each path is obtained using the k-shortest path algorithm with $k = 5$. Two different approaches for protection are considered, Dedicated Protection (DP) and Shared Protection (SP). Slice resource assignment is carried out as shown in Algorithm 2. For each possible primary path in $Pair$ (line 2)), the algorithm checks if latency constraint is met, and if there are enough resources (bandwidth and PU) available on both the primary path ($primary_i$) and at the candidate target node (i.e., to which the candidate primary path connects the source node to) to allocate the slice (line 3). The selection of the backup path and target node $backup_j$ (line 5) depends on the specific protection strategy. DP uses dedicated backup resources. As a result, a procedure identical to the one used for the selection of the primary path and target node is used. With SP, backup resources (bandwidth and/or PUs) can be shared at no additional cost if their respective primary paths and target nodes are disjoint. Otherwise, new backup resources are assigned.

If enough resources are available on the considered primary/backup pair, and latency constraint is met, a cost is calculated as:

$$cost_{i,j} = \alpha * \sum_{i \in Links} Band_i + \beta * \sum_{j \in Nodes} Compute_j \tag{6.1}$$

where α and β are coefficients used to balance the two contributions, $Band_i$ is the connectivity requirement (in Gbps) of the slice and $Compute_j$ is the PU required by the slice for the virtual baseband and the service (line 6). The value of $cost_{i,j}$ is saved together with the primary-backup pair (line 7) and, after exploring all possible pairs, the smallest cost pair is chosen (line 10). The resources needed for the slice are then booked into the network (line 11), and the slice is allocated. Resources are released after the expiration time. If there are no candidate pairs due to lack of resources, the slice is rejected (lines 8-9).

6.2 Numerical results

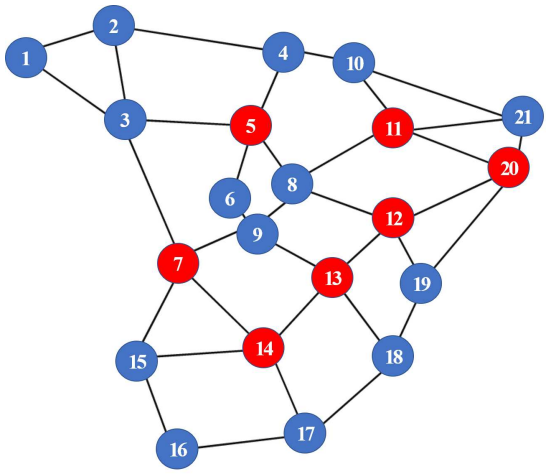


Figure 6.1: Reference network with source (blue) and target (red) nodes.

Table 6.1: Average number of PUs per slice: SP savings compared to DP for different load values.

Load	SP Savings (PU=500)	SP Savings (PU=2000)
150	35.62%	35.41%
160	35.88%	35.50%
170	36.16%	36.61%
180	36.47%	36.71%
190	36.75%	36.79%
200	37.03%	37.85%

The network considered in the analysis is depicted in fig. 6.1. It consists of 14 source (blue) and 7 target nodes (red). The target nodes are chosen among those with nodal degree equal to 4, to allow better accessibility from the source nodes. The links in the transport are bidirectional, are assumed to have the same length, and have a capacity of 1000 Gbps.
This study considers the deployment of slices with low latency and strict reliability requirements. The maximum number of hops allowed for a slice (to satisfy the latency requirements) is set to 4. The slice connectivity

requirement is 24 Gbps and the compute one is 12 PUs (i.e., for baseband and service processing adopting split option 8 [B24]). The distribution of the average inter-arrival frequency is exponentially distributed with $\lambda = 1$ per time unit. The average lifetime of the slice (θ) is also exponentially distributed, varied to consider different values of the network load ($A_0 = \lambda * \theta$), with α and β setted to 1. The results compare the two different protection schemes (DP and SP) in two network configurations, one with 500 PU (case 1) and the other with 2000 PU (case 2) available at each target node. The following quantities are introduced for evaluation:

$$B_S = \frac{\sum_{i=1}^{N} \frac{B_i}{S_i}}{N} \tag{6.2}$$

$$PU_S = \frac{\sum_{i=1}^{N} \frac{PU_i}{S_i}}{N} \tag{6.3}$$

$$B_U^{j,k} = \frac{\sum_{i=1}^{N} B_i^{j,k}}{N} \tag{6.4}$$

$$PU_U^j = \frac{\sum_{i=1}^{N} PU_i^j}{N} \tag{6.5}$$

where B_S represents the average bandwidth occupied by a slice in the network, defined as the ratio between the total bandwidth used in the transport network B_i and the number of active slices S_i when slice i is accepted, averaged over the number of events "slice accepted" N. PU_S represents the average number of PUs per slice, defined in (6.3) in a similar way, where PU_i indicates all the PUs used when slice i is accepted. The average bandwidth per link $j-k$ ($B_U^{j,k}$) and the average number of PUs per node j (PU_U^j) used in the network are represented by (6.4) and (6.5), respectively, where $B_i^{j,k}$ is the bandwidth allocated in the link that connects the node j and k, and PU_i^j indicates the PUs allocated in node j when slice i is accepted.

The blocking probability is represented by (3), where N_r is the number of events "slice request", Bl_i is equal to 1 if the slice is rejected or 0 if it is accepted.

$$\pi_b = \frac{\sum_{i=1}^{N_r} Bl_i}{N_r} \tag{3}$$

Figure 6.2 compares SP and DP in terms of blocking probability. SP outperforms DP, in particular when PU resources are scarce (case 1).

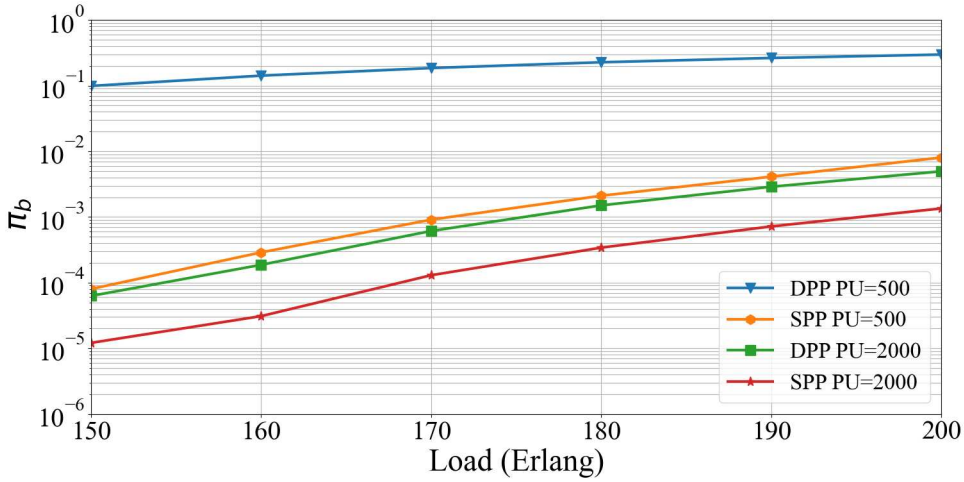


Figure 6.2: Blocking probability as a function of the load.

Figure 6.3 reports the value of B_S as a function of load. SP allows savings of up to 28% and 8.6% in case 1 and 2, respectively. While in case 2 there is almost no effect with different load conditions, in case 1 DP requires 9.6% additional bandwidth when passing from low to high load. This is because resources in the target nodes are scarce and saturates, forcing DP to try to reach nodes with available PU that are further away. This is shown in Fig. 6.4, where the average number of PUs per node is reported for load = 200.

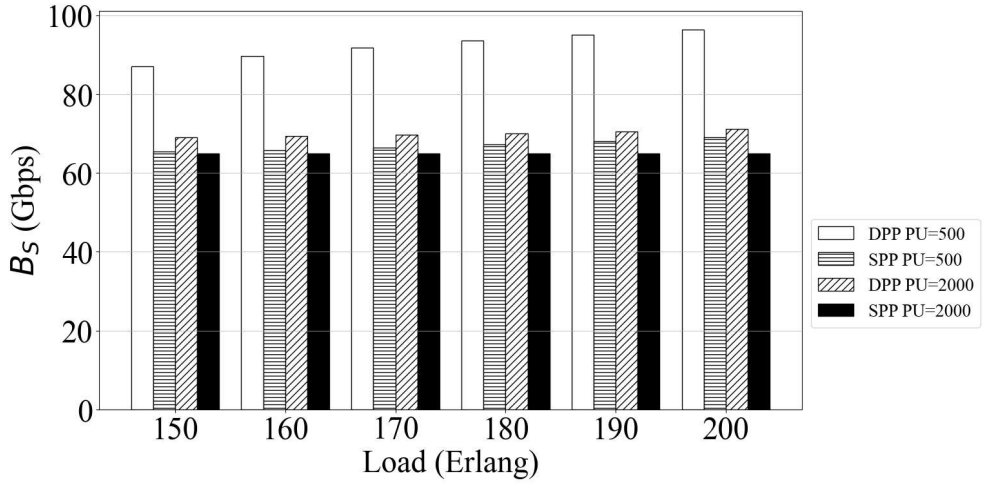


Figure 6.3: Average bandwidth per slice (B_S) as a function of the load.

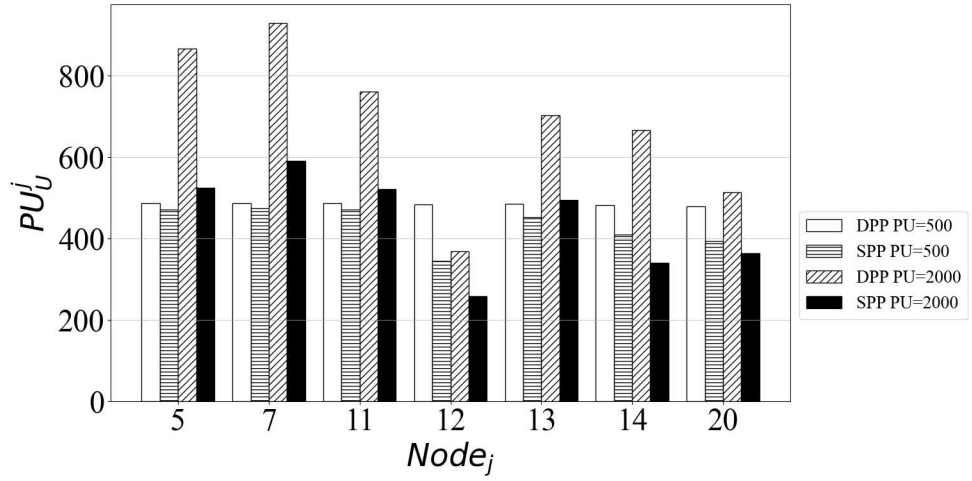


Figure 6.4: Average PU per target node j (PU_U^j) with load = 200.

In case 1, DP uses on average all resources in all the nodes. SP is able to use resources more efficiently, thus the lower blocking probability. In case 2, the average PU usage is below 50% for both DP and SP. The reason for the higher blocking shown by DP is bandwidth availability over some links. This can be seen in Fig. 6.5, where the bandwidth used per link is, on average, close to saturation.
Table 6.1 shows how many PU can be saved, on average and per slice, using SP compared to DP.

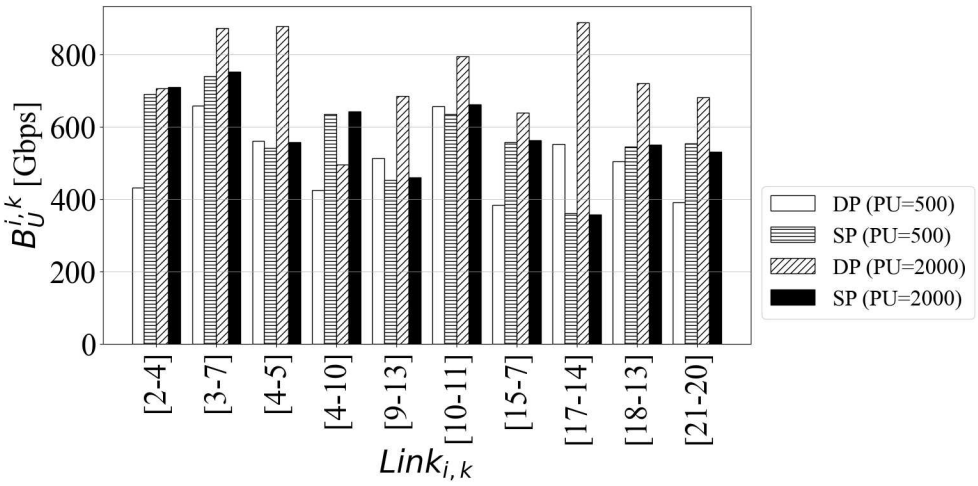


Figure 6.5: Average bandwidth per link $j-k$ ($B_U^{j,k}$) with load = 200. Only links with utilization $> 60\%$ are reported.

As the load increases, the savings slightly increase, reaching 37% savings per slice. This is due to the higher number of slices activated simultaneously, which allows sharing a larger number of backup PUs.

Chapter 7

Dynamic slicing optimization for 5G vehicular network

In this chapter, experimental changes to network management protocols will be proposed, always aimed at optimizing the distribution of resources, focused both on the sharing of bandwidth between links in the case of management of dynamic bandwidth resources, and on the sharing of PUs. In the case of dynamic management of the processing units. Finally, a model for network reconfiguration will be presented, based on different policies. An ILP model will also be presented, still being tested.

7.1 Management of dynamic network resources: Dynamic Bandwidth

To study the dynamic management of the band, reference was made to a simulator written in Python. Dynamic network management means the management (and therefore the allocation) of the bandwidth in the links. In particular, two different methodologies were studied:

- **Case A**: The bandwidth is assigned based on the number of outgoing links per node. Given a total bandwidth B to be allocated to links in the networ , the band is divided by the number of nodes in the network (N_{tot}). Each node is assigned an amount of bandwidth equal to B_{tot}:

$$B_{tot} = \frac{B}{N_{tot}} \tag{7.1}$$

This amount B_{tot} must be divided on each outgoing link.

– **Case B**: The band is assigned based on the type of node, if the node is a destination node or if the node is a source node. Considering the total band B, the ratio with which the band is assigned is 3 to 2. Therefore the partial band to be assigned to the nodes is:

$$B_{tot} = \frac{B}{5} \tag{7.2}$$

If the links entering the target nodes are $L_{t,tot}$, the bandwidth per target link is equal to:

$$L_t = \frac{B_tot * 3}{L_{t,tot}} \tag{7.3}$$

If the links entering source nodes are in total $L_{s,tot}$, the bandwidth per source link is equal to:

$$L_s = \frac{B_tot * 2}{L_{s,tot}} \tag{7.4}$$

The L_t and L_s bands are assigned to all links in the network.

In this scenario, therefore, the band is no longer statistically and equally assigned to the network nodes, but is distributed based on the network topology, making the assignment dynamic

7.1.1 Scenario and Results

The reference network is the one in the figure 7.1, where the number of nodes in the network is 21. In this case, the blue nodes represent the source nodes, the red nodes are the target nodes. The total bandwidth available in the network 70,000 Gbps. In the reference scenario, where all links have the same amount of bandwidth, 1000Gbps are assigned to each link. For Case A , the amount of bandwidth assigned to each node is 3333Gbps, which will be shared equally by the links leaving the node. For Case B, the bandwidth available for the links entering the target

nodes is 2333 Gbps, for the links entering the source nodes it is 538.5 Gbps Two different scenarios will be shown, based on the number of PUs present in the target nodes, respectively 500 and 2000 PU. Only a shared protection is considered. The other network settings are the same as in the chapter 6

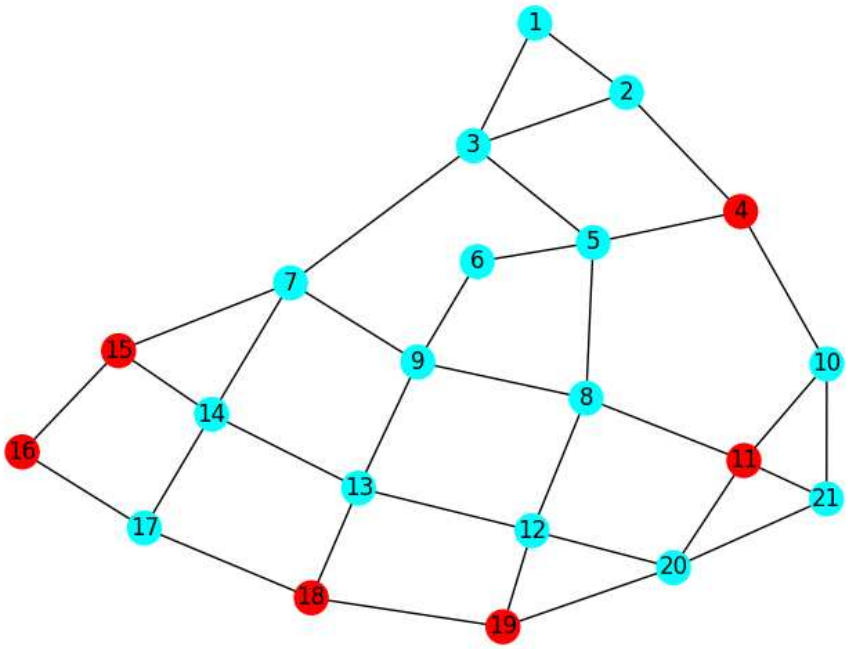


Figure 7.1: Reference network with source (blue) and target (red) nodes.

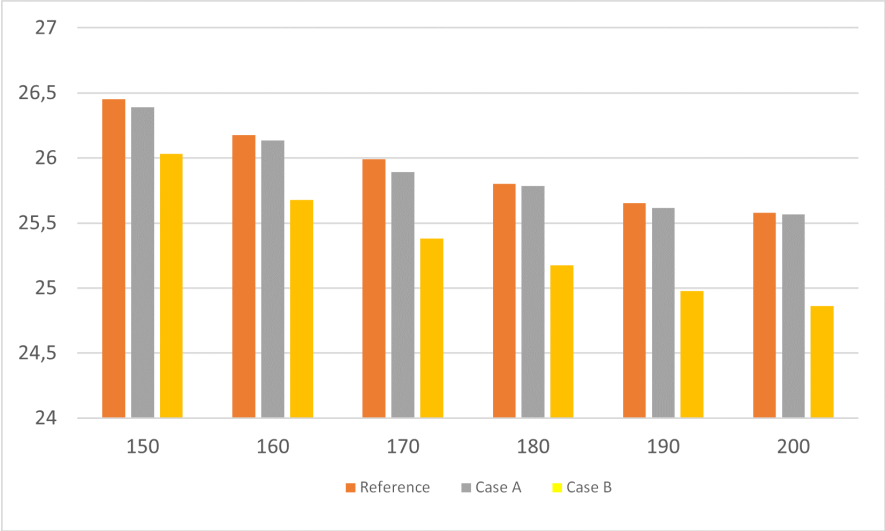


Figure 7.2: Total bandwidth used on average (B_u) per link, case 500 PU.

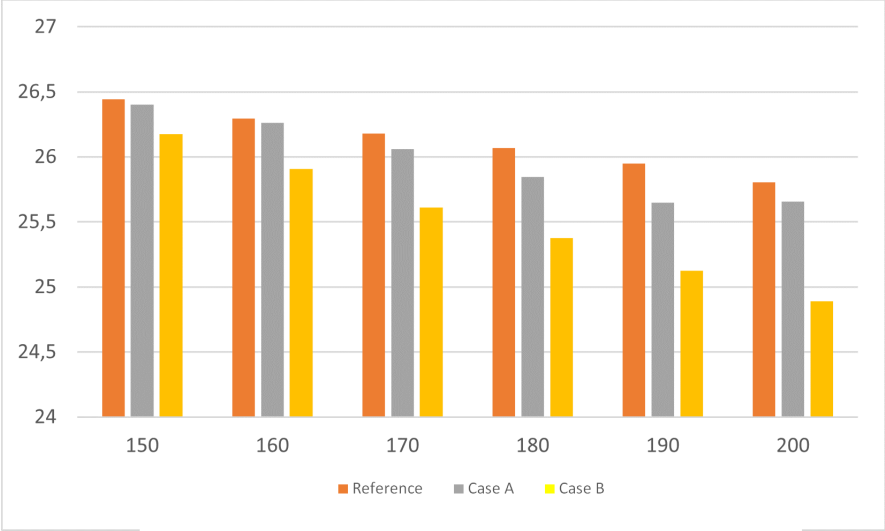


Figure 7.3: Total bandwidth used on average (B_u) per link, case 2000 PU.

Fig 7.2 and fig 7.3 show a small Gbps saving, in the two cases (A and B), of 1% after the first and 4% after the second, in both configurations. It can also be noted that as the load increases, the savings in bandwidth used also increases. This happens due to the large number of simultaneously active slices, which therefore allows to share a larger amount of backup resources. It is therefore possible to optimize the use of

bandwidth resources with subsequent changes. With the reference condition, from the moment in which some links became saturated, specifically those affluent to the target nodes, longer paths were sought to allocate the slices. With this type of optimization, however, the aforementioned links do not saturate and the bandwidth per slice is reduced.

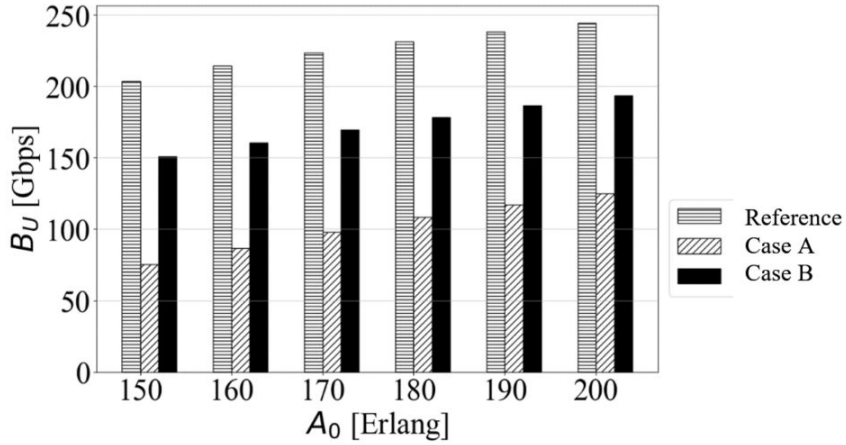


Figure 7.4: Active slices on average in the link, with 200 Erlang load and 500 PU configuration. Links with¿ 50% or ¡5% usage are considered.

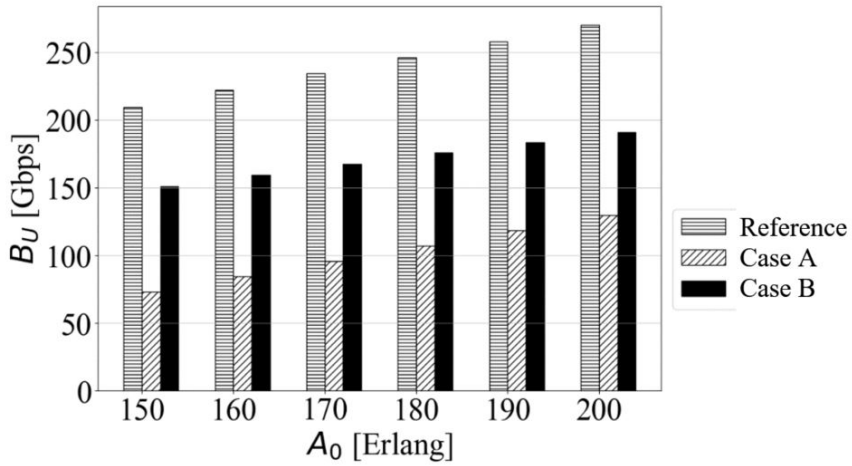


Figure 7.5: Active slices on average in the link, with 200 Erlang load and 2000 PU configuration. Links with¿ 50% or ¡5% usage are considered.

With fig 7.4 and fig 7.5 highlight the difference between the results of the case A and those of the other two variants. The low value, which would suggest a great saving of bandwidth per link, actually derives from the implemented bandwidth distribution algorithm, which provides resources in proportion to the number of links outgoing from the node. In this case, nodes such as "1", which will be discussed in detail later, find themselves having a large amount of unused band available, which in the final calculation greatly influences the average. Being it marginal, it actually manages to reach a few target nodes through certain paths and, consequently, the aforementioned paths available to it end up becoming saturated quickly, as they have further connections.

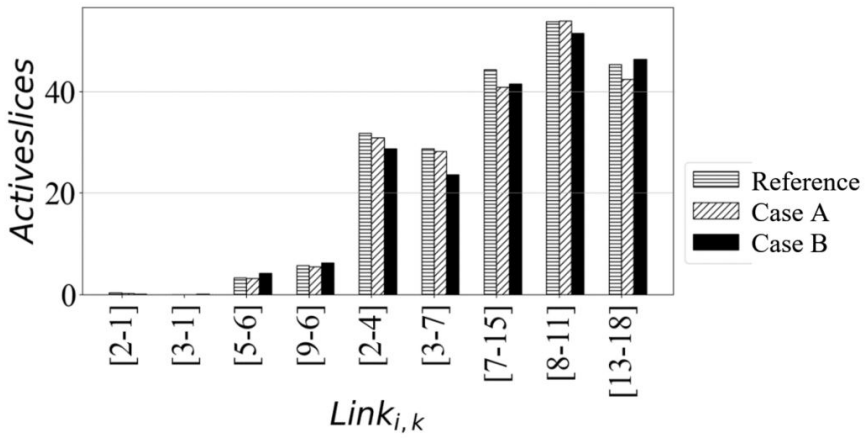


Figure 7.6: Backup band used on average per slice (Bs), in the SPP configuration with 500 PUs available

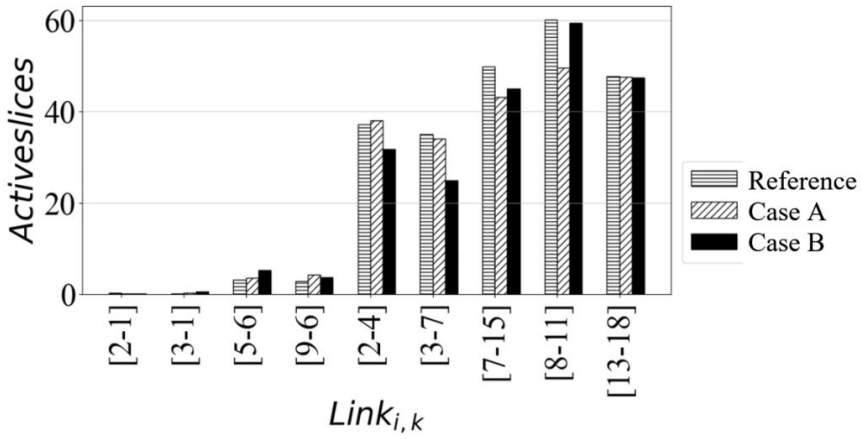


Figure 7.7: Backup bandwidth used per slice (B_s) on average, in the SPP configuration with 2000 PUs available

Fig 7.6 and fig 7.7 show the active slices in a link, on average, for the duration of the simulation. The average bandwidth usage per slice was used as a discretion value to choose which links to focus on, placing the use greater than 50% and less than 5% of the resources available as points of interest, evaluated on the reference condition at maximum load (200 Erlang). These limit values were taken in order to highlight two particular factors of the network configuration: the large load supporting the links directly connected to the target nodes and, on the contrary, the scarce use of resources assigned to marginal links. In particular, the nodes “1” and “6”, two source links, are xamples of the scarce use of the assigned network. Note, the links taken into consideration are entering these nodes, so such a low use of resources means little use in finding a path for other source nodes. This is identified in the fact that node ”1”, being marginal and connected only to two other non-target nodes, will almost never be a favorable alternative, for other source nodes, to be used as a passage in a path towards a target node . In fact, the two incoming links (2-1) and (3-1) have an average use of 0.5% in all three implementations of the algorithm, except at 2000 PU, in the case of the case B, where the The use of (3-1) sees an increase of 1%, mainly due to the dynamic distribution of the band. On the contrary, node “6”, although more central in the network, is in turn connected to only two non-target nodes; therefore follows the same type of reasoning. As can be

expected, however, on average the two connections (5-6) and (9-6) have a higher use than the first two, of 8.95% and 6.4% respectively, being still central to the network. In fact, they were selected because, in the reference configuration of the 2000 PU simulator (Fig.7.7), they saw 4.3% and 2.8% of use. The reason for the higher average is found once again in the case B, where is possible to see, compared to the reference condition, an increase in the percentage respectively of 13% and 5%, again thanks to the distribution of bandwidth (fig. 7.6). The remaining five links were taken into consideration, as links entering the target node, due to the large presence of slices active on average throughout the duration of the simulation, which translates into an average use of bandwidth for high, if not critical, links. The lower limit of 50% has been set to monitor, in the case A and B, if it becomes lighter or heavier, again with respect to the reference condition at 2000 PU with a 200 Erlang load. After the case A, it's possible to see for the link (2-4) a use of less than 5-6%, due to the fact that being marginal and having few outgoing connections (three), node "2" ended up with 1111 Gbps to be shared between its links, compared to the reference 1000 Gbps. The remaining links have instead all undergone an increase in the percentage of total use, in some cases exceeding 90%, undergoing a 5% increase compared to the reference condition, up to 94.3%, as in the case of the link (8-11) to 2000 PU. This can be identified in the fact that all the remaining nodes to be analyzed are nodes with four outgoing links, therefore the value shared between them drops from 1000 Gbps to 833.5 Gbps per link.

It is important to note that these considerations, together with the previous ones on nodes "1" and "6", following the case A, gave the idea of unequally distributing the bandwidth between links entering target and non-target nodes. target, to prevent the former from reaching critical levels and, at the same time, recovering unused bandwidth from the latter. In fact, in the case B, all the connections taken into consideration with use on average greater than 50%, which were recorded have almost touched the total use of the available resources, with the new distribution now remain below 40%, reaching the maximum to 36.5%.
Finally, the probability of system blocking is to be analyzed. In the case with 500 PUs provided (fig. 7.8), it can be seen that, between the

reference condition and the case A, the probability remains relatively unchanged. On the other hand, there is a deterioration in reliability after the case B in the case of lighter loads, but by increasing the latter the values begin to converge. Combining these results with those of the average band per link, it can be identified that, at high loads, the performance of the case B is slightly better, since it achieves a saving on bandwidth usage, while maintaining a similar blocking probability.

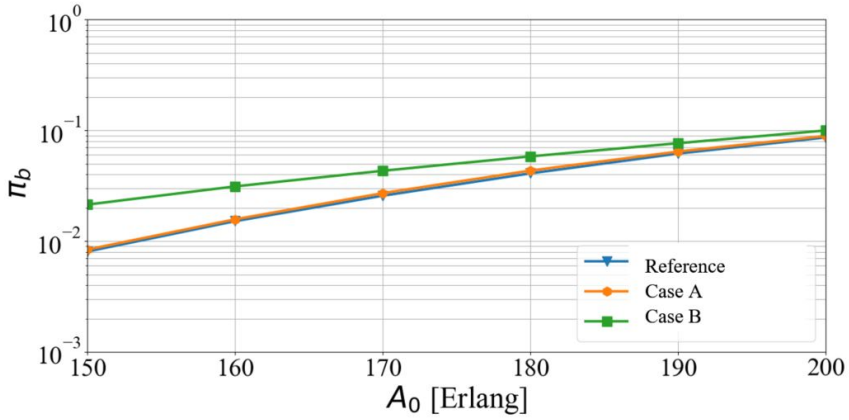


Figure 7.8: Probability of blocking depending on the load, case 500 PU.

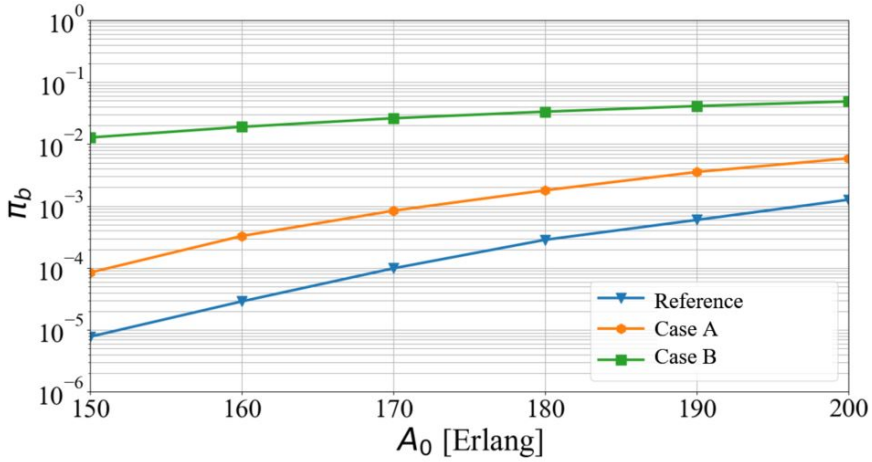


Figure 7.9: Probability of blocking as a function of the load, case 2000 PU.

On the other hand, in the case where 2000 PUs are available at the

target nodes, the situation changes considerably (fig. 7.9). The reference condition sees the probability of lower block, similar to that recorded in the experiments prior to this work, in particular at low load. The first change records a significant loss compared to the reference condition, due to the burden of work on the already critical links highlighted in the previous lines. Finally, the probability of blocking after the case B, even with processing units in larger quantities, remains very similar to the previous case. This is because it has been recorded that some links entering central non-target nodes, such as "8", which previously maintained a usage level of 40-45%, with the variation on distribution have maintained the same content, but receiving 36, 5% -52.5% less bandwidth (538.5 Gbps compared to 833.5 Gbps / 1111 Gbps, in cases with 4 or 3 outgoing links from the node), thus filling up much faster.

7.2 Management of dynamic network resources: Processing Units

To study the dynamic management of the processing units in the destination nodes, the same simulator of the previous section was used. The processing units are assigned based on the nodal degree of the target node, then based on the number of links entering the target node. The greater the number, the greater the tributary links, and the greater the amount of processing units assigned. If the total number of units to be inserted into the network is P_{tot}, and L_tot is the number of total links entering all target nodes, the amount of resources to be partitioned for each link is equal to P_{part}:

$$P_{part} = \frac{P_{tot}}{L_{tot}} \tag{7.5}$$

If a node has a nodal degree equal to 2, this node will be assigned 2 * P_{part}. If the nodal degree is 4, double (4 * P_{part}) will be assigned.

7.2.1 Scenario and Results

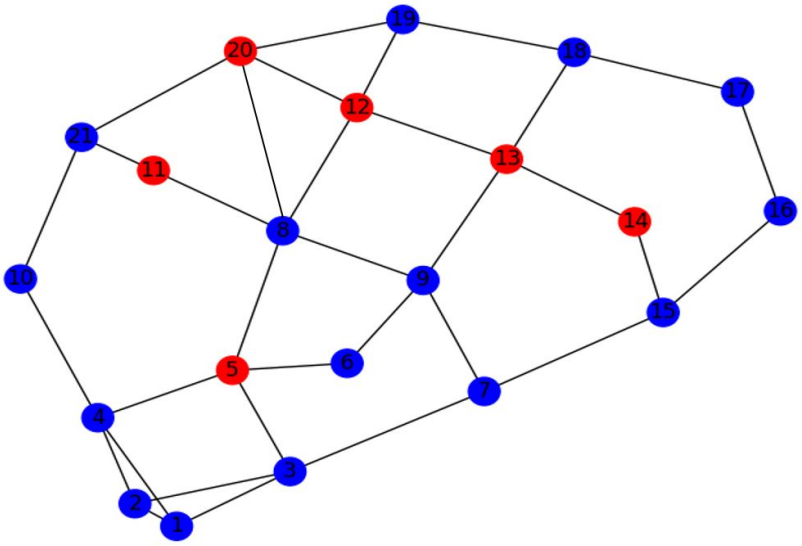


Figure 7.10: Reference network with source (blue) and target (red) nodes.

The reference network is the one in the figure 7.10. The blue nodes are source nodes and red nodes are target nodes. Having two target nodes with a nodal degree equal to 2 and 4 target nodes with a nodal degree equal to 4, there will be respectively 1200 PU available for the first and 2400 PU for the second. In the reference conditions, the PUs are instead distributed equally, so each target node will have 2000 PUs at its disposal. the amount of bandwidth assigned to each node is 1000Gbps. The rest of the network settings are equivalent to the model presented in the chapter 6. The comparison in between shared protection and dedicated protection in a reference case and in dynamic case.

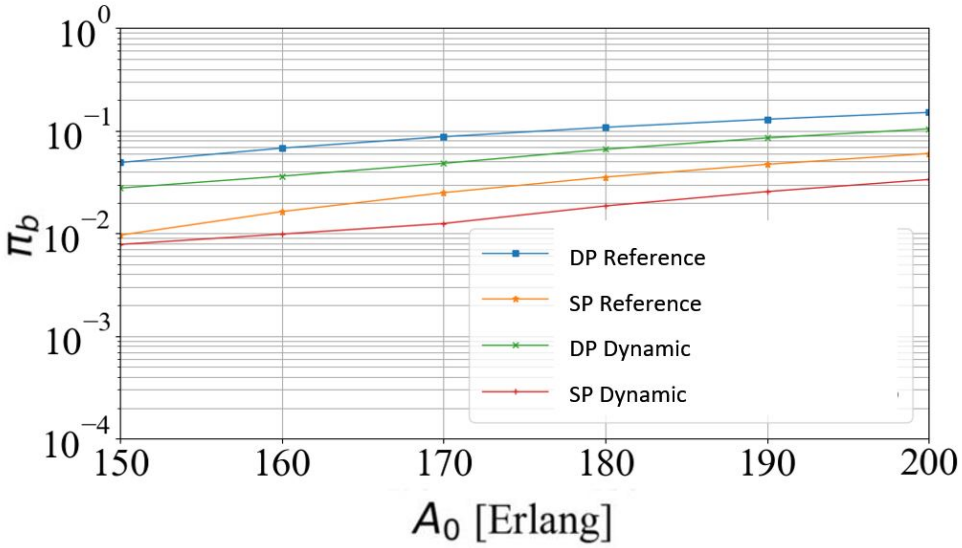


Figure 7.11: Probability of blocking depending on the load.

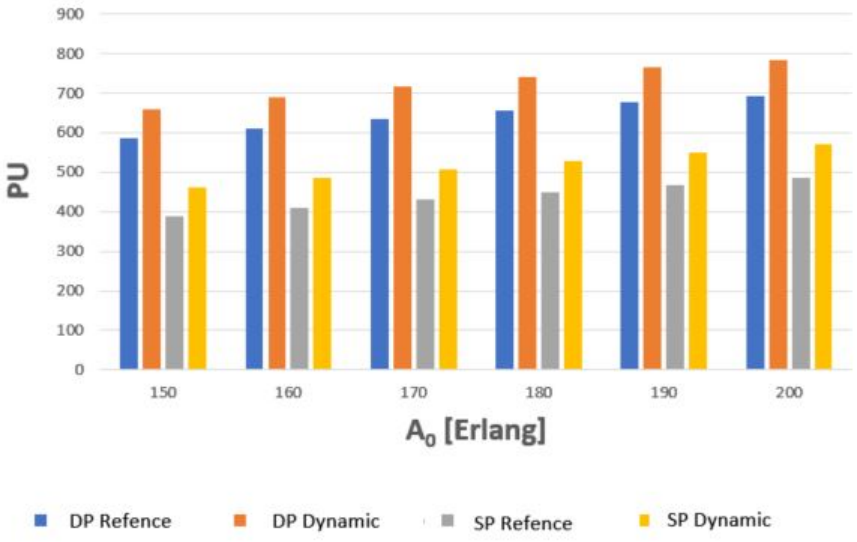


Figure 7.12: Average PU used per node.

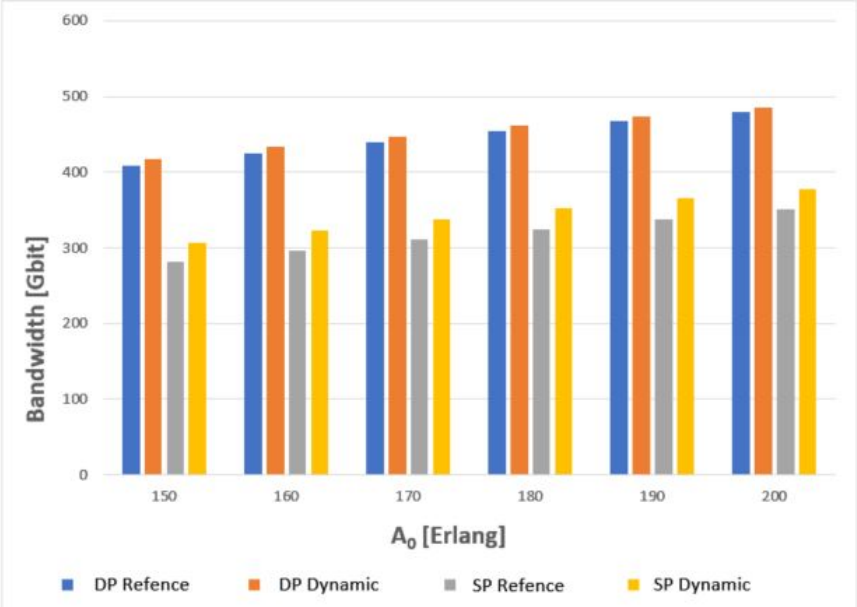


Figure 7.13: Average bandwidth used per node.

As can be seen in fig. 7.11, the probability of blocking is lower in the dynamic case than in the reference case and this means that fewer slices in the network are rejected. This result falls on the number of active PUs in the nodes which, as can be observe from fig 7.12, there is an increase in the dynamic case, and in the amount of average bandwidth used in the links, which in turn increases in the dynamic case, as can be seen in fig 7.13.

The comparison between the two configuration scenarios denotes how the network in the dynamic case is more efficient than in the reference case, this leads to the conclusion that assigning a different maximum number of PUs to the nodes according to the number of links connected to them leads to an evident improvement in the use of bandwidth in the links and to a greater number of PUs active in the nodes.

7.3 Reconfiguration

Reconfiguration is a strategy that allows you to reallocate backup flows within the network, to lead to better optimization. This strategy is usually applied to the triggering of a condition, a policy that is set according

to the needs of the network. The reconfiguration techniques are therefore created to be able to respond to the problems of a network, from the probability of blocking, to the saturation of resources in a part or in the whole network.
Usually a reconfiguration is applied in case the provisioning technique is softer, and does not foresee the calculation time of the best allocation. It was therefore decided to introduce a new provisioning, easier to implement and more immediate in the allocation: the Shortest Path First Fit (SPFF). This type of provisioning allows to allocate the primary and backup in the shortest available paths. If Hop is the number of nodes that a path must cross to reach the target node, the path chosen will be the one with the least number of hops available. Since provisioning is not aimed at optimizing network performance, but only at optimizing provisioning times, network performance cannot exceed those of exhaustive provisioning (such as that shown in the chapter 6. For the following result, the nodes have 2000 PUs available, the links have 1000 GBPS available, the network is the one in fig 6.1 in chapter 6. The probability of blocking shown in the fig. 7.14 relates the two types of dedicated and shared protection in the case of comprehensive provisioning or SPFF, and it is noted how the performance of the SPFF in the case of dedicated resources remains the same as that of exhaustive provisioning. This result arises as a consequence of the choice of the cost function in the exhaustive case: since the search is tempted to look for the path with the lowest cost for the network (see chapter 6), the paths that are chosen in both provisioning are the same. In the case of resource sharing, however, it is possible to see an increased of blocking probability.
To improve performance, especially in the case of resource sharing, reconfiguration has been introduced. Two methodologies are under study: an ILP and a heuristic. The implemented heuristics works as an exhaustive search, where the goal of the backup allocation is to find the lowest cost primary-backup pair, where the cost depends only on the backup and is defined as:

$$cost^b_{i,j} = \alpha * \sum_{i \in Links} Band^b_i + \beta * \sum_{j \in Nodes} Compute^b_j \tag{7.6}$$

where α and β are coefficients used to balance the two contributions,

$Band_i^b$ is the connectivity requirement (in Gbps) of the backup slice and $Compute_j^B$ is the PU required by the backup slice for the virtual baseband and the service. Since the primary is already allocated and unmovable, the only possibility is to find a backup that allows the sharing of network resources. As a search by slice, the results and behavior of this heuristic depend on when the search is started and the order in which the slices are reallocated. It is possible to see in the tab 7.1 the bandwidth and PUs saved on average. Thanks to the reconfiguration, it is possible to reduce backup resources by up to 16%, and PUs by up to 4%.

Table 7.1: Band and PU saved in SP protection using reconfiguration.

Load	Band Saved	PU Saved
150	16.45%	1.60%
170	16.85%	2.52%
200	17.34%	4.14%

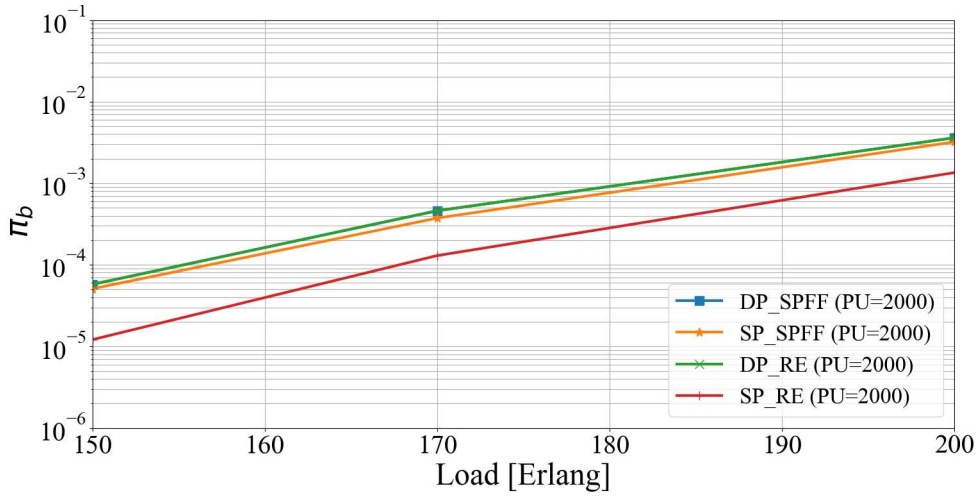


Figure 7.14: Probability of blocking depending on the load. RE refers to Exhaustive Research, SPFF refers to Shortest Path Best Fit.

The ILP is still in study, but the model has been made. The ILP takes as input the nodes traversed by the primary path for each slice ($p_{s,n}$), a set of candidate target nodes (D) to place the backup compute resources, a set of candidate backup paths (K_s) for each slice (there may be more

than one candidate path for each target node, pre-computed using the k-shortest path and disjoint from the primary path), the maximum amount of compute and bandwidth resources per node and link respectively. Tabs 7.2 and 7.2 shown the list of variables and parameters.

Table 7.2: List of cost parameters.

	Parameters:
S	Set of slices.
N	Set of the network nodes.
L	Set of links.
$p_{s,n}$	1 if the primary path for slice slice $s \in S$ traverses node $n \in N$, 0 otherwise.
K_s	Set of backup candidate paths for each $s \in S$. $K = \bigcup_{s \in S} K_s$. Each path $k \in K_s$ is disjoint from primary path used by slice $s \in S$ Max hops length already satisfied.
$a_{k,j}$	1 if target node $j \in N$ is assigned as backup for path $k \in K_s$, 0 otherwise.
$e_{k,l}$	1 if link $l \in L$ is assigned as backup for path $k \in K_s$, 0 otherwise.
z_s	Element of set N that corresponds to the source node of each slice $s \in S$.
MC_j	Maximum amount of compute available at node $j \in N$.
MW_l	Maximum bandwidth available at the link $l \in L$.
r_s	Number of PU for slice $s \in S$.
bw_s	Bandwidth requirement for slice $s \in S$.
α	Activation cost for the PU.
β	Activation cost for the bandwidth.

Table 7.3: List of cost variables.

	Variables:
$b_{s,k}$	1 if slice $s \in S$ is using path $k \in K_s$ as backup, 0 otherwise.
y_j	Number of PU required at node $j \in N$ for backup purposes.
x_l	Amount of bandwidth required at link $l \in L$ for backup purposes.
c_{snj}	1 if slice $s \in S$ is using node $n \in N$ as primary and $j \in N$ as backup target node; 0 otherwise.
d_{snl}	1 if slice $s \in S$ is using node $n \in N$ as primary and $l \in L$ as backup link; 0 otherwise.

Objective function:

$$Minimize\ \alpha \cdot \sum_{j \in D} y_j + \beta \cdot \sum_{l \in L} x_l \tag{7.7}$$

Constraints:

$$\sum_{k \in K_s} b_{s,k} = 1, \forall s \in S \tag{7.8}$$

$$y_j \leq MC_j, \forall j \in D \tag{7.9}$$

$$c_{snj} \geq p_{s,n} + b_{s,k} \cdot a_{k,j} - 1, \forall k \in K_s, s \in S, j \in D, n \in N, n \neq j, n \neq z_s \tag{7.10}$$

$$y_j \geq \sum_{s \in S} c_{snj} \cdot r_s, \forall j \in D, n \in N, n \neq j, n \neq z_s \tag{7.11}$$

$$x_l \leq MW_l, \forall l \in L \tag{7.12}$$

$$d_{snl} \geq p_{s,n} + b_{s,k} \cdot e_{k,l} - 1, \forall k \in K_s, s \in S, n \in N, n \neq z_s, l \in L \tag{7.13}$$

$$x_l \geq \sum_{s \in S} d_{snl} \cdot bw_s, \forall n \in N, n \neq z_s, l \in L \tag{7.14}$$

The objective function 7.7 is used to minimize the amount of bandwidth and PU used for backup path, exploiting the sharing of resources to obtain a better result.
The constraint of Equation (7.8) serves to guarantee one backup for each slice.
The constraint of Equation (7.9) limits the amount of compute in each target nod.
The constraints of Equations (7.10) and (7.11) count the amount of backup compute with sharing, where $n \in N$ is a node, $p_{s,n}$ is 1 if primary path for slice s includes node n. c_{snj} is 1 if slice s is traversing node n with the primary path and terminates its backup path at node j.
The constraint of Equation (7.12) limits the amount of bandwidth in each link, where x_l amount of bandwidth used over link l for backup purposes.
The constraints of Equations (7.13) and (7.14) count the amount of backup bandwidth with sharing, where d_{snl} is 1 if slice s is traversing node n with the primary path and includes link l in the backup path. $a_{k,l}$ 1 if link $l \in L$ is used in the backup path $k \in K_s$, 0 otherwise.
ILP is still under study and results will be available soon.

Chapter 8

Conclusions

This thesis highlights the results obtained during three years of research during the Ph.D program. Vehicle network optimization techniques were proposed to support the use cases required by the automotive sector. In chapter 3 an optimization of functional split for URLLC service has been presented for DPP and SPP based on Integer Linear Programming. Two different sets of constraints have been defined with the objective to minimize the number of active nodes in the optical aggregation networks. The effectiveness of the algorithms has been shown also in terms of active nodes and bandwidth usage which is sensibly reduced with repsect to the conventional centralized approach and allows statistical multiplexing gain for potential allocation of multiple slices. The further saving related to SPP has been shown in comparison with DPP. The SPP algorithm has some scalability limitations that, at this moment has reached optimization for a more limited size network with respect to DPP. In any case the evaluations result suitable for most metro contexts.
In chapter 4 another optimization model is defined for differentiated reliability in URLLC and eMBB slices. Resource savings has been shown with respect to the fully dedicated protection scheme, with 13% of PU saving. In addition, the sequential methodology allows better scalability while achieving bandwidth saving between 25 and 30 % using the AV strategy. Almost the same saving can be achieved with the ACT strategy, which also saves links, providing that the SPP is applied first. Further investigations for larger networks can better show the effectiveness of the approach and its scalability.

Chapter 5 addresses the problem of providing low latency and reliable services in vehicular scenarios in a cost-efficient way using 4G and 5G networks. Baseband resources of C-RAN can be co-located with MEC resources to achieve target service requirements. An ILP model for the cost-efficient deployment of baseband and edge cloud resources with reliability against single node failure is proposed. In addition, a heuristic technique is also proposed to reduce computational complexity of the ILP model by proper selection of a subset of edge nodes for the optimization phase. Results show that the hybrid approach provides similar results to the ILP ones while considerably reducing the solving time.
In chapter 6 presented a performance comparison between dedicated and shared protection schemes for dynamic slice provisioning in a 5G metro network context where processing resources per node are typically scarce. Results show that, especially in these conditions, the SP leads to considerably lower blocking probability than DP. SP is shown to save up to 37% PUs and 28% bandwidth per slice with respect to DP.
In chapter 7 other dynamic resource optimization techniques are shown. In particular, it was shown how the different and dynamic management of network resources impacts on network performance. The available bandwidth and available computational resources must be intelligently distributed within the network to achieve better performance. Furthermore, a reconfiguration, in the case of using simple provisioning, can be an excellent resource for obtaining a greater saving of resources, in order to allocate other slices in the network.
The works presented offer an important support to research, to optimize vehicle networks and prepare the 5G network in an appropriate way, to obtain high reliability without incurring waste of network resources. The prospects in this research field are excellent, and it may be interesting to further explore different optimization techniques, such as Machine Learning (ML) and Deep Learning (DL).

Acronyms

3GPP Third Generation Partnership Project.

BBU Baseband Unit.

BS Base Station.

C-RAN Centralized Radio Access Network.

C-V2X Cellular V2X.

CPRI Common Public Radio Interface.

CU Centralized Unit.

DPP Dedicated Path Protection.

DU Distributed Unit.

eMBB Enhanced Mobile Broadband.

ILP Integer Linear Programming.

IoT Internet of Things.

LTE Long Term Evolution.

MAC Medium Access Control.

mMTC Massive Machine Type Communication.

NFV Network Function Virtualization.

RRU Remote Radio Unit.

RU Radio Unit.

SDN Software-Defined Network(ing).

SPP Shared Path Protection.

URLLc Ultra Reliable Low Latency Communication.

V2I Vehicle to infrastructure.

V2N Vehicle to network.

V2P Vehicle to pedestrian.

V2V Vehicle to vehicle.

V2X Vehicle to everything.

VEC Vehicular Edge Computing.

VRU Vulnerable Road User.

Bibliography

[B1] *TR 21.914 V14.0.0*. Tech. rep. 2018.

[B2] SAMSUNG. *The Next Hyper-Connected Experience for All*. Tech. rep. 2020.

[B3] GSMA. *Connecting vehicles today and in the 5G era with C-V2X*. Tech. rep. London, UK, 2019.

[B4] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeadally. "5G for Vehicular Communications". In: *IEEE Communications Magazine* 56.1 (2018), pp. 111–117.

[B5] H. Zolfaghari, D. Rossi, W. Cerroni, H. Okuhara, C. Raffaelli, and J. Nurmi. "Flexible Software-Defined Packet Processing Using Low-Area Hardware". In: *IEEE Access* 8 (2020), pp. 98929–98945.

[B6] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina. "V2X Access Technologies: Regulation, Research, and Remaining Challenges". In: *IEEE Communications Surveys Tutorials* 20.3 (2018), pp. 1858–1877.

[B7] *TR 22.186 v16.0.0*. Tech. rep. 2018.

[B8] L. M. P. Larsen, A. Checko, and H. L. Christiansen. "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks". In: *IEEE Communications Surveys Tutorials* 21.1 (2019), pp. 146–172.

[B9] P. Q. e. a. Liu L. Chen C. "Vehicular Edge Computing and Networking: A Survey". In: (2020).

[B10] 3. G. P. Project. *Technical Specification Group Services and System Aspects; Summary of Rel-14 Work Items*. Tech. rep. 2018.

[B11] E. T. S. Institute. *Intelligent Transport Systems (ITS); Access Layer Specification for Intelligent Transport Systems Operating in The 5 GHz Frequency Band*. Tech. rep. 2012.

[B12] Z. Ning, X. Wang, and J. Huang. "Mobile Edge Computing-Enabled 5G Vehicular Networks: Toward the Integration of Communication and Computing". In: *IEEE Vehicular Technology Magazine* (2019).

[B13] *eCPRI V2.0 Specification*. Tech. rep. 2019.

[B14] D. M. M. Pioro. *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Morgan Kaufmann, 2004.

[B15] A. Hmaity, M. Savi, F. Musumeci, M. Tornatore, and A. Pattavina. "Protection strategies for virtual network functions placement and service chains provisioning". In: *Networks* 70.4 (2017), pp. 373–387.

[B16] *Small Cell Forum - Functional splits and use cases*. 2016.

[B17] *IBM ILOG CPLEX Optimization Studio V12.6.3*.

[B18] D. Harutyunyan and R. Riggio. "Flex5G: Flexible Functional Split in 5G Networks". In: *IEEE Transactions on Network and Service Management* 15.3 (2018), pp. 961–975.

[B19] H. Chang, B. Qiu, C. Chiu, J. Chen, F. J. Lin, D. de la Bastida, and B. P. Lin. "Performance evaluation of Open5GCore over KVM and Docker by using Open5GMTC". In: *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*. 2018, pp. 1–6.

[B20] B. M. Khorsandi, D. Colle, W. Tavarnier, and C. Raffaelli. "Adaptive function chaining for efficient design of 5G Xhaul". In: *2019 International Conference on Optical Network Design and Modeling ONDM*. 2019.

[B21] J. Zhang, K. Zhu, H. Zang, N. S. Matloff, and B. Mukherjee. "Availability-Aware Provisioning Strategies for Differentiated Protection Services in Wavelength-Convertible WDM Mesh Networks". In: *IEEE/ACM Transactions on Networking* 15.5 (2007), pp. 1177–1190.

[B22] F. Tonini, B. M. Khorsandi, E. Amato, and C. Raffaelli. "Scalable Edge Computing Deployment for Reliable Service Provisioning in Vehicular Networks". In: *Journal of Sensor and Actuator Networks* 8.4 (2019).

[B23] C.-L. I, H. Li, J. Korhonen, J. Huang, and L. Han. "RAN Revolution With NGFI (xhaul) for 5G". In: *Journal of Lightwave Technology* 36.2 (2018), pp. 541–550.

[B24] F. Tonini, E. Amato, and C. Raffaelli. "Optimization of Optical Aggregation Network for 5G URLLC Service". In: *2019 IEEE Global Communications Conference (GLOBECOM)*. 2019, pp. 1–6.

[B25] *Global Status Report on Road Safety 2018*. Tech. rep. 2018.

[B26] K. I. Mueck M. *Networking Vehicles to Everything: Evolving Automotive Solutions*. De Gruyter, 2018.

[B27] H. Deliverable D1.2. *5G-TRANSFORMER Initial System Design–Project Grant No. 761536*. Tech. rep. 2018.

[B28] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann. "Cloud RAN for Mobile Networks—A Technology Overview". In: *IEEE Communications Surveys Tutorials* 17.1 (2015), pp. 405–426.

[B29] F. van Lingen, M. Yannuzzi, A. Jain, R. Irons-Mclean, O. Lluch, D. Carrera, J. L. Perez, A. Gutierrez, D. Montero, J. Marti, R. Maso, Rodriguez, and J. Pedro. "The Unavoidable Convergence of NFV, 5G, and Fog: A Model-Driven Approach to Bridge Cloud and Edge". In: *IEEE Communications Magazine* 55.8 (2017), pp. 28–35.

[B30] P. Öhlén, B. Skubic, A. Rostami, M. Fiorani, P. Monti, Z. Ghebretensaé, J. Mårtensson, K. Wang, and L. Wosinska. "Data Plane and Control Architectures for 5G Transport Networks". In: *Journal of Lightwave Technology* 34.6 (2016), pp. 1501–1508.

[B31] J. C. Nobre, A. M. de Souza, D. Rosário, C. Both, L. A. Villas, E. Cerqueira, T. Braun, and M. Gerla. "Vehicular software-defined networking and fog computing: Integration and design principles". In: *Ad Hoc Networks* 82 (2019), pp. 172–181.

[B32] E. T. S. Institute. *Cloud RAN and MEC: A Perfect Pairing*. Tech. rep. 2018.

[B33] B. M. Khorsandi, F. Tonini, and C. Raffaelli. "Design methodologies and algorithms for survivable C-RAN". In: *2018 International Conference on Optical Network Design and Modeling (ONDM)*. 2018, pp. 106–111.

[B34] *TR 38 801*. Tech. rep. 2017.

[B35] E. T. S. Institute. *MEC in 5G Networks*. Tech. rep. 2018.

[B36] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines. "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges". In: *Computer Networks* 167 (2020), p. 106984.

[B37] X. Zhou, R. Li, T. Chen, and H. Zhang. "Network slicing as a service: enabling enterprises' own software-defined cellular networks". In: *IEEE Communications Magazine* 54.7 (2016), pp. 146–153.

[B38] F. Tonini, C. Natalino, M. Furdek, C. Raffaelli, and P. Monti. "Network Slicing Automation: Challenges and Benefits". In: *2020 International Conference on Optical Network Design and Modeling (ONDM)*. 2020.

[B39] A. Marotta, D. Cassioli, M. Tornatore, Y. Hirota, Y. Awaji, and B. Mukherjee. "Reliable Slicing with Isolation in Optical Metro-Aggregation Networks". In: *2020 Optical Fiber Communications Conference and Exhibition (OFC)*. 2020, pp. 1–3.

[B40] N. Shahriar, S. Taeb, S. R. Chowdhury, M. Zulfiqar, M. Tornatore, R. Boutaba, J. Mitra, and M. Hemmati. "Reliable Slicing of 5G Transport Networks With Bandwidth Squeezing and Multi-Path Provisioning". In: *IEEE Transactions on Network and Service Management* 17.3 (2020).

[B41] H. D. Chantre and N. L. Saldanha da Fonseca. "The Location Problem for the Provisioning of Protected Slices in NFV-Based MEC Infrastructure". In: *IEEE Journal on Selected Areas in Communications* 38.7 (2020).

[B42] B. M. Khorsandi, F. Tonini, and C. Raffaelli. "Centralized vs. distributed algorithms for resilient 5G access networks". In: *Photonic Network Communications* 37.3 (2019).

Publications

[P1] B. M. Khorsandi, F. Tonini, E. Amato, and C. Raffaelli. "Dedicated Path Protection for Reliable Network Slice Embedding Based on Functional Splitting". In: *2019 21st International Conference on Transparent Optical Networks (ICTON)*. 2019, pp. 1–4.

[P2] F. Tonini, E. Amato, and C. Raffaelli. "Optimization of Optical Aggregation Network for 5G URLLC Service". In: *2019 IEEE Global Communications Conference (GLOBECOM)*. 2019, pp. 1–6.

[P3] F. Tonini, B. M. Khorsandi, E. Amato, and C. Raffaelli. "Scalable Edge Computing Deployment for Reliable Service Provisioning in Vehicular Networks". In: *Journal of Sensor and Actuator Networks* 8.4 (2019). URL: https://www.mdpi.com/2224-2708/8/4/51.

[P4] E. Amato, F. Tonini, and C. Raffaelli. "Differentiated Protection in 5G Vehicular Networks". In: *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*. 2020, pp. 1–6.

[P5] E. Amato, F. Tonini, C. Raffaelli, and P. Monti. "A Resource Sharing Method for Reliable Slice as a Service Provisioning in 5G Metro Networks". In: *t 2021 International Conference on Optical Network Design and Modelling (ONDM)*. 2021.