Università degli Studi di Padova

Department of Information Engineering

Ph.D. in Information Engineering

Science and Information Technology Curriculum

XXXIV Series

Design and Evaluation of Models, Algorithms, and Architectures for Next-generation Cellular Networks

Coordinator
Prof. Andrea Neviani

Supervisor
Prof. Michele Zorzi

Ph.D. Candidate
Tommaso Zugno

Academic year 2020 / 2021

Abstract

The always-increasing number of mobile subscribers, the growing demand for mobile data, and the emergence of new applications require cellular systems to be constantly improved. The last generation of cellular networks, i.e., 5G, stands out for its high performance and extreme flexibility, making it possible to support multiple use cases with diverse and stringent service requirements. One of the main novelties is represented by the possibility to communicate at millimeter wave (mmWave) frequencies, providing access to an unprecedented amount of radio resources which can theoretically enable extremely high data rates. However, signals propagating at these frequencies experience harsh conditions, posing several challenges for the realization of efficient mmWave cellular systems.

The grand objective of this thesis is to provide innovative solutions to overcome the limitations of mmWave communications and exploit the potential of this technology in the context of 5G and beyond cellular networks. In particular, we (i) present novel simulation tools, including channel, antenna, and beamforming models for the accurate characterization of next-generation cellular systems; (ii) identify the potential and challenges for the realization of wireless-backhauled mmWave deployments, and present a semi-centralized resource partitioning scheme for this type of networks; (iii) analyze the cross-layer challenges arising from the integration of Hybrid Beamforming (HBF) and Multi User MIMO (MU-MIMO) in mmWave cellular systems; (iv) introduce a novel framework to enable network slicing in mmWave Radio Access Networks (RANs); and (v) evaluate the feasibility of providing vehicular communication services by means of mmWave communications. We adopt a system-level approach that allow us to properly characterize the network behavior, considering the full protocol stack and all the elements that have an impact on the performance of the end-users. Our results demonstrate the effectiveness of the proposed solutions, breaking new ground towards more efficient and high-performance mmWave cellular systems.

Sommario

A causa del sempre maggior numero di utenti, della crescente domanda di dati mobili e della nascita di nuove applicazioni, le reti cellulari necessitano di costante aggiornamento. L'ultima generazione di reti mobili, le reti 5G, è caratterizzata da elevate prestazioni ed estrema flessibilità, grazie alle quali è possibile supportare vari casi d'uso con diversi requisiti di servizio. Le comunicazioni a frequenze millimetriche rappresentano una delle principali novità dello standard 5G, perché consentono l'utilizzo di una vasta porzione di risorse radio ed il raggiungimento di elevate velocità di trasmissione. Tuttavia, la realizzazione di sistemi cellulari operanti a tali frequenze è soggetta a numerose problematiche che derivano dalle severe condizioni di propagazione dei segnali radio.

Questa tesi si pone l'obiettivo di fornire soluzioni innovative per risolvere le problematiche realizzative e sfruttare appieno i benefici di questa tecnologia. Nello specifico, (i) vengono presentati nuovi strumenti per la simulazione delle reti di prossima generazione, tra cui un modello di canale e modelli per la caratterizzazione delle antenne e delle operazioni di beamforming; (ii) vengono identificati i benefici e le problematiche relative alla realizzazione di reti millimetriche con backhaul senza fili e viene presentato un efficiente meccanismo di ripartizione delle risorse; (iii) viene analizzata l'interazione cross-layer che deriva dall'utilizzo congiunto di soluzioni HBF e MU-MIMO; (iv) viene introdotto un sistema innovativo per l'implementazione del paradigma di network slicing all'interno di una rete di accesso a frequenze millimetriche e, infine, (v) viene valutata la possibilità di supportare servizi di comunicazioni veicolare attraverso comunicazioni a frequenze millimetriche. L'approccio di tipo "system-level" utilizzato in questa tesi permette di caratterizzare il comportamento della rete in modo adeguato, prendendo in considerazione l'intero stack protocollare e tutti gli elementi che influenzano le prestazioni degli utenti finali. I risultati ottenuti dimostrano l'efficacia delle soluzioni proposte, aprendo nuove strade per la realizzazione di reti cellulari più efficienti e performanti.

Contents

1 Introduction

Over the past two decades, the increase of mobile subscriptions and the growth of data traffic generated by mobile devices triggered the need for more efficient and versatile cellular networks, and this trend is expected to continue in the future. Indeed, as reported in the latest Ericsson Mobility Report [1], the number of unique subscriptions is expected to grow from 5.9 billions in 2021 to 6.5 billions in 2026. 76% of the subscribers correspond to smartphone devices, each generating an average traffic volume of more than ∼10 GB/month. In 5 years, this value is expected to reach 35 GB/month. Currently, the main source of mobile traffic is video streaming, which accounts for 66% of the overall data consumption, but emerging use cases, such as virtual and augmented reality, Industrial Internet of Things (IIoT), and autonomous vehicles are expected to drive the trend in the next years.

Such increasing demand for mobile data traffic and the need to support new use cases require cellular systems to be continuously updated and improved. The latest generation of cellular networks, i.e., 5G, has been designed to support [2]:

- peak data rate of 10 Gbit/s and up to 20 Gbit/s in certain conditions and scenarios;
- user experienced data rate of 100 Mbit/s in wide area coverage cases and even higher in hotspot cases;

- spectrum efficiency three times higher compared to Long Term Evolution (LTE)-Advanced networks;
- area traffic capacity up to 10 Mbit/s/m^2;
- energy consumption not higher than current networks;
- over-the-air latency of 1 ms;
- high mobility, up to 500 km/h with acceptable Quality of Service (QoS);
- connection density of up to 10^6 devices/km^2.

Moreover, 5G is required to handle several use cases with diverse communication requirements, which may possibly coexist within the same network. They can be divided into three main categories [2]:

- Enhanced Mobility Broadband (eMBB), which includes human-centric applications aiming at providing fast and seamless access to media contents, services, and data (e.g., virtual and augmented reality, and video streaming);
- Ultra-Reliable Low-Latency Communication (URLLC), including use cases that require the support of critical services with very stringent requirements in terms of latency and reliability (e.g., wireless control of industrial manufacturing or production processes, remote medical surgery, transportation safety);
- Machine Type Communications (MTC), which includes use cases involving a large number of connected devices characterized by sporadic transmissions and low data volumes, but with stringent requirements in terms of costs and battery life (e.g., monitoring of industrial processes, smart grids, health data collection).

Notably, these use cases involve different types of devices other than smartphones, such as vehicles, sensors, and wearables. Differently from the previous generations, 5G networks have to provide seamless and ubiquitous connectivity to both humans and machines [3, 4].

Over the last decade, researchers and the industry have worked to meet the target requirements and address the emerging market trends. Among the many contributions that have been provided, we can identify four main research directions that have shaped the design of the 5G technology [5]:

- **mmWave communications**
  The high capacity required to support the envisioned use cases triggered the need for new spectrum resources, because the frequency bands used by previous generations systems are limited and already saturated. mmWave communications solve this issue by enabling the usage of new portions of the spectrum, roughly between 30 and 300 GHz *, which are currently underutilized [6, 7].

- **Densification of cellular deployments**
  Another way to increase the capacity of cellular systems is through network densification, defined as a combination of spatial densification (i.e., increase the number of cell sites) and spectral aggregation (i.e., by exploiting the availability of multiple frequency bands) [8]. While 4G deployments usually have a flat network structure with a single layer of macro cells, 5G networks can assume hierarchical structures with multiple layers of macro and small cells operating at different frequencies and with different coverage capabilities.

- **Massive Multiple Input, Multiple Output (MIMO)**
  Massive MIMO makes it possible to improve the spectral efficiency of cellular systems, therefore increasing the throughput and reducing the radiated power [9]. In a Massive MIMO system, base stations and user terminals are equipped with antenna arrays made of multiple, physically small, radiating elements and can exploit suitable beamforming techniques to achieve the spatial multiplexing of the radio resources.

- **Network softwarization**
  5G systems are required to be flexible enough to support multiple services with diverse communication requirements. This flexibility can be achieved

*The mmWave spectrum was originally defined as the band between 30 and 300 GHz, however, the industry loosely considers as mmWave any band above 10 GHz [6]

through the virtualization of the network infrastructure (referred to as network function virtualization), by replacing dedicated hardware with virtualized instances running on generic hardware. Moreover, software defined networking paradigms enable an easy and dynamic configuration of the network, by deploying new network function instances in real time [10]. Finally, network slicing allows operators to instantiate multiple virtual networks on top of the same physical infrastructure, allowing multiple services to be delivered by the same system simultaneously [11].

To merge the newly developed solutions within a unique communication standard, the 3rd Generation Partnership Project (3GPP), i.e., the main standardization body for cellular networks, carried out a significant effort which resulted in the publication of a new set of technical specifications referred to as 5G NR [12]. The effort started in 2015 and can be divided into two main phases. The first phase (Release 15) ended in March 2019 with the release of the first set of 3GPP specifications for 5G. With the second phase (Release 16), which ended in July 2020, the 3GPP completed the specifications for a full 5G system. Currently, the 3GPP is working on Releases 17 and 18, which will put the basis for 5G-Advanced, the evolution of 5G.

1.1 5G Cellular Networks

5G NR defines a new Radio Access Technology (RAT), featuring a new air interface and a new RAN architecture, referred to as Next Generation Radio Access Network (NG-RAN) [12].

While still based on Orthogonal Frequency Division Multiplexing (OFDM), the air interface has been completely redesigned and includes several novelties which make 5G much more flexible, scalable, and efficient with respect to 4G LTE [13]. One of the key features is the support to operations at mmWave frequencies, between 24.25 and 52.6 GHz, other than at traditional sub-6 GHz frequencies, between 410 MHz and 7.125 GHz [14]. The high spectrum availability provides the possibility to use very wide transmission bandwidths, up to 400 MHz per carrier, which can be further increased by means of Carrier Aggregation (CA) [15]. Moreover, the bandwidth can be adapted through a novel adaptation mechanism, allowing the User Equipment (UE) to use just a portion of the overall system

bandwidth. Also, NR enables a flexible configuration of the Physical (PHY) layer numerology by selecting one of the five predefined settings, each providing a different sub-carrier spacing (ranging from 15 to 240 kHz) and frame structure (from 1 to 16 slots per subframe) [16]. To support low latency applications, it introduces the concept of "mini-slot," which enables transmissions over a portion of the slot [17], and of "dynamic TDD," which leaves to the scheduler the possibility to dynamically assign resources for uplink or downlink transmissions in different portions of the slot [18]. Massive MIMO and beamforming operations are natively supported. Finally, NR provides ad hoc support for Vehicle-to-Everything (V2X) services, including Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) transmissions both at sub-6 GHz and mmWave frequencies [19].

With respect to the architecture of the access network, NG-RAN includes several advancements with respect to the LTE RAN. Given the need for high density cellular deployment, NG-RAN has been equipped with a number of technological solutions to facilitate the deployment of small cells, build hierarchical network structures, and ease the replacement of old network equipment. Indeed, NG-RAN can operate in two different configurations, namely Standalone (SA) and Non Standalone (NSA). The SA mode is equivalent to a traditional deployment, where a 5G NR device is served by a Next Generation Node Base (gNB) (i.e., a 5G base station) and Internet access is provided by the 5G Core (5GC) (i.e., the 5G core network) [12]. In the NSA mode, instead, the gNB is assisted by an LTE base station to access the core network functions, thus removing the need for a 5GC instance, which might not be available in early deployments. Moreover, NG-RAN enables multi-connectivity, by which a primary (e.g., macro) and a secondary (e.g., small) cells can be used in conjunction to serve the same user, possible through different RATs (i.e., LTE and 5G NR) [20]. To decrease the installation costs of new cell sites, it supports the Integrated Access and Backhaul (IAB) technology, which enables the deployment of base stations with a wireless backhaul connection [21]. Another novelty is the possibility to split the gNB functionalities into multiple entities, i.e., (i) the Central Unit (CU), which implements the higher layers of the protocol stack, (ii) the Distributed Unit (DU), containing the lower layers, and (iii) the Radio Remote Unit (RRU), which handles the RF operations. This paradigm, referred to as network disaggregation, enables a flexible network customization and optimization [22].

1.2 mmWave Communications

mmWave communications represent one of the main novelties introduced by 5G. The large amount of radio resources available at these frequencies enables the usage of large transmission bandwidths and provides the possibility to reach data rates in the order of gigabits per second [7]. Despite these promises, however, communications at mmWaves introduces several challenges that need to be addressed to ensure robustness and reliability to the end users, which are summarized in the following sections.

Free Space Loss and Directionality As described by the Friis equation, the power loss of a signal propagating in free space is proportional to the frequency and the distance between the transmitter and the receiver [23]. In principle, this phenomenon limits the possibility to communicate at mmWaves, because of the high attenuation experienced at these frequencies. However, the smaller wavelength of mmWave signals make it possible to pack more antenna elements within a small form factor and realize antenna arrays with reduced dimensions [6]. Coupled with beamforming techniques, the use of antenna arrays can increase the channel gain by focusing the transmit power into narrow beams, hence increasing directivity. While directivity can (at least partially) compensate for the high pathloss experienced at mmWaves and reduce interference, it requires transmitter and receiver beams to maintain alignment, making it difficult to handle high-mobility scenarios [24].

Atmospheric and Environmental Phenomena mmWave propagation is susceptible to several environmental phenomena. Certain mmWave bands, such as the bands around 22 and 60 GHz, are subject to strong attenuations caused by water vapor and oxygen molecules present in the air [25]. Moreover, mmWave signals are susceptible to rain, since their wavelength is proportional to the size of rain droplets. However, the loss induced by this phonomenon causes severe problems only in case of extreme conditions, such as monsoons [26]. Finally, the presence of vegetation can also induce an attenuation on radio signals propagating at mmWave frequencies [27].

Shadowing and Blockage Another characteristic of mmWave propagation is that signals do not propagate through obstacles. Walls, windows or room forniture can cause shadowing or even complete blockage [28]. As a consequence, ensuring uniform coverage is challenging, and achieving seamless connectivity in indoor environments through outdoor base stations is not possible [29].

Human-induced Attenuation The presence of humans can also impact the propagation of the mmWave signals, causing ad additional loss of up to 40 dB [30, 31]. This effect mainly depends on the "shape" of the occluding bodies and the antenna configurations, with just a weak dependence on the number of people present in the environment, since signals are mainly reflected and not absorbed [32]. Therefore, the presence of long and deep fades is not uncommon and needs to be taken into account.

Doppler Effect In case of mobility, the Doppler effect causes a dispersion of the signal in the frequency domain, which is proportional to the speed and the carrier frequency. The spread experienced at mmWaves is 10 to 20 times higher than at 3 GHz, but can be considerably reduced by introducing directional antenna arrays [33].

As pointed out in [34], the peculiar propagation conditions experienced at mmWaves opens new challenges at all layers of the protocol stack and requires the system architecture to be revised.

1.3 Simulation of Cellular Systems

The design of new solutions for cellular systems typically involves three different methodologies, which have different purposes and are usually exploited jointly in order to fully understand the system behavior, namely analytical modeling, real-world measurements, and simulation. In particular, analytical modeling provides a general characterization of the system and a preliminary evaluation, usually through the derivation of bounds and/or approximations for the system performance. However, it may be difficult to devise mathematical models capable of capturing all the relevant dynamics, and several assumptions may be needed to

make these models tractable. On the other hand, measurement campaigns on prototypes or real systems could provide very accurate results, but they are very expensive and difficult to conduct. Sometimes the realization of a working prototype may not even be feasible, because the required hardware may not be available on the market. Finally, simulations consist in mimicking the system operations by means of one or multiple models describing the system and the phenomena influencing its behavior, with a certain level of abstraction. The latter is chosen according to the desired evaluation accuracy, allowing the testing of the system performance at different scales and in different situations. Also, simulations make it possible to arbitrarily set the operating conditions under which the system has to be tested, and to reproduce them at any time. Therefore, simulation allows the comparison of the performance of different variants of the system or different configuration options. This tool has been exploited throughout this thesis to assess the effectiveness of novel algorithms and architectures for mmWave cellular networks.

Nonetheless, the reliability of the evaluation results mainly depends on the quality of the simulation models, which have to be detailed enough to include the characterization of all the phenomena of interest, and this may be challenging when the system is too complex. For this reason, part of this thesis has been devoted to the development of models for the accurate simulation of next-generation cellular systems.

1.4 Contributions and Thesis Organization

The grand objective of this thesis is to provide innovative solutions to overcome the limitations of mmWave communications and fully exploit their potential in the context of 5G and beyond cellular networks. We can identify five main contributions, each corresponding to an individual chapter, which are summarized in the following.

Simulation Tools for Next-generation Wireless Networks

Network simulators are fundamental tools to assess the effectiveness of novel designs, architectures, and algorithms for networking problems, offering the possibility to monitor the performance of the overall system in a controlled environment,

with different scenarios and parameter settings, and without the need for a real deployment [35]. In Chapter 2, we present novel simulation tools for the accurate evaluation of next-generation wireless networks. In particular, we propose a new spatial channel model for the ns-3 network simulator [36], which is based on the 3GPP specifications, featuring a wide frequency range and several propagation scenarios. Moreover, we introduce a new model for antenna arrays and beamforming operations which improves the realism in the simulation of mmWave cellular systems.

Integrated Access and Backhaul in 5G mmWave Networks

The high propagation loss experienced at mmWaves limits the achievable communication range and requires operators to realize dense cellular deployments. This paradigm is often referred to as network densification [8]. However, the high capital and operational expenditures needed to deploy a large number of base stations within small areas make this effort usually impractical [37, 38]. One of the most critical parts is the backhaul, i.e., the wired interface which connects the base station to the core network. In Chapter 3, we consider wireless backhauling as a cost-effective solution towards dense mmWave deployments [39]. This technology, which has been formalized by the 3GPP, enables the deployment of base stations with a wireless backhaul connection, thus removing the need for new fiber drops. In particular, we review the 3GPP standardization activities on IAB and evaluate the performance of IAB mmWave deployments in different scenarios. Moreover, we propose a novel, semi-centralized resource management scheme based on Maximum Weighted Matching (MWM), able to improve the performance of IAB mmWave networks.

Full-stack Evaluation of Hybrid Beamforming in 5G mmWave Networks

mmWave systems make use of antenna arrays and beamforming techniques to compensate for the high pathloss experienced at these frequencies. Beamforming can be realized through either analog, hybrid, or digital hardware architectures, each providing different characteristics in terms of cost, complexity, and performance [40]. In particular, hybrid solutions represent a middle ground between

low cost/complexity and high performance and, for this reason, are widely used for the realization of mmWave systems [41]. In Chapter 4, we analyze the integration of HBF in the 5G radio protocol stack, focusing on the interplay between well-established beam design methods and the higher layers. Our results reveal novel issues in the interaction between scheduling and beamforming operations, which pose new challenges for the realization of efficient mmWave HBF cellular systems.

RAN Slicing in mmWave Cellular Networks

5G and beyond cellular systems are required to support multiple simultaneous service classes with diverse communication requirements. Network slicing is one of the main enablers to satisfy this requirement, defined as the concept of running multiple virtual networks (i.e., slices) on top of the same physical infrastructure [11]. The implementation of such concept in 5G networks gives birth to a variety of challenges, in particular, one of the biggest adversities lies in the allocation of the available resources to the different slices of the RAN [11]. In Chapter 5, we propose a RAN slicing framework for 5G cellular networks operating at mmWave frequencies. This framework is based on CA [42], a multi-connectivity technique which allows users to establish multiple connections with the base stations, and exploits the presence of different wireless links to support multiple slices simultaneously.

Towards Millimeter Wave Vehicular Networks

The exploitation of mmWave frequency bands is being considered also for new uses cases, such as vehicular communications. Notably, the multi-gigabit-per-second throughput that can be achieved at mmWaves has been seen as an opportunity for bandwidth-hungry applications in a vehicular context, where sensors (e.g., RADARs, LiDARs, cameras) and infotainment systems are expected to generate data in the order of hundreds of megabits per second [43]. In Chapter 6, we discuss the potential and challenges of mmWave communications to provide next-generation vehicular services. Notably, we introduce MilliCar, a novel simulation tool for the evaluation of mmWave V2V networks, which is based on the 3GPP NR V2X standard. Moreover, we present a performance analysis carried out using

this tool, which evaluates the impact of several system-level parameters on the end-to-end communication performance.

Finally, Chapter 7 concludes this thesis by summarizing the presented contributions and highlighting the main results.

2

Simulation Tools for Next-generation Wireless Networks

2.1 Introduction

A correct and reliable testing and performance evaluation of next-generation wireless networks becomes of paramount importance to identify the critical elements of the system before commercializing it, and to understand which algorithms and network architectures can provide the best quality of service to the end users. Simulation will play a fundamental role in this, as testbeds for 5G and next-generation Wireless Local Area Networks (WLANs) are still in the making [44, 45]. Additionally, simulations can adapt better than testbeds to the large number of evolving use cases and deployment scenarios that such networks will serve. ns-3 is well positioned to be an important simulation tool for future wireless networks, thanks to the already available modules for mmWaves and NR [46, 47], IEEE 802.11ad/ay [48, 49], and to the activity to extend the `wifi` module to also support IEEE 802.11ax [50].

Nonetheless, ns-3 is currently lacking common channel model Application Programming Interfaces (APIs) that can be used by all the aforementioned modules, to provide results based on the same channel abstraction, or to test the coexistence of different technologies in the same frequency spectrum. These modules,

indeed, currently use different channel modeling techniques, included in the modules themselves [51, 52, 53], not directly comparable with each other, and not designed with a modular and extensible approach. ns-3, on the other hand, provides a number of propagation models, and a flexible abstraction for the spectrum usage of single and multi carrier systems [54], but is lacking a fading model that can be integrated with multi-antenna wireless technologies.

The channel model, however, is one of the most important components of a wireless network simulator, as the results can only be as accurate as the channel abstraction [55]. In particular, when it comes to mmWaves, the harsh propagation conditions may severely impact the performance of the higher layers of the protocol stack, much more so than at traditional sub-6 GHz frequencies [56]. Moreover, mmWave systems generally exploit beamforming to increase the link budget of the communication, and this element has to be introduced in the overall modeling process of the channel. Additionally, when considering MIMO systems, an exact characterization of the rank of the wireless channel is necessary for a proper evaluation of how many parallel streams can be supported [57].

In the first part of this chapter, we review modern channel modeling efforts and present the implementation of a spatial channel model for future wireless networks that has been recently included in ns-3. Notably, we implemented a Spatial Channel Model (SCM) for the `spectrum` module, which characterizes the channel through a matrix $\mathbf{H}$, in which each single entry models the channel between two antenna elements at the transmitter and the receiver [58]. The channel realization is computed using the 3GPP stochastic model for 5G networks between 0.5 and 100 GHz [59]. Additionally, we extended the `propagation` module to support the models in [59], with a different characterization for Line of Sight (LOS) and Non Line of Sight (NLOS) states (according to whether the direct path between the transmitter and the receiver is blocked or not). The implementation of the channel model equations is based on that in [53], but the code has been refactored and redesigned to be as modular as possible, with a clear separation of the propagation model, the fading, the antenna, and the beamforming. It can be easily extended to support other fading models based on the computation of a channel matrix. Indeed, we introduce an extension which enables the simulation of wireless channels in vehicular environments, and an interface which can be used to read channel traces obtained with ray-tracing tools. Moreover, we present three

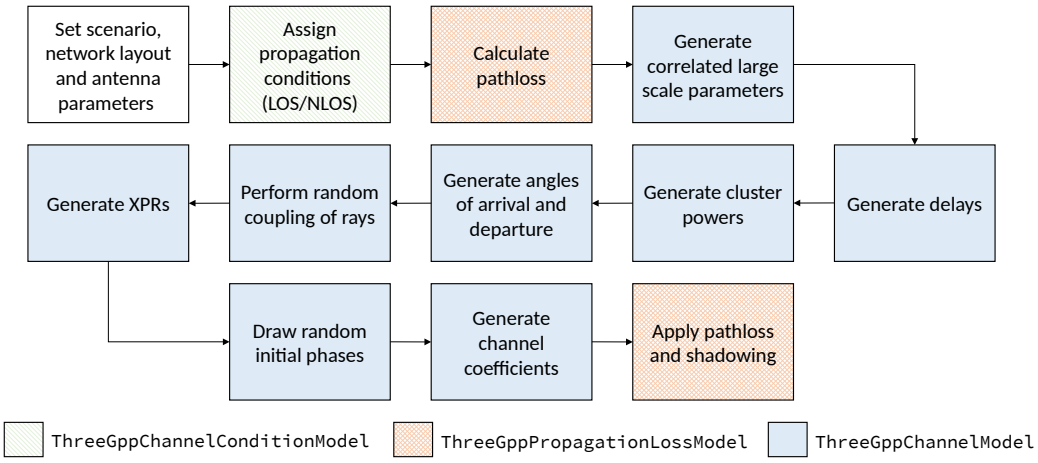


Fig. 2.1: Diagram Representing the Channel Generation Procedure; the Colors Indicate the Classes where the Steps are Accomplished

examples which demonstrate the usage of this model, and comment on possible use cases.

In the second part, we introduce novel simulation models for antenna arrays and beamforming, which adds on top of our channel model to further improve the support to the simulation of next-generation wireless systems with ns-3. This framework features (i) a flexible antenna module, comprising of multiple parametric antenna elements as well as a generic interface for phased antenna arrays, and (ii) a Beamforming (BF) module including codebook- and Singular Value Decomposition (SVD)-based BF algorithms. It is integrated with the ns-3 spatial channel model, building a complete, flexible, and accurate simulation tool. To demonstrate the capabilities of this framework, we present the full-stack evaluation of different antenna and beamforming configurations. We believe that the simulation tools described in this chapter represent a substantial and timely contribution to the wireless research community that uses ns-3 to study next-generation wireless networks.

The rest of the chapter is organized as follows. In Section 2.2 we describe the new SCM for ns-3 and its extensions, while in Section 2.3 we present the simlation framework for antenna arrays and beamforming. Finally, we conclude the chapter in Section 2.4.

2.2 Modeling of Wireless Channels

Channel modeling is a fundamental activity for the design and evaluation of future wireless networks. The authors of [55] claim that the new features of cellular and WLAN networks call for new approaches in channel modeling. Large antenna arrays and the deployment of MIMO techniques require the addition of the spatial dimension in the channel, with a full 3D model, capable of characterizing the diversity of the channel paths for each pair of antenna elements between the transmitter and the receiver. Moreover, the channel in the new frequency bands of 3GPP NR and IEEE 802.11ad/ay (i.e., mmWaves) needs proper understanding, especially with respect to multipath fading and blockage. Finally, new deployments (e.g., vehicular networks) introduce additional modeling requirements for network simulations.

These challenges have motivated several efforts in channel modeling, especially when considering mmWave frequencies [33]. Multiple measurement campaigns in these frequency bands have strived to accurately model the propagation and fading in different scenarios [7, 60, 61], highlighting how mmWaves are characterized by high propagation loss, sensitivity to blockage, and a reduced impact of small scale fading with sparsity in the angular domain. These measurement campaigns have then led to different families of channel models for future wireless networks, generally given by the combination of propagation loss and fading models. The different modeling approaches differ for their degree of abstraction, simplicity and accuracy. Analytical studies for 5G generally use simple propagation loss models, combined with Nakagami-m or Rayleigh fading [62]. These models are computationally efficient, but fail to capture the spatial dimension of the channel and cannot be combined with realistic beamforming models. Quasi-deterministic channels, developed, for example, for IEEE 802.11ad/ay [63], are instead designed to be as accurate as possible in specific scenarios, but are much more complex and require a precise characterization of the environment [64].

3GPP TR 38.901 For the evaluation of NR, the 3GPP has adopted a 3D SCM [59], which represents a tradeoff between the two aforementioned channel modeling approaches: it is generic, thanks to its stochastic nature, but at the same time can model interactions with beamforming vectors. An SCM, indeed,

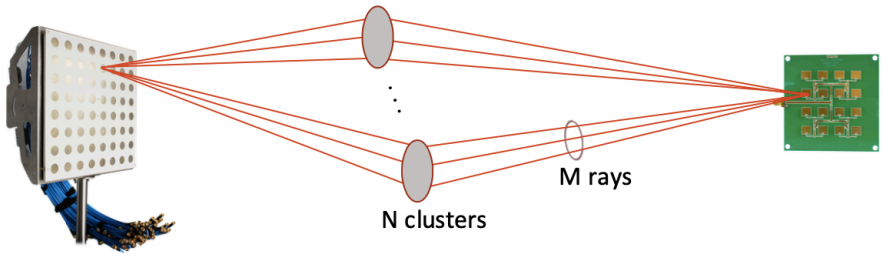


Fig. 2.2: Schematic Representation of Multipath Components.

models the channel through a channel matrix $\mathbf{H}(t, \tau)$, with as many rows and columns as the number of transmit (U) and receive (S) antenna elements. Each entry $H_{u,s}(t, \tau)$ corresponds to the impulse response of the channel between the s-th element of the Base Station (BS) antenna and the u-th element of the User Terminal (UT) antenna at delay τ at time t. $H_{u,s}(t, \tau)$ is generated by the superposition of N different clusters, representing groups of multipath components that arrive and/or depart the antenna arrays with certain angles (Figure 2.2). The multipath components impact the receiving array with different delays, and the power will be scaled according to a delay-based profile. If present, an LOS cluster is modeled with the strongest power and the minimum delay. The other clusters, instead, represent reflections from the scattering environment.

The 3GPP channel modeling framework is described in TR 38.901 [59] and represents the extension of TR 38.900, which was targeted for above-6 GHz bands only. It supports the modeling of wireless channels between 0.5 and 100 GHz by means of a stochastic SCM, in which a single instance of the channel matrix $\mathbf{H}(t, \tau)$ is computed according to random distributions for large scale fading parameters (i.e., the delay profile, the angles of arrival and departure, and the shadowing) and for the small scale fading (i.e., for small variations in the channel, for example, as given by the Doppler spread). To enable the simulation of signal propagation in different environments, it specifies four scenarios, with different parameters for the random distributions underlying the channel:

- RMa (Rural Macro), targeting rural deployments with continuous wide area coverage;
- UMa (Urban Macro), intended to model urban areas with macrocells mounted above the rooftops of the surrounding buildings;

- UMi (Urban Micro) Street Canyon, similar to UMa but with base stations mounted below the rooftops;

- Indoor Hotspot (InH) Mixed and Open Office, to model indoor environments.

For each scenario, this model provides the characterization of the LOS/NLOS channel condition, the propagation loss, and the small scale fading due to the effect of Doppler and multipath. Also, it defines a radiation model to account for the non-isotropic behavior of real antennas.

The channel matrix generation procedure, represented in Figure 2.1, accounts for both large (i.e., pathloss and shadowing) and small scale (fast fading) propagation phenomena, and provides the possibility to select different models and parameters depending on the scenario of interest. The pathloss model describes the signal attenuation between the transmitter and the receiver as a function of the 3D positions and the carrier frequency. The shadowing model provides the statistical characterization of the attenuation due to the presence of obstacles between the transmitter and the receiver. The small scale fading accounts for the signal phase and amplitude variations due to small changes in the spatial separation between the transmitter and the receiver, and for the Doppler effect introduced by a moving terminal. While the large scale propagation effects are considered to be constant within the frequency band of interest, the small scale fading has a frequency-selective behavior, thus introducing a gain which varies within the band.

In the following, we describe the 3GPP SCM for 5G networks that has been implemented in ns-3, providing details on the pathloss and channel condition computations, the channel matrix generation procedure, and the antenna model that can be associated to such matrix.

2.2.1 A Spatial Channel Model for ns-3

In ns-3, the modeling of the signal propagation through the wireless channel is handled by the `spectrum` module, which includes the abstract classes `SpectrumPhy` and `SpectrumChannel`. Devices communicating through the same wireless channel have their own `SpectrumPhy` instances, which are in charge of creating the

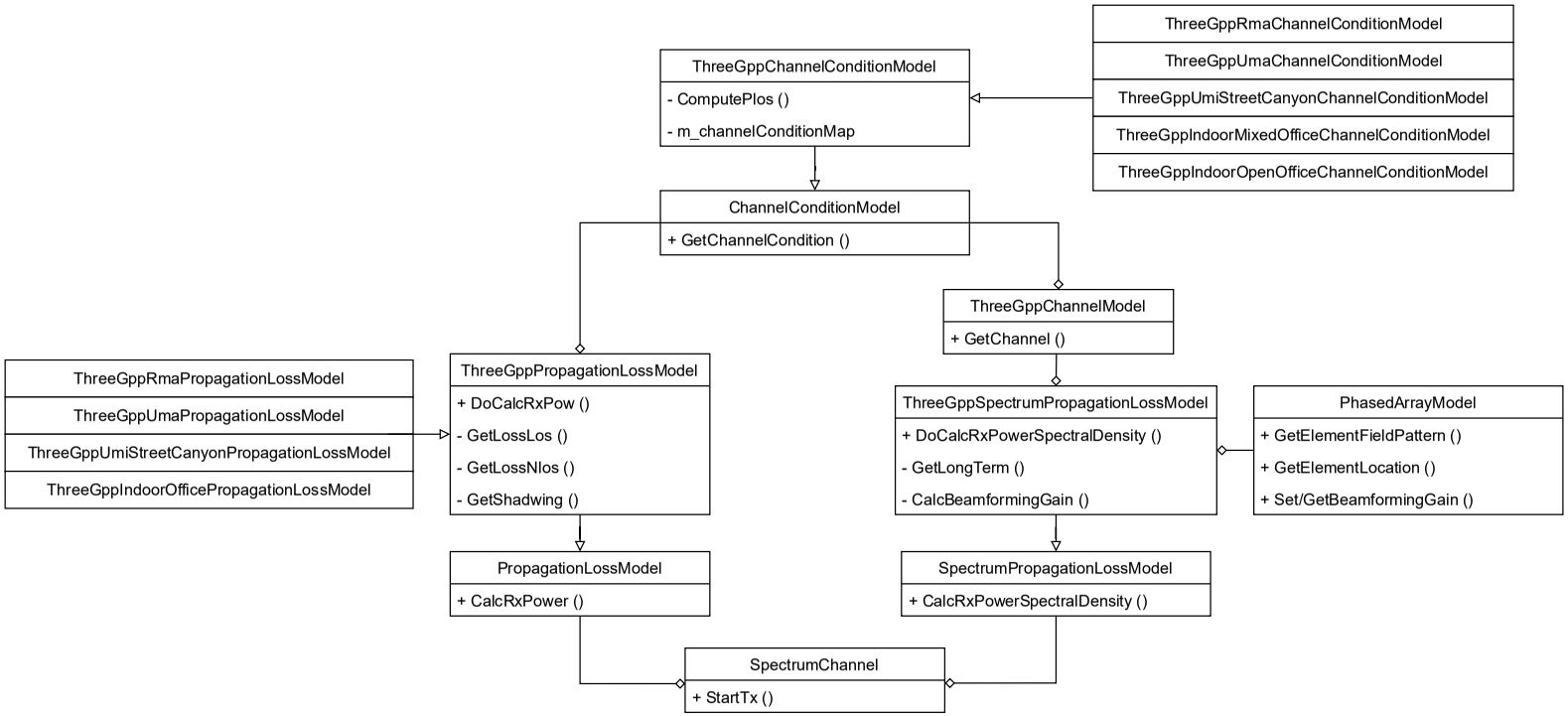


Fig. 2.3: Simplified Unified Modeling Language (UML) Diagram of the SCM Implementation

Power Spectral Density (PSD) of the transmitted signals. The different `SpectrumPhy` instances are attached to the same `SpectrumChannel` object which dispatches the transmissions among the devices. At each transmission, `SpectrumPhy` calls the method `SpectrumChannel::StartTx` which notifies each receiver and computes the corresponding PSDs of the received signals. To account for the power attenuation and fading due to the propagation of the signal through the environment, `SpectrumChannel` relies on two standard interfaces, i.e., `PropagationLossModel` and `SpectrumPropagationLossModel`. The former models slow fading, in which the loss is constant over the frequency band of the signal, while the latter is used for fast fading models, which introduce frequency-selective losses.

The 3GPP SCM can be divided into four main components, namely, (i) channel condition models, used to determine the LOS/NLOS channel state, (ii) propagation loss models, including pathloss and shadowing, (iii) the fast fading model, and (iv) the antenna model. The objective of this project was to implement these components using, whenever possible, the interfaces provided by the `spectrum` and `propagation` modules [54], without compromising the support of existing models and ensuring an easy integration in the main code base. We decided to implement each component as a separate class in order to achieve a flexible and re-usable architecture, enabling the possibility to easily replace, modify or include new parts. In this section, we focus on the first three components, while the antenna model will be described in Section 2.3. Figure 2.3 reports a simplified UML diagram for the classes involved in the channel model implementation.

LOS Probability Models

The first step for the generation of the channel matrix is to determine the LOS/NLOS channel condition. 3GPP TR 38.901 provides stochastic models to determine the channel state in all the scenarios of interest, taking into account the distance between the communication endpoints and the characteristics of the propagation environment, e.g., the presence of buildings and obstacles.

Since ns-3 lacks a general way to account for the channel state, we developed the class `ChannelCondition`, which stores the state information related to a certain channel. Also, we proposed a new interface, called `ChannelConditionModel`, which can be extended to implement any specific channel condition model, either

stochastic or deterministic. The main method is `GetChannelCondition`, which accepts as argument the positions of the two nodes and returns a pointer to the corresponding `ChannelCondition` instance.

To include the channel condition models defined in the 3GPP TR 38.901, we developed five different classes, i.e., `ThreeGppRmaChannelConditionModel`, `ThreeGppUmaChannelConditionModel`, `ThreeGppUmiStreetCanyonChannelConditionModel`, `ThreeGppIndoorOpenOfficeChannelConditionModel` and `ThreeGppIndoorMixedOfficeChannelConditionModel`, each handling a different scenario. All the new classes derive from the same base, called `ThreeGppChannelConditionModel`, which extends the `ChannelConditionModel` interface and provides caching functionalities for the periodic update of the states. When the method `GetChannelCondition` is called for the first time, the channel state is computed and its value is stored in a map, together with the generation time. Then, at subsequent calls, the method checks if the state has to be updated or not based on the time expired since its generation and, if so, a new state is independently generated, without accounting for any temporal correlation. The update interval can be tuned by the user with the attribute `UpdatePeriod`, with the possibility of never updating the channel condition if the attribute is set to 0.

Pathloss and Shadowing Models

The pathloss models defined in 3GPP TR 38.901 can be expressed through the general form of Eq. (2.1), where d is the 3D distance between the two endpoints, f_C is the carrier frequency, A, B and C are model parameters, and X is an optional loss term.

$$PL = A \log_{10}(d) + B + C \log_{10}(f_C) + X \quad \text{[dB]} \tag{2.1}$$

In particular, A represents the pathloss exponent and accounts for the dependence on the distance between the receiver and the transmitter, while C determines the relation between the pathloss and the carrier frequency. A, B, C and X take different values depending on the propagation conditions, such as the scenario, the LOS/NLOS channel state and the break point distance d_{BP}, as defined in [59].

Also, to account for the variations of the received signal power due to blockage

events, a log-normal shadowing component is added to the mean pathloss. Adjacent fading values are correlated with an exponential autocorrelation function, and their correlation depends on the spatial separation between the two positions. As for the pathloss, the standard deviation of the shadowing component, as well as the autocorrelation function, depend on the specific propagation conditions.

Moreover, 3GPP TR 38.901 specifies a model to account for the outdoor-to-indoor penetration loss due to buildings or cars, which however was not considered in this work and is planned for future development.

To include the pathloss and shadowing model defined in 3GPP TR 38.901, we developed the base class `ThreeGppPropagationLossModel`, which extends the `PropagationLossModel` interface and implements the general logic used to handle the computation of the mean pathloss and the shadowing component. Then, we extended this class by developing four subclasses, i.e., `ThreeGppRmaPropagationLossModel`, `ThreeGppUmaPropagationLossModel`, `ThreeGppUmaStreetCanyonPropagationLossModel` and `ThreeGppIndoorOfficePropagationLossModel`, which define the models for the different channel scenarios. Since the propagation loss depends on the LOS/NLOS channel state, the `ThreeGppPropagationLossModel` class is paired with a channel condition model through the `ChannelConditionModel` interface. The main method is `DoCalcRxPower`, which returns the power received at the receiver side based on the positions of the communicating nodes. It makes use of the methods `GetLossLos` and `GetLossNlos` to compute the mean pathloss in the LOS and NLOS states, respectively, and of the method `GetShadowing` to apply the shadowing model. Two other functions, namely `GetShadowingStd` and `GetShadowingCorrelationDistance`, are used by `GetShadowing` to retrieve the standard deviation of the shadowing component and the correlation distance, a parameter which defines the autocorrelation function.

Fast Fading Model

The fast fading model included in 3GPP TR 38.901 accounts for the changes in the phase and amplitude of the transmitted signal due to the effect of multipath propagation, i.e., the presence of multiple signal components that propagate over different paths. It provides the possibility to set the model parameters depending

on the scenario of interest, thus enabling the modeling of multiple propagation environments.

Eq. (2.2) represents the overall channel impulse response $H_{u,s}(t,\tau)$. As mentioned in Section 2.2, it is obtained by the superposition of $M \times N$ rays, grouped in N clusters. Rays belonging to the same cluster experience the same power P_n and propagation delay τ_n, present similar angles of arrival $(\theta^A_{n,m}, \phi^A_{n,m})$ and departure $(\theta^D_{n,m}, \phi^D_{n,m})$, and have uniformly distributed initial phases $\Phi_{n,m}$. Each ray accounts for the antenna field patterns $\mathbf{F}(\theta_{n,m}, \phi_{n,m})$ and for the power distribution among the vertical and horizontal polarizations through the term $K_{n,m}$. The terms $exp(j\bar{\mathbf{k}}^T\bar{\mathbf{d}})$ represent the array responses of the transmitting and receiving antennas, where $\bar{\mathbf{k}}$ is the wave vector and $\bar{\mathbf{d}}$ is the element location vector. In case of user mobility, each ray is subject to a phase shift $\nu_{n,m}$ due to the Doppler effect. In the LOS case, a Ricean factor is added to the direct path.

$$
\begin{aligned}
H_{u,s}(t,\tau) = & \sum_{n=1}^{N} \sqrt{\frac{P_n}{M}} \sum_{m=1}^{M} \bar{\mathbf{F}}_{rx}(\theta^A_{n,m}, \phi^A_{n,m}) \\
& \times \begin{bmatrix} e^{j\Phi^{\theta,\theta}_{n,m}} & \sqrt{K^{-1}_{n,m}} e^{j\Phi^{\theta,\phi}_{n,m}} \\ \sqrt{K^{-1}_{n,m}} e^{j\Phi^{\phi,\theta}_{n,m}} & e^{j\Phi^{\phi,\phi}_{n,m}} \end{bmatrix} \\
& \times \bar{\mathbf{F}}_{tx}(\theta^D_{n,m}, \phi^D_{n,m}) \\
& \times e^{j\bar{\mathbf{k}}^T_{rx,n,m}\bar{\mathbf{d}}_{rx,u}} e^{j\bar{\mathbf{k}}^T_{tx,n,m}\bar{\mathbf{d}}_{tx,s}} \\
& \times e^{j2\pi\nu_{n,m}t}\delta(\tau - \tau_n)
\end{aligned}
\tag{2.2}
$$

Our implementation follows the same approach described in [53], but introduces some changes to improve the modularity of the code and includes the latest updates with respect to [65]. As in [53], to reduce the model complexity, we assumed that all rays within a cluster are subject to the same Doppler shift (ν_n), corresponding to that of the central ray. Thus, the channel impulse response can be expressed as:

$$
H_{u,s}(t,\tau) = \sum_{n=1}^{N} H_{u,s,n} e^{j2\pi\nu_n t}\delta(\tau - \tau_n), \tag{2.3}
$$

where $H_{u,s,n}$ represents all the terms of the impulse response except for the

Table 2.1: Main Entries of `ThreeGppChannelMatrix`

ThreeGppChannelMatrix	
`m_channel`	the channel coefficients $H_{u,s,n}$
`m_delay`	the clusters delays τ_n
`m_angle`	the clusters arrival and departure angles
`m_generatedTime`	a time stamp indicating the generation time
`m_nodeIds`	IDs of the transmitter and receiver nodes

Doppler contribution.

We developed the class `ThreeGppChannelModel`, which computes the coefficients $H_{u,s,n}$ as described in Section 7.5 of [59] and handles their periodic update. It is associated with an instance of `ChannelConditionModel`, used to determine the LOS/NLOS channel state. The main method is `GetChannel`, which takes as input the mobility models of the transmitter and receiver nodes and the associated antenna objects, and returns an instance of `ThreeGppChannelMatrix`. As represented in Table 2.1, the structure `ThreeGppChannelMatrix` contains entries to store the channel coefficients $H_{u,s,n}$, the propagation delays τ_n, the angles of arrival and departure, and a time stamp indicating the generation time. The first time a channel is generated, the corresponding `ThreeGppChannelMatrix` is cached in a map together with identifiers for the transmitting and receiving nodes. When the same channel is requested again, the method `GetChannel` retrieves the `ThreeGppChannelMatrix` from the map and checks whether the channel coefficients have to be updated or not, depending on the expired time and the occurrence of LOS-NLOS transitions. If so, it recomputes the coefficients, otherwise it returns the old realization. Moreover, the class `ThreeGppChannelModel` provides attributes to enable an easy configuration of the model parameters, such as carrier frequency, channel scenario and update period. In particular, the choice of the update period should consider (i) the channel coherence time, i.e., the time duration over which the channel response does not vary, which depends on several factors, such as frequency, user mobility and propagation environment, and (ii) the time granularity of the simulation, which should be fine enough to capture the channel dynamics.

Blockage Model

3GPP TR 38.901 also provides an optional feature that can be used to model the blockage effect due to the presence of obstacles, such as trees, cars or humans, at the level of a single cluster. This differs from a complete blockage, which would result in an LOS to NLOS transition. Therefore, when this feature is enabled, an additional attenuation is added to certain clusters, depending on their angle of arrival. There are two possible methods for the computation of the additional attenuation, i.e., stochastic (Model A) and geometric (Model B). In this work, we used the implementation provided by Zhang et al. in [53], which uses the stochastic method. In particular, we extended the class `ThreeGppChannelModel` by including the method `CalcAttenuationOfBlockage`, which computes the additional attenuation. Also, we defined attributes to enable/disable the blockage feature and to configure the model parameters.

Computation of the PSD

The PSD of the received signal is computed as:

$$S_{rx}(t,f) = S_{tx}(t,f)\mathbf{w}_{rx}^{T}\mathcal{H}(t,f)\mathbf{w}_{tx}, \tag{2.4}$$

where $S_{tx}(t,f)$ is the PSD of the transmitted signal, $\mathbf{w}_{rx}$ and $\mathbf{w}_{tx}$ are the transmitting and receiving beamforming vectors, and $\mathcal{H}(t,f)$ is the channel matrix in the frequency domain. Applying the Fourier transform to channel coefficients expressed as in Eq. (2.3), $S_{rx}(t,f)$ can be rewritten as:

$$\begin{aligned} S_{rx}(t,f) &= S_{tx}(t,f)\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{u=1}^{U} w_{rx,u}H_{u,s,n}w_{tx,s}e^{j2\pi\nu_n t}e^{j2\pi\tau_n f} \\ &= S_{tx}(t,f)\sum_{n=1}^{N} L_n e^{j2\pi\nu_n t}e^{j2\pi\tau_n f}, \end{aligned} \tag{2.5}$$

where L_n represents the long-term component of cluster n, as defined in [53].

In our implementation, the computation of $S_{rx}(t,f)$ is handled by the class `ThreeGppSpectrumPropagationLossModel`, which extends the `SpectrumPropagationLossModel` interface. This class interacts with `ThreeGppChannelModel` to re-

trieve the channel coefficients and holds a map containing the objects representing the antennas of all the devices. The main method is `DoCalcRxPowerSpectralDensity`, which takes as input the mobility models of transmitter and receiver nodes, and returns the PSD of the received signal, computed using Eq. (2.5). In particular, it relies on the private methods `GetLongTerm`, to calculate the long term components, and `CalcBeamformingGain`, to account for the Doppler and the propagation delay. To reduce the computational load, all the long term components associated with a certain channel are cached and recomputed only when the channel realization is updated. Also, `ThreeGppSpectrumPropagationLossModel` provides the method `SetChannelModelAttribute`, which can be used to configure the model parameters, such as carrier frequency and channel scenario.

2.2.2 Extension for Vehicular Scenarios

Table 2.2: Channel Models Defined for Different V2X Links

Type of link	**Model**
V2P, P2P, V2R, R2R	Urban: TR 37.885 V2V-Urban Highway: TR 37.885 V2V-Highway
V2B, B2R	Urban: TR 38.901 UMa Highway: TR 38.901 RMa LOS
P2B	Urban: TR 38.901 UMa Highway: TR 38.901 RMa

3GPP TR 38.901 [59] describes a stochastic modeling framework which enables the simulation of 3D MIMO channels. Despite representing one of the most general tools in this field, this framework still has some limitations which may prevent its applicability in certain contexts. For instance, as a design choice, it supports mobility at a single end of the link, and therefore is not suitable for the simulation of V2V or device-to-device communications, where both end points can move. To overcome this constraint, it is possible to follow the guidelines described in Section 6.3 of TR 37.885 [66], which extends TR 38.901 by adding the possibility to model V2V links. The extended model supports mobility of both end terminals and specifies an additional Doppler component to account for the presence of scattering in high mobility environments. For a better modeling of vehicular blockages, the standard introduces a new channel state, i.e., Non Line of Sight-v (NLOSv), which represents a situation in which the direct path between the transmitter and the receiver is obstructed by a vehicle.

Moreover, it defines two new scenarios to model V2V propagation, i.e., *V2V-Urban*, which targets vehicular channels in urban environments, and *V2V-Highway*, which instead targets vehicular channels in highway environments. For each scenario, new channel condition models, propagation models, and fast fading parameters capturing the characteristics of the two environments have been defined.

TR 37.885 also indicates which channel models (including pathloss, shadowing, and fast fading modeling) are used for V2X links different than V2V. This includes all the different node pairs that can be encountered in V2X scenarios, i.e., V2P (vehicle to pedestrian), P2P (pedestrian to pedestrian), V2R (vehicle to road side unit), R2R (road side unit to road side unit), V2B (vehicle to base

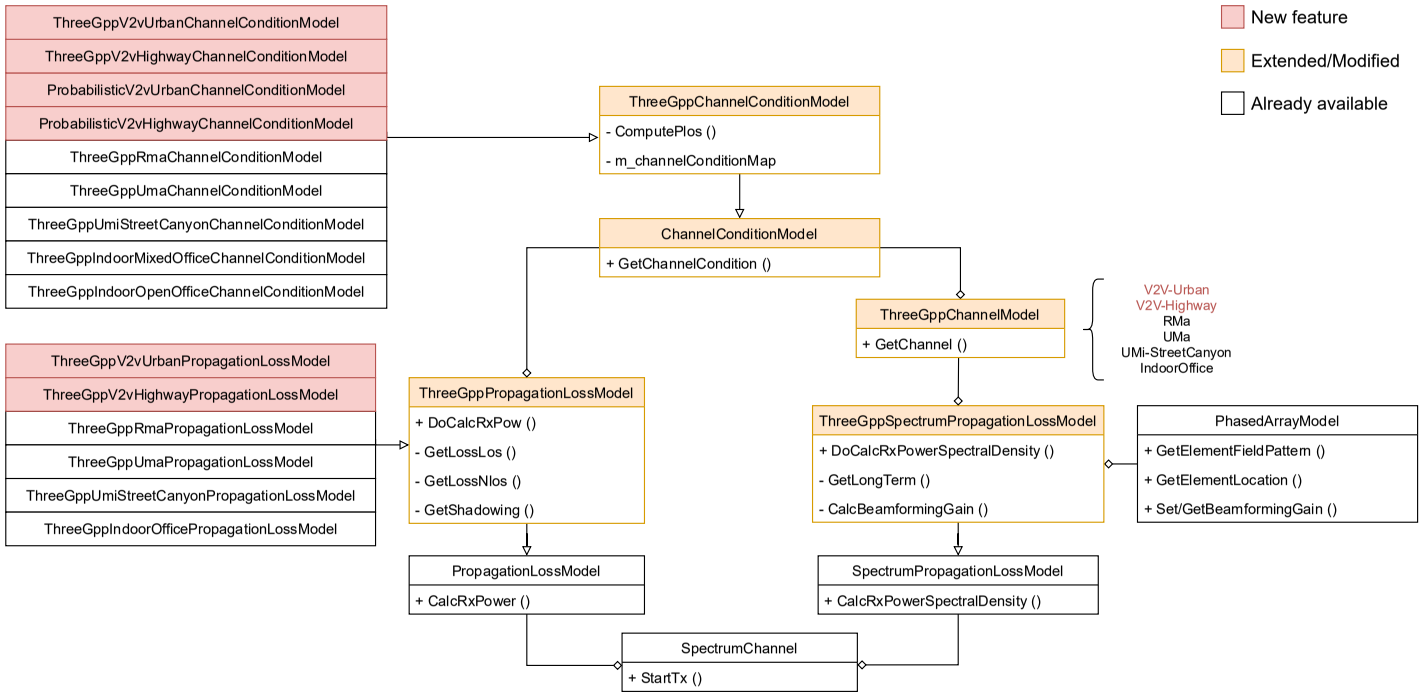


Fig. 2.4: Code Architecture

station), P2B (pedestrian to base station), and B2R (base station to road side unit). In particular, TR 37.885 defines the models for V2V, and then states the V2X links for which the V2V model applies and which models are used otherwise. A summary is provided in Table 2.2.

In Section 2.2.1, we introduced a new spatial channel model for ns-3 which implements the guidelines described in TR 38.901. In the following, we will describe how the different components of the model have been modified or extended to enable the simulation of vehicular propagation environments. Thanks to this extension, ns-3 now supports the modeling of the different V2X links reported in Table 2.2, both in urban and highway scenarios. An overview of the changes we made is provided in Figure 2.4, which shows a UML diagram representing the code architecture, where we highlighted the new classes and those that have been modified.

Channel Condition Models

The 3GPP TR 38.901 framework characterizes the wireless channel between two nodes using a two-state definition. The LOS state represents a situation in which the direct path between the transmitter and the receiver is not obstructed, while the NLOS state accounts for blockages due to buildings. The state is randomly drawn with a certain LOS probability, which depends on the propagation scenario and on the distance between the two nodes.

To capture the peculiarities of signal propagation in vehicular environments, TR 37.885 extends this definition by introducing a new state, referred to as NLOSv, whose aim is to represent a situation in which the direct path is blocked by a vehicle. Moreover, it defines a new procedure to determine the channel state. First, the model checks whether the direct path is blocked or not by looking at possible obstructions due to buildings. If the path intercepts one or more buildings, the channel is in the NLOS state. Instead, if there are no buildings along the path, the model computes the LOS probability and randomly chooses between LOS and NLOSv states. As can be seen in Table 2.3, the LOS probability depends on the distance between the communicating vehicles (d) and the formula is different depending on the scenario of interest.

To represent the state of the channel between two nodes, the ns-3 SCM uses

the class `ChannelCondition`, which has been extended to account for the new NLOSv state. Also, we extended the `ThreeGppChannelConditionModel` interface by developing two new classes, `ThreeGppV2vUrbanChannelConditionModel` and `ThreeGppV2vHighwayChannelConditionModel`, which implement the procedure to determine the channel state for the *V2V-Urban* and *V2V-Highway* scenarios, respectively.

To determine the presence of obstructions due to buildings, these classes exploit the functionalities provided by the ns-3 `buildings` module. Indeed, by creating an instance of the `Building` class, ns-3 makes it possible to model the presence of a building in the scenario. The proposed models look through the list of `Building` objects that have been instantiated, and compute the interceptions between the perimeter of each building and the path connecting the communicating nodes.

Clearly, the computation time needed to determine the channel state increases with the number of buildings, making it difficult to simulate large scale scenarios. To overcome this limitation, we developed the classes `ProbabilisticV2vUrbanChannelConditionModel` and `ProbabilisticV2vHighwayChannelConditionModel` which implement the fully probabilistic model described in [67], thus removing the need for the deterministic characterization of the NLOS state. These classes can be used as an alternative to `ThreeGppV2vUrbanChannelConditionModel` and `ThreeGppV2vHighwayChannelConditionModel`.

Pathloss and Shadowing

The pathloss and shadowing models included in TR 37.885, reported in Table 2.4, are specifically designed to account for the propagation loss of V2V wireless links as a function of the carrier frequency (f) and the distance between the two nodes (d). While the characterization of the LOS and NLOSv state is different for *V2V-Urban* and *V2V-Highway*, the NLOS pathloss equation is the same for both scenarios. Moreover, NLOSv is modeled as LOS, but with the addition of a loss component defined as $\widetilde{B}_v = \max\{0, B_v\}$, where B_v is a random variable with log-normal distribution and whose parameters depend on the height of the blocking vehicle. If the blocker is taller than both vehicles, B_v has mean 12.5 dB and standard deviation 4.5 dB. Instead, if the blocker is taller than only one of the vehicles, B_v has mean 5 dB and standard deviation 4 dB. Finally, if the

Table 2.3: Equations for LOS and NLOSv Probabilities

	V2V-Urban	***V2V-Highway***
P_{LOS}	$\min\{1, 1.05 \times \exp(-0.0114 \times d)\}$	$\begin{cases} \min\{1, 2.1013 \times 10^{-6} \times d^2 - 0.002 \times d + 1.0193\} & d \leq 475m \\ \max\{0, 0.54 - 0.001 \times (d - 475)\} & d > 475m \end{cases}$
P_{NLOS_v}	$1 - P_{LOS}$	$1 - P_{LOS}$

Table 2.4: Equations for Pathloss and Shadowing [dB]

	V2V-Urban	***V2V-Highway***
LOS	$38.77 + 16.7\log_{10}(d) + 18.2\log_{10}(f) + X$	$32.4 + 20\log_{10}(d) + 20\log_{10}(f) + X$
NLOSv	$38.77 + 16.7\log_{10}(d) + 18.2\log_{10}(f) + \widetilde{B}_v + X$	$32.4 + 20\log_{10}(d) + 20\log 10(f) + \widetilde{B}_v + X$
NLOS	$36.85 + 30\log_{10}(d) + 18.9\log_{10}(f) + X$	$36.85 + 30\log_{10}(d) + 18.9\log_{10}(f) + X$

blocker does not intercept the direct path between the communicating vehicles, no additional loss is considered. The blocker height is selected between 1.6 m, if it is a passenger vehicle, or 3 m, if it is a truck. The choice is random and depends on the percentage of trucks in the scenario, which is a model parameter.

The shadowing effect is characterized through the addition of the log-normal loss component X, with zero mean and a standard deviation of 3 dB if the channel state is LOS or NLOSv, or 4 dB if the state is NLOS. Also, the shadowing component is spatially correlated with an exponential autocorrelation function which accounts for the distance between vehicles, as well as the channel state and the propagation environment.

To implement the pathloss and shadowing models described in TR 37.885 we used the base class `ThreeGppPropagationLossModel`, provided by the ns-3 SCM, which handles the main logic for the computation of the different loss components. In particular, we developed the classes `ThreeGppV2vUrbanPropagationLossModel` and `ThreeGppV2vHighwayPropagationLossModel` which extends the base class and implements the models for *V2V-Urban* and *V2V-Highway* scenarios, respectively. The attribute `PercType3Vehicles`, common to both classes, can be used to specify the percentage of trucks in the scenario.

Fast Fading

TR 38.901 includes a fast fading model able to characterize the effect of multi-path propagation using a stochastic approach. Although this fast fading model targets cellular deployments, where only users can move while base stations are fixed, it can be easily extended to consider vehicular scenarios. Indeed, the TR 37.885 specification extends it by providing new sets of parameters for V2V channels, which have been obtained from measurement campaigns in urban and highway deployments. Moreover, it removes the single-end mobility constraint and includes an additional Doppler component for a better modeling of the environmental scattering.

In particular, the channel impulse response is still obtained through Equation 2.2, but the term $\nu_{n,m}$, which accounts for the phase shift caused by the

Doppler effect, is re-defined as:

$$\nu_{n,m} = \begin{cases} \frac{\hat{r}^T_{rx,n,m}\bar{v}_{rx}+\hat{r}^T_{tx,n,m}\bar{v}_{tx}}{\lambda_0} & \text{if } n \text{ is the LOS cluster} \\ \frac{\hat{r}^T_{rx,n,m}\bar{v}_{rx}+\hat{r}^T_{tx,n,m}\bar{v}_{tx}+2\alpha_{n,m}D_{n,m}}{\lambda_0} & \text{otherwise} \end{cases}$$

where $\hat{r}^T_{rx,n,m}$ and $\hat{r}^T_{tx,n,m}$ are the spherical unit vectors corresponding to the arrival and departure angles, $\bar{v}_{rx}$ and $\bar{v}_{tx}$ are the velocity vectors of the receiver and the transmitter, $D_{n,m}$ is a random variable with uniform distribution in $[-v_{scatt}, v_{scatt}]$ (v_{scatt} corresponds to the maximum speed of the vehicles in the layout), $\alpha_{n,m}$ is a random variable with uniform distribution in $[0, 1]$, and λ_0 is the wavelength corresponding to the carrier frequency. The addition of a random component in the Doppler term for the reflected paths has been made to consider the strong environmental scattering that can be experienced in vehicular scenarios, where the metal coating of the cars may completely reflect the signal.

To deal with the 3GPP fast fading model, the ns-3 SCM provides the classes `ThreeGppChannelModel`, implementing the procedure to compute the channel impulse response, and `ThreeGppSpectrumPropagationLossModel`, which interacts with the fast fading and antenna models to compute the channel gain. We modified these classes to include the features provided by TR 37.885, without changing the default model behavior to ensure complete backward compatibility. In particular, we included the new sets of parameters for V2V channels, which can be selected by setting the attribute `Scenario` to "*V2V-Urban*" or "*V2V-Highway*." Also, we updated the computation of the Doppler component and defined the attribute `Vscatt` in the class `ThreeGppSpectrumPropagationLossModel`, through which users can configure the parameter v_{scatt}. The default value is set to 0, so that no additional random component is considered in the Doppler term.

2.2.3 Extension for Trace-based Channel Modeling

To further improve the accuracy of channel models, it is possible to rely on quasi-deterministic approaches, mixing deterministic ray-tracing with stochastic models for diffuse scattering [68]. We implemented an add-on for ns-3, making it possible to read channel traces obtained, for example, by a quasi-deterministic mmWave channel simulator. Channel traces between each pair of nodes encode information on path loss, delay, phase, and angles of arrival and departure for each valid ray.

Both the quasi-deterministic channel simulator* and the ns-3 add-on† are available and open source, making it possible for the community to further improve their results with realistic channel modeling, while keeping the simulation complexity under control [69].

2.2.4 Examples

In this section, we present a preliminary performance evaluation which showcases the use of the proposed model and discusses possible use cases.

In our evaluation, we focused on corroborating the quality of the implemented model while, at the same time, highlighting the available features. To do so, in the following sections we present

- a basic example which shows how the different classes presented above have to be configured to build the entire 3GPP channel modeling framework;
- the results obtained with `three-gpp-v2v-channel-example`, a script that has been included among the ns-3 examples to demonstrate how the proposed model can be used to simulate vehicular propagation scenarios and extract channel metrics;
- the results of a full-stack simulation using MilliCar [70], the ns-3 module for NR V2X networks.

Basic Example

The example `three-gpp-channel-example.cc` included in the `spectrum` module demonstrates the usage of the proposed framework. It involves two devices, a transmitter and a receiver, placed at a certain distance from each other and communicating over a wireless channel. At regular intervals, we simulate a transmission between the two nodes and estimate the Signal to Noise Ratio (SNR) perceived at the receiver node. The script provides the possibility to configure the distance between the two nodes, the channel model parameters, as well as the transmission power and the receiver noise figure. Also, it produces an output trace containing the experienced propagation loss and the SNR estimate. As an

*https://github.com/wigig-tools/qd-realization
†https://github.com/signetlabdei/qd-channel

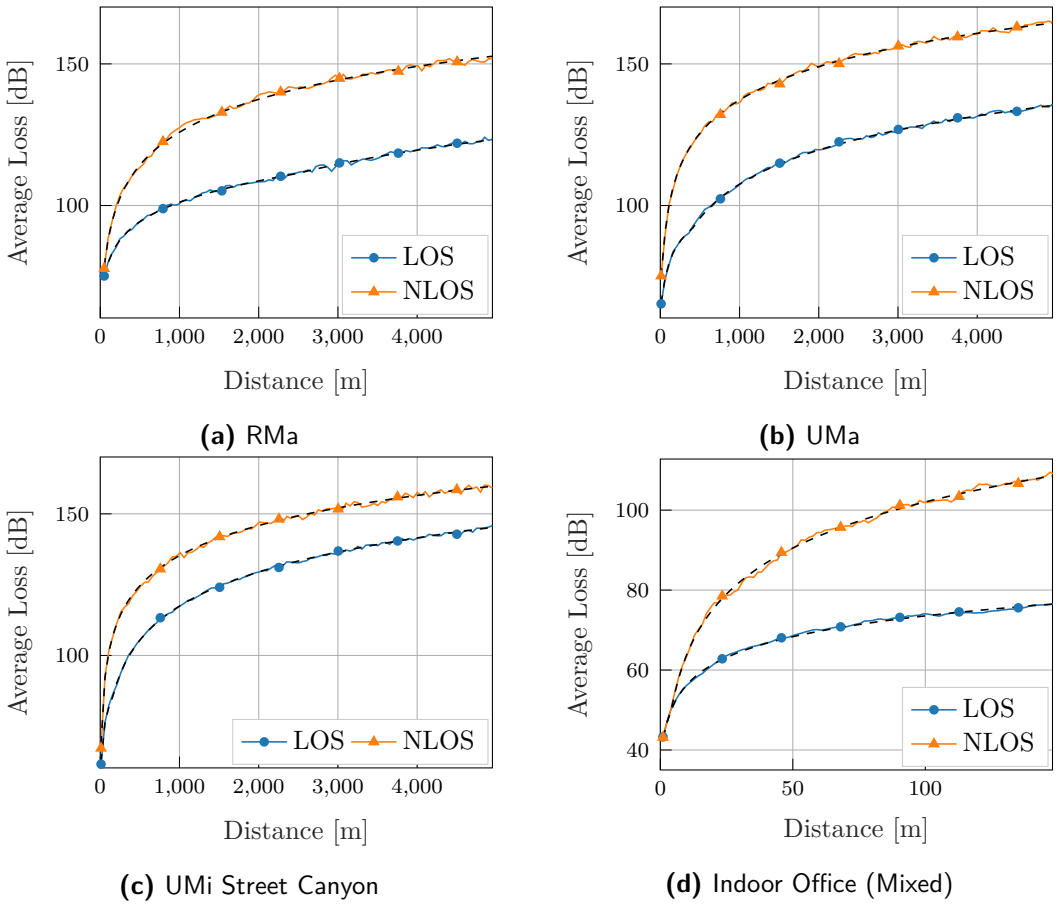


Fig. 2.5: Average Propagation Loss vs Distance between the Nodes

example, in Figure 2.5 we reported the average propagation loss (solid lines) over the distance between the two devices operating at 2.1 GHz for different scenarios and channel conditions, obtained by averaging the results of 100 independent runs of this script. The black dashed lines represent the pathloss value computed from the models defined in 3GPP TR 38.901.

In the following, we review the procedure used to create and configure the channel model classes, assuming that the UMa scenario is selected:

1. create an instance of the class `ThreeGppUmaChannelConditionModel`;
2. create an instance of `ThreeGppUmaPropagationLossModel`, configure the carrier frequency through the attribute "`Frequency`", and set the channel condition model through the attribute "`ChannelConditionModel`";

3. create an instance of `ThreeGppSpectrumPropagationLossModel`, set the UMa scenario, the carrier frequency and the channel condition model through the method `SetChannelModelAttribute` and using the attributes "`Scenario`", "`Frequency`" and "`ChannelConditionModel`";

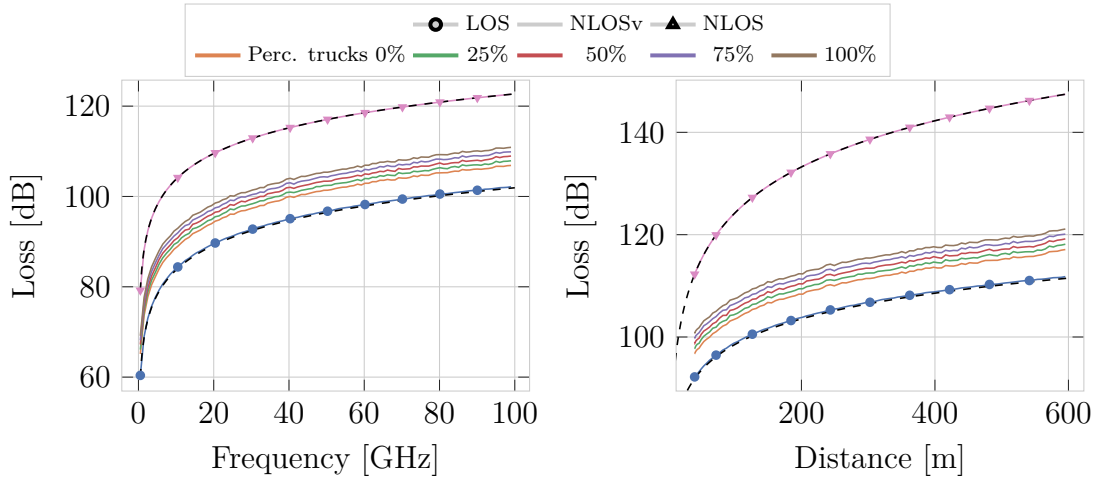
4. for each device, create an instance of `ThreeGppAntennaArrayModel` and inform the `ThreeGppSpectrumPropagationLossModel` class about the device-antenna associations by calling the method `ThreeGppSpectrumPropagationLossModel::AddDevice`.

Besides being implemented in the `three-gpp-channel-example.cc` script, these steps could be included in a helper class for an ns-3 module that aims at using this channel model. For a proper usage of this model, users may need to set the transmission time granularity of the simulation based on the channel coherence time of the scenario of interest and use an error model that accounts for the non-Additive White Gaussian Noise (AGWN) behavior of fast fading channels. For example, the error model in [71], which has been developed according to TR 38.901, could be used in combination with the proposed channel model for New Radio (NR) system-level simulations, provided that the channel coherence time is larger than the slot length of the NR frame structure.

Vehicular Example

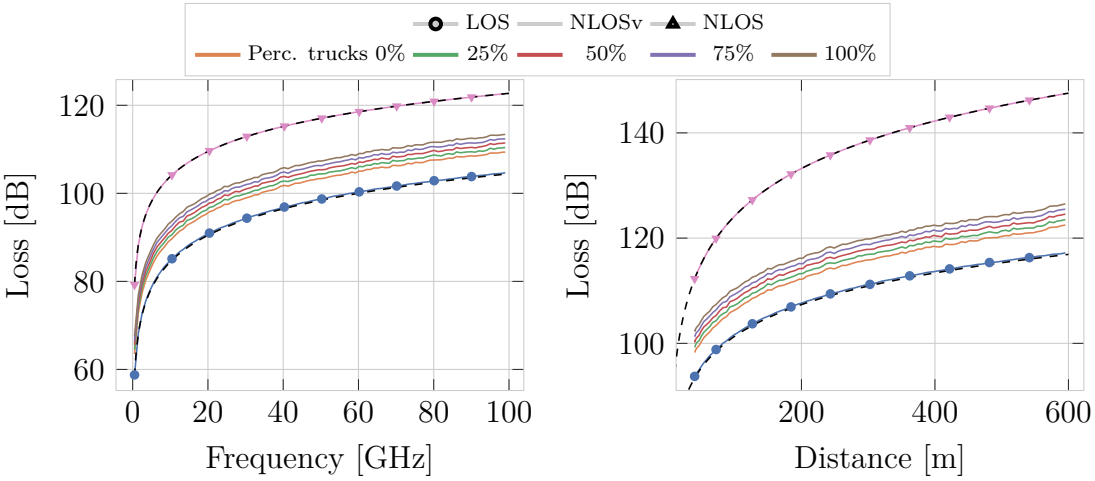
The script `three-gpp-v2v-channel-example` demonstrates how the proposed model can be used to simulate vehicular propagation scenarios and extract channel metrics. It provides the possibility to configure the distance between the communicating nodes, carrier frequency, transmission power, noise figure, and select the *V2V-Urban* or *V2V-Highway* propagation scenario. By making use of the classes described in Section 2.2.1, it computes the channel state, the propagation loss (which includes both pathloss and shadowing), and the SNR.

In our evaluation, we considered a transmitter with transmission power of 30 dBm and antenna height of 1.7 m, and a receiver with noise figure of 9 dB and height of 1.5 m. First, we placed the two nodes at a distance of 40 m and computed the average propagation loss by varying the carrier frequency between 0.5 and 100 GHz. Then, we set the carrier frequency to 28 GHz and evaluated



(a) Average Propagation Loss vs. Frequency **(b)** Average Propagation Loss vs. Distance

Fig. 2.6: Average Propagation Loss for *V2V-Urban*

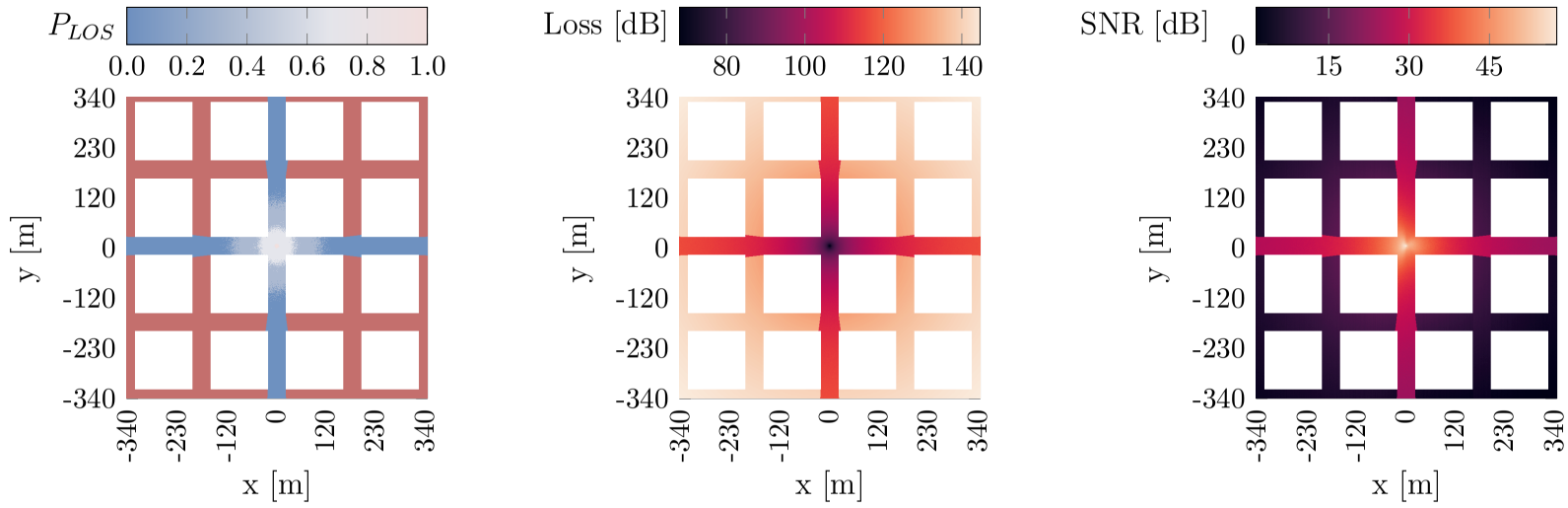


(a) Average Propagation Loss vs. Frequency **(b)** Average Propagation Loss vs. Distance

Fig. 2.7: Average Propagation Loss for *V2V-Highway*

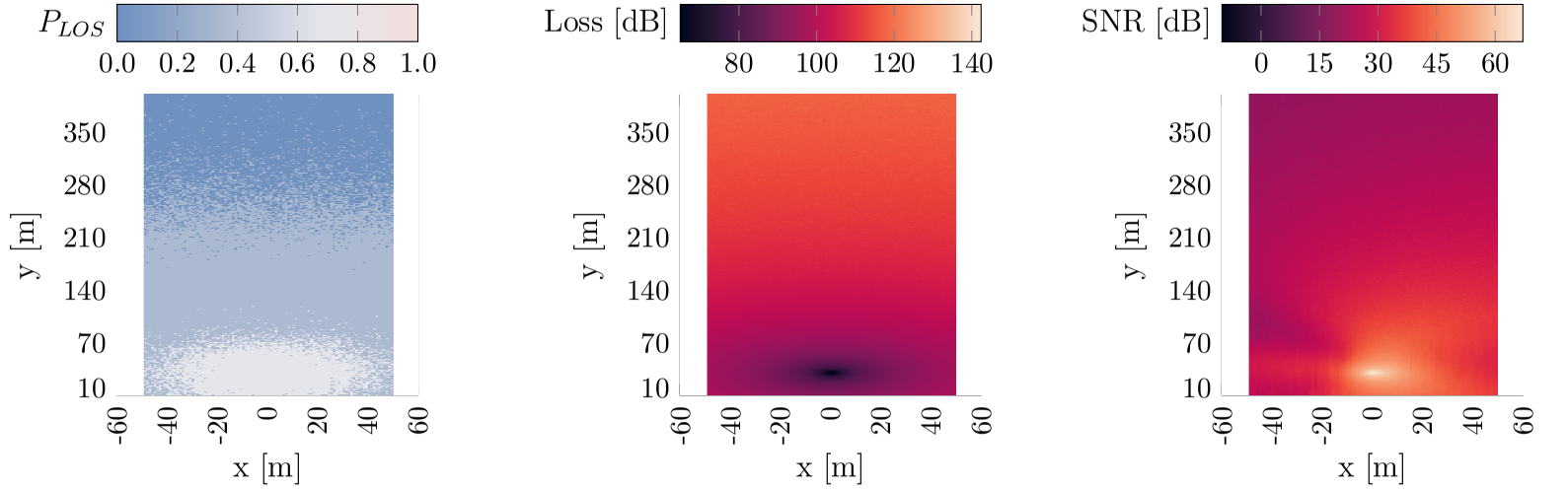
the same metric by varying the distance between the nodes, from 40 to 600 m. In both cases, we performed 1000 independent simulations.

In Figures 2.6 and 2.7, we show the results obtained for different channel states in the *V2V-Urban* and *V2V-Highway* scenarios, respectively. It can be seen that the propagation loss increases with the distance and the carrier frequency, and is highest in NLOS conditions. The propagation loss experienced in NLOSv is



(a) LOS Probability **(b)** Propagation Loss **(c)** SNR

Fig. 2.8: Low Level Metrics for *V2V-Urban*



(a) LOS Probability **(b)** Propagation Loss **(c)** SNR

Fig. 2.9: Low Level Metrics for *V2V-Highway*

between the LOS and NLOS states, as expected, and increases with the percentage of trucks in the scenario. To check the correctness of our implementation, we compared the average propagation loss in LOS and NLOS with the pathloss curves obtained using the equations in Table 2.4 and plotted as black-dashed lines.

Moreover, we used the same example to visualize the average LOS probability, propagation loss and SNR as 2D heatmaps. In this case, we performed 150 independent simulations. For the *V2V-Highway* scenario, we considered a road segment of 400 m, while for the *V2V-Urban* scenario we considered a 680× 680 m grid with 16 buildings of size 150×150 m and with a height of 10 m. In both cases, the transmitter was placed at position (0,30) for *V2V-Highway* and (0,0) for *V2V-Urban*. The results we obtained are reported in Figures 2.8 and 2.9.

It has to be highlighted that a LOS probability equal to 0 corresponds to the NLOSv condition, as NLOS is determined in a deterministic way, based on the presence of buildings, as indicated by the red areas in Figure 2.8a. We notice that as the distance between the communicating devices increases, the LOS path will almost certainly be blocked by another vehicle and, as a consequence, pathloss grows and SNR sinks.

Full Stack Example

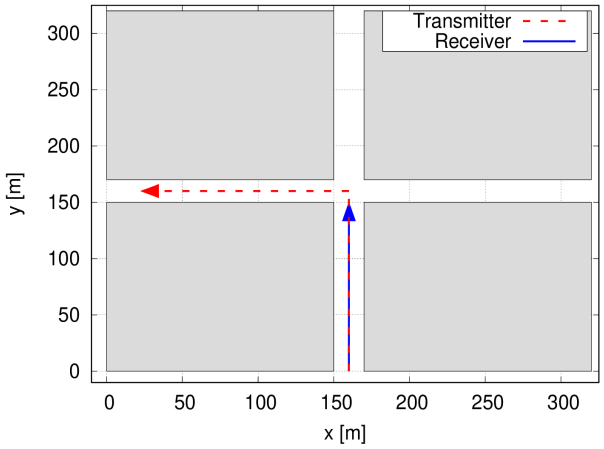


Fig. 2.10: Simulation Scenario

Millicar is an ns-3 module for the simulation of V2V communications based on the 3GPP NR V2X standard. This module will be thoroughly described in

Chapter 6, in this section we use it as a tool to demonstrate how our channel model can be used to perform full-stack simulations.

In our evaluation, we considered two devices located in an urban grid, as represented in Figure 2.10. The transmitter generates User Datagram Protocol (UDP) packets of 1024 bytes with an inter-packet interval of 60 μs, for an overall traffic of 137 Mbps. The channel was configured to operate at a frequency of 28 GHz, using a 100 MHz bandwidth, and was set to use the *V2V-Urban* equations. During the simulation, the Modulation and Coding Scheme (MCS) is adapted based on the channel estimates. The two vehicles start at a distance of 1 m, the transmitter in front of the receiver positioned at (0,0). The receiver proceeds at a 30 km/h speed, while the transmitter at 60 km/h, increasing in this way the inter-vehicle distance as the simulation goes on. At the first crossroads, after 10.3 s from the beginning of the simulation, the transmitter turns to the left and a building obstructs the LOS path.

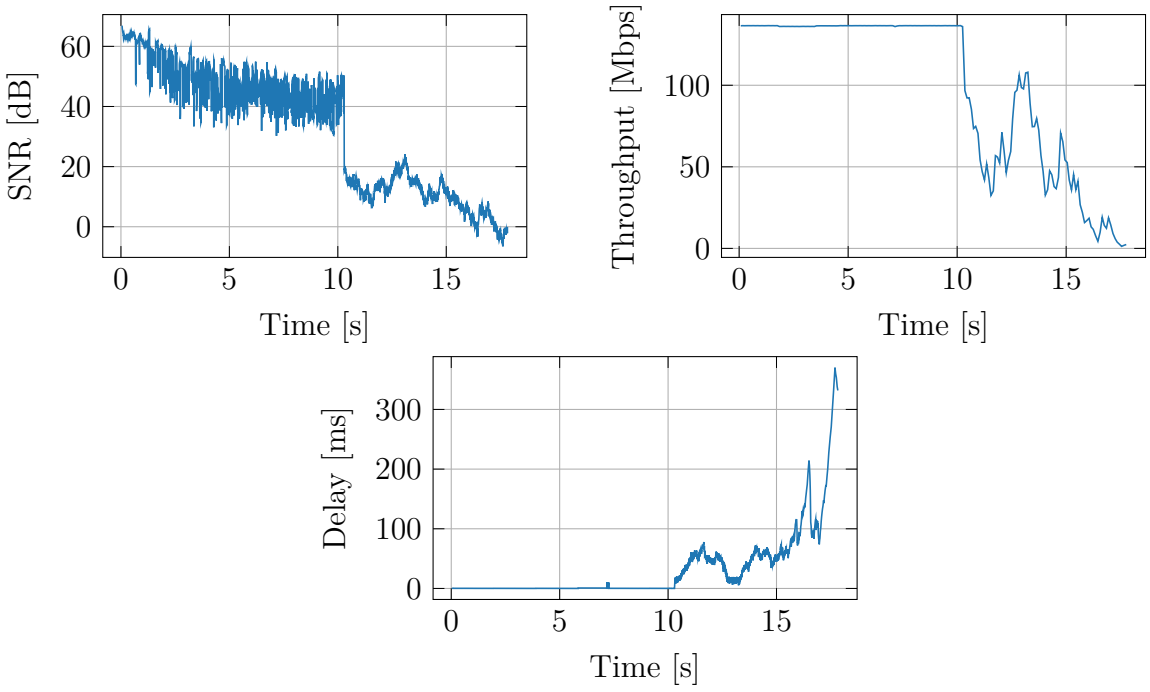


Fig. 2.11: Results for the Full Stack Example

In Figure 2.11 we represented the behavior of the SNR, as well as the end-to-end throughput and delay, experienced during the simulation. The SNR decreases as

the distance between the vehicles increases, and experiences variations that are caused by the multipath and Doppler effects. When the transmitter turns left, the SNR suddenly drops from 50 to 19 dB because the channel state condition passes from LOS to NLOS. A degradation of the SNR leads to a higher packet loss, which causes a deterioration of the end-to-end communication performance: indeed, the throughput decreases by as much as 70%, while the delay increases progressively up to 300 ms. In addition, small scale fading variations introduce further variability in the SNR, which translates in sudden changes in the used MCS as we can see from the throughput spike around 13 s.

2.2.5 Use Cases

The main target of the developed model is to enable system-level simulations of 3GPP scenarios through a 3GPP-compliant channel and antenna model. As such, it is a requirement for any **3GPP LTE and NR**-based system-level simulation that aims to properly model and evaluate the performance of physical layer techniques using appropriate channel modeling, both in the sub-6 GHz bands and in mmWave bands.

Moreover, it enables the simulation and coexistence studies of different technologies that share the spectrum resources, such as 3GPP and IEEE RATs in unlicensed/dedicated spectrum bands. For example, it can be used to evaluate the **3GPP and IEEE RATs coexistence** of:

- IEEE 802.11b/g/a/n/ac/ax (Wi-Fi) and 3GPP LTE-LAA (Licen-sed-Assisted Access) in unlicensed sub-6 GHz bands [72];
- IEEE 802.11b/g/a/n/ac/ax (Wi-Fi) and 3GPP NR-U in unlicensed sub-6 GHz bands [73, 74];
- IEEE 802.11ad/ay (WiGig, directional multi-Gigabit Wi-Fi) and 3GPP NR-U in unlicensed 60 GHz bands [75, 76];
- IEEE 802.11p/bd, 3GPP C-V2X (Cellular V2X) and 3GPP NR V2X (Vehicle-to-Everything) in dedicated sub-6 GHz bands [77];
- IEEE 802.11bd and 3GPP NR V2X in dedicated mmWave bands [19];

- Wi-Fi, IEEE 802.11p/bd, 3GPP C-V2X and 3GPP NR V2X in unlicensed sub-6 GHz bands [77];
- WiGig, IEEE 802.11bd and 3GPP NR V2X in unlicensed 60 GHz bands [19].

Also, the proposed model provides a common framework for simulations of spectrum sharing solutions through either **spectrum refarming or dynamic spectrum sharing**, for example, if different 3GPP RATs of the same operator share the licensed spectrum for some long period of time until one of the RATs becomes obsolete (spectrum refarming). This happens in low frequency bands (e.g., 900, 1800 MHz) that are essential for 3GPP NR to achieve coverage, but in which 3GPP LTE is already deployed and operational, and cannot thus be migrated to other frequency bands. As such, a key example of spectrum refarming is that of 3GPP LTE and 3GPP NR in licensed sub-6 GHz bands. Another example of spectrum sharing is when different operators share the spectrum by means of coordination policies. In this regard, the research community has also recently proposed solutions based on spectrum sharing [78] and spectrum pooling [79] for mmWave bands, which exploit coordination among different cellular network operators to improve the spatial reuse, and which could be tested from an end-to-end perspective on top of the proposed framework.

In addition, the developed model is also useful to evaluate the **3GPP and IEEE interworking** through a common channel modeling. 3GPP and IEEE interworking considers core network and radio access network integration by means of aggregating 3GPP-based RATs in licensed bands and Wi-Fi in unlicensed bands. Examples for which the developed model could be used include:

- Wi-Fi and 3GPP LTE interworking, e.g., through LTE-WLAN Aggregation (LWA) and LTE-WLAN Radio Level Integration with IPsec Tunnel (LWIP) [80],
- IEEE 802.11ax and 3GPP NR interworking [81].

2.3 Modeling of Antenna Arrays and Beamforming

Antennas are a fundamental component of every wireless system. Given that the size of an antenna is in the order of the wavelength of the desired carrier frequency [82], when operating at mmWave frequencies antennas sizes are in the order of a few millimeters [83], thus allowing to design compact antenna arrays with tens or hundreds of elements.

The most common and scalable approach is the use of phased antenna arrays, which allows us to focus the transmission in the desired direction. By carefully introducing phase shifts to the transmitted signal of each antenna element, it is possible to create complex radiation patterns, a technique called Beamforming (BF). In the simplest case, the overall radiation pattern of the antenna system is computed from the type of antenna element used, the placement of the antenna elements in the 3D space, and the BF scheme used.

Typically, antenna and BF design is carried out as an independent task, by means of real-world experiments or link-level simulations, without considering it as part of the overall system optimization. However, the solutions obtained with this approach may not be able to achieve optimal system-level performance, because they are designed without considering the interactions between the antenna systems and the higher layers of the protocol stack.

To go beyond this standalone block-level design perspective, new tools able to properly consider all the relevant aspects of the cellular system are required. For instance, the authors of [84] verified the importance of carrying out system-level simulations for the design of an 8×8 hybrid beamformer, since the cross effects between the different system blocks may have a strong impact on the overall performance. In [85], the authors present novel antenna array and BF solutions for mmWave MIMO systems based on lens antennas, and evaluate the end-to-end performance through system level simulations based on ray-tracing. Also, in [86] the authors investigate the possibility of co-designing the antennas and the Radio Frequency (RF) blocks in the front-end using a system-level platform. Although these works tackle antenna and/or BF design with a system-level approach, they make use of closed-source software or unavailable tools, specifically developed for a single application.

In this work, we propose new models for the end-to-end performance evaluation

of antenna and BF designs targeted for mmWave cellular systems. Thanks to the integration with ns-3, these models allow users to evaluate the impact of novel antenna and BF solutions on the end-to-end system behavior. Section 2.3.1 describes the antenna array model, Section 2.3.2 describes the antenna element model, while Section 2.3.3 describes the BF model.

2.3.1 Antenna Array Model

Phased antenna arrays can have extremely diverse geometries, from which their BF capabilities are derived. While it would be possible to create a generic class for arbitrary phased arrays, some geometries (e.g., uniform linear and planar arrays) are extremely popular and deserve specialized methods. For this reason, we created a generic interface for phased antenna arrays, specifying the polarized element field pattern, the locations of the elements (from which it is possible to compute the phase difference experienced by each antenna element for a transmitting or receiving signal), and the BF vector (the phase shifts and amplifications applied to every single element necessary to obtain the desired beam shape). This interface is implemented by the class `PhasedArrayModel` (see Figure 2.3) and provides the methods `GetElementFieldPattern`, which accepts as argument the azimuth and zenith angles of arrival and returns `std::pair` containing the element field components, `GetElementLocation`, which accepts as argument the index of the antenna element and returns the corresponding location vector $\bar{\mathbf{d}}$, and `Set/GetBeamformingVector` to store and retrieve the beamforming vector $\mathbf{w}$.

For this work, we considered the model described in the 3GPP specifications TR 38.901 [59]. The standard describes a uniform planar array, meaning that antenna elements are equal and are placed in an equally-spaced $M \times N$ rectangular lattice with vertical spacing d_V and horizontal spacing d_H, which form a panel. In our implementation we consider the simpler case of vertically polarized elements and only a single-panel configuration.

2.3.2 Antenna Element Model

Phased antenna arrays are composed of multiple antenna elements capable of radiating and receiving electromagnetic signals. Every antenna element has a specific radiation and polarization pattern due to its specific design. Different

antennas are needed in different contexts, e.g., directional elements can be used in multi-sector devices (e.g., gNBs), while quasi-isotropic antennas may be used for devices with no preferred communicating direction (e.g., UTs with a single-antenna panel).

A large number of antenna element designs exist in practice, leading us to creating a generic interface allowing users to add their own antenna models. In general, it is possible to create antenna elements with pattern measured from real devices to further increase the simulation accuracy. We implemented three of the most common antenna element models, with directivity pattern in dBi D_{dB} in the θ (inclination) and ϕ (azimuth) directions:

- *Isotropic antenna element*

$$D_{\text{dB}}(\theta, \phi) = 0$$

- *3GPP antenna element* [59]

$$D_{\text{v,dB}}(\theta) = -\min\left\{12\left(\frac{\theta - 90^\circ}{\theta_{\text{3dB}}}\right)^2, SLA_V\right\}$$
$$D_{\text{h,dB}}(\phi) = -\min\left\{12\left(\frac{\phi}{\phi_{\text{3dB}}}\right)^2, A_{\max}\right\}$$

$$D_{\text{dB}}(\theta, \phi) = G_{E,\max} - \min\left\{-(D_{\text{v,dB}}(\theta) + D_{\text{h,dB}}(\phi)), A_{\max}\right\}$$

  where the side-lobe attenuation in the vertical direction $SLA_V = 30$ dB, the maximum attenuation $A_{\max} = 30$ dB, the vertical and horizontal 3 dB beamwidths are respectively $\theta_{\text{3dB}} = \phi_{\text{3dB}} = 65^\circ$, and the maximum directional gain of the antenna element is $G_{E,\max} = 8$ dBi.

- *Cosine antenna element*

$$D_{\text{dB}}(\theta, \phi) = G_{\max} + 20\log_{10}\left(\cos^{\alpha_h}\left(\frac{\phi}{2}\right)\cos^{\alpha_v}\left(\frac{90^\circ - \theta}{2}\right)\right),$$

  where the exponents $\alpha_{h/v}$ can be computed from the beamwidths $\text{BW}_{\text{h/v}}$ as

$\alpha_{h/v} = \frac{-3}{20 \log_{10} \cos \frac{\mathrm{BW_{h/v}}}{4}}$, and the maximum gain $G_{\max}$ can be computed from the directivity formula found in [82].

2.3.3 Beamforming Model

Multiple BF architectures exist, which are commonly divided into three main categories, namely analog, digital, and hybrid. In analog architectures, a network of phase shifters is used to connect the antenna elements to a single RF chain, enabling a passive control of the beam by acting on the elements' phases. In digital architectures, instead, each antenna element is connected to an independent RF chain to provide digital control of the BF using baseband processing. The presence of multiple RF chains enables MU-MIMO operations, i.e., independent data streams can be transmitted and received simultaneously, possibly serving multiple users at the same time. Finally, hybrid architectures represent a middle ground between analog and digital approaches, in which the array is divided into multiple sections, each including multiple elements connected to an independent RF chain.

Although digital and hybrid architectures have the potential to achieve higher spectral efficiencies, several technological and economic issues still make analog BF a valuable choice, also due to its relatively low complexity. For this reason, in this work we consider the analog architecture and leave the study of the other two categories as future work.

Analog BF is achieved by controlling amplitude and phase shift of each antenna element of the phased array; this corresponds to assigning a complex number to each element, which is often identified as a *BF vector*. Several algorithms exist to compute such vectors, each affecting the directivity pattern in a unique way. Some try to maximize the gain in given directions, some try to suppress side lobes, some try to regulate the beamwidth, some others try to optimize the performance for a given channel estimate, and some others even try to also take into account the interference generated to other users.

In general, two main approaches exist: those based on a channel estimate, and those based on BF codebooks.

For the first approach, we implemented an algorithm originally proposed in [53] based on the MIMO *Maximum Ratio Transmission* scheme, in which the optimal

weight vectors correspond to the singular vectors associated with the largest singular values of the SVD of the estimated channel matrix. For a perfect channel estimate in interference-free environments, this method ensures optimal performance. Unfortunately, good channel estimates are hard and expensive to obtain, especially when dealing with large antenna arrays. The SVD decomposition is itself an expensive operation, and sending feedback information comprises a difficult trade-off between accuracy and overhead. For this work, we assume that the channel matrix is perfectly known, thus posing the BF algorithm in ideal conditions.

For the second approach, we implemented a generic interface for codebooks, allowing the user to create custom ones (for the sake of our evaluation, we used the tool available at [‡]). We also implemented a file-based codebook, allowing to create complex codebooks using other custom and highly-specialized softwares, avoiding the computation of sophisticated algorithms in ns-3. As a first step, the implemented codebook-based BF computes the Signal to Interference plus Noise Ratio (SINR) for every pair of TX/RX BF vector, choosing the pair with the best performance. The advantages over the previous approach are many, in particular, no channel estimation nor complex matrix decomposition is performed and the only feedback needed is the index corresponding to the best performing BF codeword. On the other hand, exhaustive search among all possible codeword pairs may be inefficient, while reducing the search to a subset of codewords might yield sub-optimal performance. We leave a more realistic and standard-compliant beam-management implementation and evaluation as future work.

2.3.4 Full-Stack Evaluation of Antenna and Beamforming Configurations

We carried out a simulation campaign to evaluate the performance of different antenna and BF configurations. To this aim, we used the ns-3 mmWave module extended with the proposed modeling framework and exploiting the ray tracing-based channel model described in Section 2.2.3. The scenario we considered, depicted in Figure 2.12, models a parking lot with multiple cars (between 1.2 and 2.25 m high) and buildings. Two mmWave BSs providing cellular coverage

[‡]https://github.com/signetlabdei/codebook-file-generator

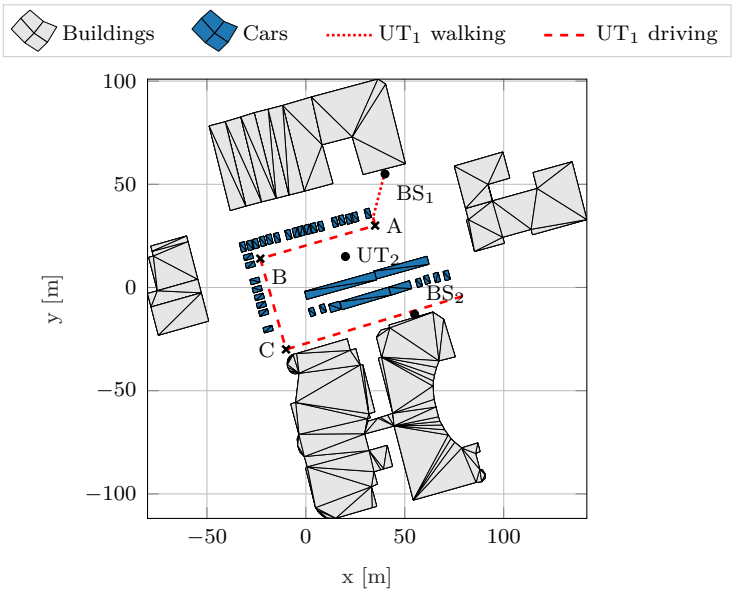


Fig. 2.12: Reference scenario.

Table 2.5: Simulation parameters.

Frequency	28 GHz
Bandwidth	400 MHz
Channel sampling period	5 ms
NR numerology index	2
Transmission power	30 dBm
Noise figure	9 dB
BS array size	8×8
UT array size	$\{1 \times 4, 4 \times 1, 4 \times 4\}$
BS element pattern	{Isotropic, 3GPP, Cosine}
UT element pattern	Isotropic
BF algorithm	{SVD, Codebook}
Codebook BF period	{10, 100, 1000} ms
APP packet size	1490 bytes
Inter-packet interval	$\{10, 1000\}$ μs
RLC mode	Acknowledge Mode (AM)

are placed on the front face of two buildings at a height of 3 m and are oriented with a bearing angle in the direction normal to the wall and with a downtilt of 12° with respect to the horizon. Two users, UT_1 and UT_2, both at a height of 1.5 m, are connected to the respective BS. During the simulation, UT_1 leaves the main building walking at 1.2 m/s up to point A and then starts driving towards the exit of the parking lot at 4.2 m/s, while UT_2 stands still at the center of the scenario. The channel was ray-traced every 5 ms considering up to 2^{nd} order specular reflections and diffuse scattering, but ignoring diffraction effects. More details on the ray-traced scenario can be found in [69]. The same type of BF schemes is used by all nodes of the scenario. We assume perfect channel knowledge for SVD BF, computed for every received and transmitted packet. Instead, to assess the impact of realistic mobility on this type of scenario, codebook-based BF is only updated to find the best codeword pair for each TX/RX node pair every {10, 100, 1000} ms.

Codebooks have been generated ensuring that adjacent beams cross at 3 dB below the maximum directivity and with no tapering across antennas. The system operates at 28 GHz with a bandwidth of 400 MHz, and is configured with NR numerology index 2. The downlink traffic is generated by a remote server which transmits UDP packets to the users at a constant rate. Table 2.5 summarizes the parameters used in our evaluation.

To evaluate the communication performance, we considered both link-level and end-to-end metrics, including SINR and SNR experienced by UT_1, respectively showing the performance with and without the interference from the second cell, and APP-layer throughput.

Simulation Results

In this section, we present and comment the results obtained. Unless explicitly stated, we consider the baseline simulation to have 4×4 arrays for the UTs, 3GPP antenna elements for the BSs, and codebook-based BF with 100 ms beam alignment, in addition to the parameters shown in Table 2.5.

In Figure 2.13, we reported the temporal evolution of the SNR and SINR experienced by UT_1. During the first part of the simulation, the SNR stays always above 50 dB and decreases as the user walks away, but the presence of

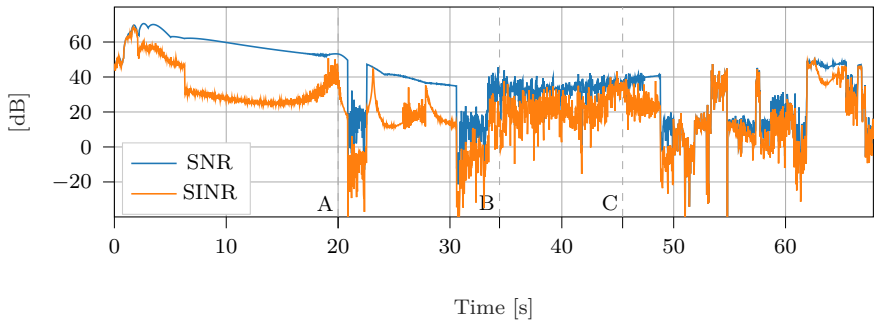


Fig. 2.13: Temporal evolution of the signal quality experienced by UT_1.

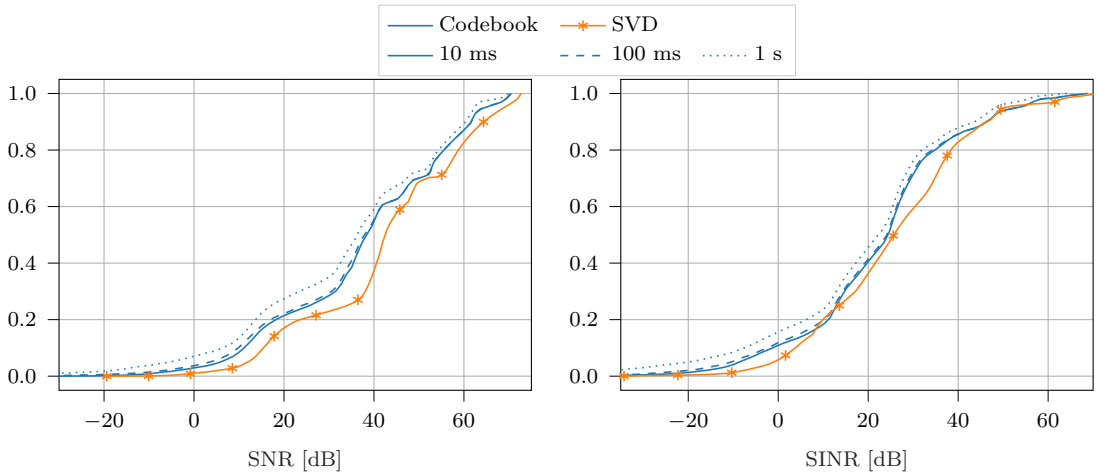


Fig. 2.14: Comparison of the SNR/SINR CDFs for different BF schemes.

interference strongly affects the channel quality, as shown by the behavior of the SINR. At time instant A, the user starts driving towards the exit of the parking lot. Shortly after 20 s and 30 s, some of the parked cars temporarily block the line of sight, making the channel quality suddenly drop. From time instant B to time instant C, both the SNR and the SINR show an oscillating behavior caused by the presence of multiple reflections with similar path losses from the surrounding cars. The last part of the simulation is characterized by multiple blockage events due to the cars parked in the bottom part of the parking lot. During this phase, the SNR and SINR exhibit similar behavior since the user is no longer subject to the inference caused by the communication between BS_2 and UT_2.

Figure 2.14 shows the Cumulative Distribution Functions (CDFs) of SNR and SINR experienced by UT_1 with different BF configurations. We can notice that

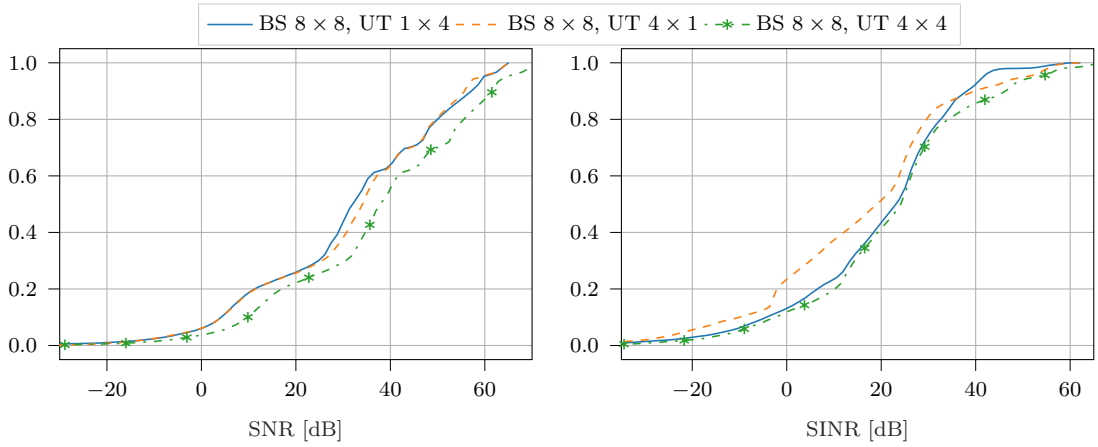


Fig. 2.15: Comparison of the SNR/SINR CDFs for different phased antenna array configurations at the UT side.

the SVD approach guarantees the best performance in terms of SNR, as supported by the theory, but not always when considering the SINR, i.e., when interference is considered. Since SVD BF does not account for interference when computing the BF vectors, while codebook BF does so when probing the different codeword pairs, the performance gap between the two approaches is reduced and SVD may even be suboptimal. Moreover, it can be seen that the value of the refresh rate used to update the weight vectors affects the behavior of the codebook-based algorithm, providing better performance for more frequent updates. Due to the geometry of the environment and the mobility, diminishing returns are clearly visible when reducing the beam alignment period from 100 ms to 10 ms making the extra overhead unnecessary.

Figure 2.15 shows a comparison between different array sizes for the UTs. Clearly, the most complex configuration represented by a 4×4 array is able to achieve the highest performance for both SNR and SINR. This is due to the higher antenna gain obtained with the larger antenna array, but also to the reduced interference due to the higher directivity. On the other hand, considering vertical 4-element Uniform Linear Arrays (ULAs) results in a very similar performance in the interference-free scenario, but vastly different performance when considering the interfering cell. In fact, a vertical array is only able to produce directivity with cylindrical symmetry around the vertical axis. Being both BSs at the same height, a good BF codeword able to improve the received power will also

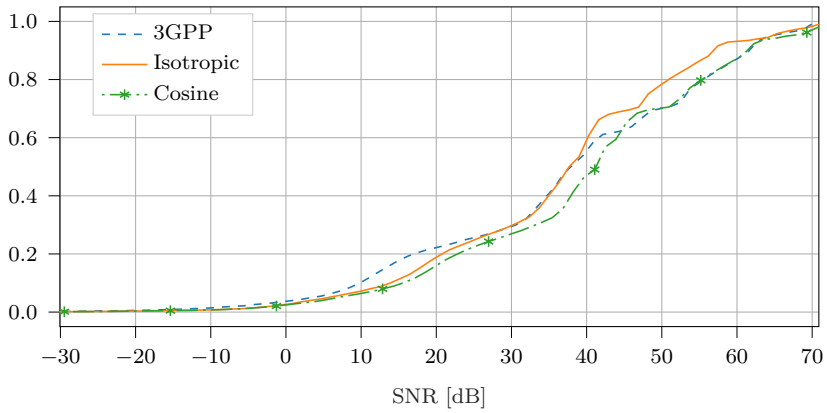


Fig. 2.16: Comparison of the SNR CDFs for different antenna element patterns.

be likely to increase the downlink interference from the second cell. On the other hand, when orienting the linear array horizontally, the cylindrical symmetry will also rotate over the horizontal axis. In this case, the geometry of the environment and the positioning of the BSs make it less likely to incur strong interference.

Figure 2.16 evaluates the impact of the element radiation pattern on the SNR experienced by the user. Isotropic elements radiate equal power in all directions, and therefore provide a low directional gain, but are able to cover a wide area. On the contrary, elements characterized by the 3GPP pattern have high directivity but small beamwidth, which implies that the transmitted power is focused in a small portion of the space. The best performance is achieved with the cosine pattern set to have a 3 dB beamwidth of 120°, thus obtaining a maximum gain $G_{\max} = 5.7$ dBi, as this represents a good compromise between directivity and beamwidth.

Figure 2.17 shows the average throughput achieved by UT_1 and UT_2 at the APP layer. With an inter-packet interval of 10 μs, the network is highly loaded and the scarcity of radio resources may prevent the recovery of the lost packets, e.g., by means of Medium Access Control (MAC) and Radio Link Control (RLC) layer retransmissions. In this situation, the choice of the BF algorithm may have a strong impact on the end-to-end performance, especially in the presence of user mobility. Indeed, as shown in Figure 2.17, the higher channel gain provided by the SVD-based algorithm allows UT_1 to achieve higher throughput, while there is no benefit for UT_2 since it stays in the same position during the entire simula-

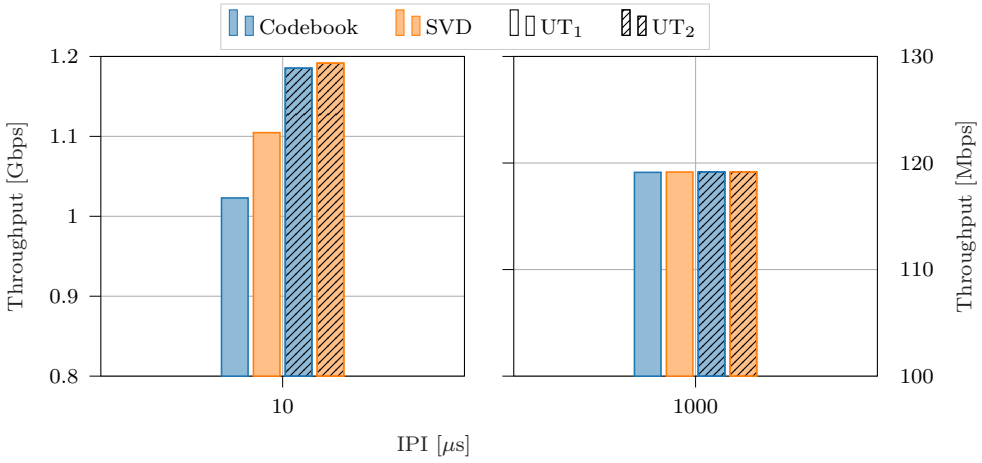


Fig. 2.17: APP-layer throughput for different inter-packet intervals.

tion. Instead, with a higher inter-packet interval, the codebook-based algorithm achieves the same performance as the SVD, since the recovery mechanisms at the MAC and RLC layers are able to compensate for the lower channel quality.

2.4 Conclusions and Future Work

In this chapter, we introduced a new channel model for ns-3 which was developed following the specifications in [59]. Moreover, we presented an extension to enable the simuation of wireless channels in vehicular environments, and an interface to read channel traces generated with ray-tracing tools. Furthermore, we introduced a new modeling framework for antenna arrays and beamforming, which interfaces with the ns-3 SCM to enable the accurate simulation of next-generation wireless networks. As a use case, we presentend the full-stack evaluation of different antenna and beamforming configurations. This work is expected to enrich ns-3 by enabling a more accurate modeling of the dynamics of wireless channels between 0.5 and 100 GHz, thus enhancing the support for the simulation of wireless systems.

After discussing the importance of channel models and simulation tools for the design of next-generation wireless systems, in Section 2.2, we explained the motivations that drive the design of more accurate channel models and described the implementation of a SCMs for ns-3. In Section 2.3, we presented the new simualtion models for antenna array and beamforming, together with a full-stack

performance evaluation.

We plan to further improve this work by (i) refining the antenna model to enable the modeling of multiple panels with dual polarization; (ii) implementing the outdoor-to-indoor penetration loss model described in [59], Section 7.4.3; (iii) implementing the additional modeling components, such as the spatial consistency procedure and the modeling of the oxygen absorption, specified in [59], Section 7.6; (iv) performing a calibration campaign to validate the model following the assumptions reported in [59], Section 7.8; and (v) lowering the computation time needed to generate a channel realization [87] through the optimization of matrix operations.

3

Integrated Access and Backhaul in 5G mmWave Networks

3.1 Introduction

Recently, wireless backhaul solutions for 5G networks have emerged as a viable strategy toward cost effective, dense mmWave deployments. Notably, the 3GPP has promoted IAB [39], i.e., a wireless backhaul architecture which dynamically splits the overall system bandwidth for backhaul and access purposes. With IAB, only few gNBs need to be connected to traditional fiber infrastructures, while the others wirelessly relay the backhaul traffic, possibly through multiple hops and at mmWave frequencies [88]. The 3GPP has recognized IAB as a cost-effective alternative to wired backhaul. Indeed, IAB has been subject of a Study Item (SI) for 3GPP NR Release 16 [39], which studies architectures, radio protocols, and the physical layer for sharing radio resources between access and backhaul links. Although 3GPP LTE and LTE-Advanced already support base stations with wireless backhaul, the SI on IAB foresees a more advanced and flexible solution, with multi-hop communications, dynamic resource multiplexing, and a plug-and-play design for low-complexity deployments. Moreover, IAB can exploit a much larger bandwidth at mmWaves than in legacy sub-6-GHz systems, and the directionality reduces the interference of concurrent access and backhaul transmissions.

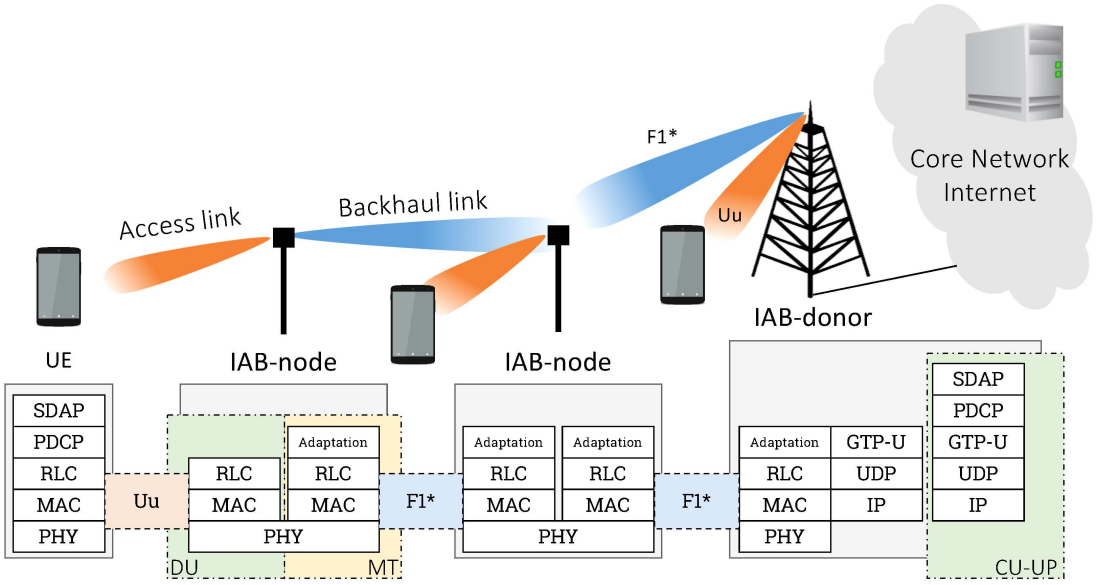


Fig. 3.1: IAB architecture. The Uu interface connects the UE and the DU in the IAB-node, while the F1* interface is used between the IAB DU and the upstream CU.

However, despite the consensus about IAB's ability to reduce costs, designing a high-performance IAB network is still an open research challenge.

This chapter reviews 3GPP standardization activities on IAB and provides actual quantitative evidence (through detailed numerical simulation results) of the performance of IAB in realistic deployments. Moreover, it introduces a novel and efficient resource partitioning framework for IAB networks.

The remainder of the chapter is organized as follows. Section 3.2 describes the 3GPP IAB activities, Section 3.3 discusses the model and the results of our end-to-end performance evaluation, while Section 3.4 identifies the potentials and challenges of this technology. Moreover, Section 3.5 introduces a novel resource management framework for IAB networks and evaluates its performance through system-level simulations. Finally, Section 3.6 concludes the chapter.

3.2 Integrated Access and Backhaul in 3GPP NR

Wireless backhaul in mobile networks has been studied extensively in the last two decades. Current cellular networks, however, do not tightly integrate access and backhaul, with the latter generally deployed through custom solutions. The

LTE specifications incorporate a relay functionality, which however has not been widely deployed, because of its reduced flexibility (it supports only a single hop, with the relay associated to a fixed parent base station, and a rigid partitioning of the access and backhaul resources).

To overcome these limitations, the 3GPP introduced the SI on IAB [39], which examines efficient solutions for integrated access and wireless backhaul over NR. This SI led to a Work Item, which will be integrated in future releases of the 3GPP specifications.

The SI considered fixed wireless relays with in-band (i.e., with the same spectrum for access and backhaul) and out-of-band backhaul (i.e., with separate bands for access and backhaul), with a focus on the former, which makes network design and management more challenging but maximizes the spectrum utilization. IAB is spectrum agnostic [39], thus it operates either in the sub-6 or in the above-6 GHz spectrum (exploiting the large bandwidth available at mmWaves), and either in SA or NSA modes. As represented in Figure 3.2 the topologies for IAB are (i) a Spanning Tree (ST), with a single parent for each IAB-node, or (ii) a Directed Acyclic Graph (DAG), where IAB-nodes may be connected to multiple upstream nodes.

In the following, we will review the innovations introduced in [39] for the network architecture, the procedures for network management, and the resource multiplexing through scheduling.

3.2.1 Architecture

The SI on IAB initially proposed five different configurations for the architecture, with various levels of decentralization of the network and backhauling functionalities. Figure 3.1 shows the logical architecture of the configuration that was eventually selected for future standardization (i.e., Architecture 1a, according to the 3GPP nomenclature), where multiple IAB-nodes use wireless backhaul, and IAB-donors have fiber connectivity towards the core network. IAB-nodes and IAB-donors can serve UEs and other IAB-nodes. Such configuration yields the most limited impact on the core network and signaling overhead, and the lowest relay complexity and processing requirements.

In this architecture, each IAB-node hosts two NR functions: (i) a Mobile Termi-

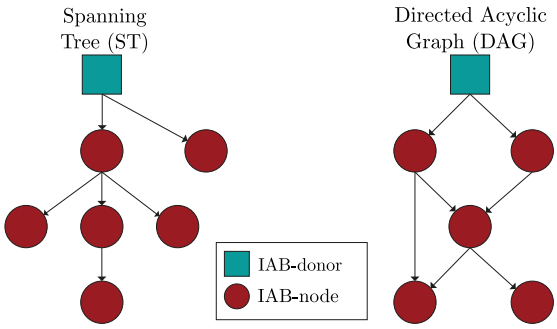


Fig. 3.2: ST and DAG topologies.

nation (MT), for the wireless backhaul connection towards an upstream IAB-node or IAB-donor, and (ii) a DU, for the access connection to the UEs or the downstream MTs of other IAB-nodes. The DU connects to a CU in the IAB-donor with the NR F1* interface running over the wireless backhaul link. Therefore, the access stack in IAB-nodes and donors serves two coexisting interfaces (the F1* and the Uu between the UEs and the DU of the gNB).

This choice implements a functional split of the radio protocol stack, with the control and upper layer in the IAB-donor CU, and the lower layers in the DUs of the IAB-nodes. The split happens at the RLC layer, therefore Radio Resource Control (RRC), Service Data Adaptation Protocol (SDAP) and Packet Data Convergence Protocol (PDCP) layers reside in the CU, while RLC, MAC and PHY are in the DUs. An additional adaptation layer manages the routing on top of RLC, hence enabling the end-to-end connection between DUs and the CU.

3.2.2 Network Procedures and Topology Management

An optimized establishment and management of the network topology is fundamental for efficient IAB operations. Indeed, the end-to-end performance of the overall network strongly depends on the number of hops between the donor and the end relay, on how many relays the donor can support, and on the procedures for network formation, route selection and resource allocation.

The topology establishment, performed during the IAB-node setup, is a critical step. Upon activation, the IAB-node first selects the upstream node. For this, the MT performs the same initial access procedure as a UE, using the NR

synchronization signals transmitted by the available cells to estimate the channel and select the parent. Moreover, although not currently supported by the specifications, we argue that it would be beneficial if the MT could retrieve additional information (e.g., the number of hops to the donor, the cell load, etc.) to select the parent cell, using more advanced path selection metrics [24] than just the channel quality, as will be discussed in Section 3.3. Then, the IAB-node configures its DU, establishes the F1* interface towards the CU in the remote IAB-donor, and can then serve UEs and other IAB-nodes. During this initial phase, the IAB-node may signal to the IAB-donor its topological location within the IAB network.

The topology is dynamically adapted for service continuity (e.g., when a backhaul link is degraded or lost), or for load balancing (e.g., to avoid congestion). Besides the signaling for the initial setup, the IAB-nodes may also transmit periodic information about traffic load and backhaul link quality. This allows the CU to be aware of the overall IAB topology, and to converge to the optimal configuration by updating the associations between the IAB-nodes.

Clearly, the ST topology exhibits less complexity but, at the same time, poses some limits in terms of network performance: the possible presence of obstacles may result in a service interruption, due to the single backhaul route available to the UEs. Greater redundancy and load balancing could be provided by a DAG topology with multi-connectivity towards multiple upstream nodes. In this case, the update of redundant routes is managed by the CU based on the propagation conditions and traffic load of each wireless backhaul link.

3.2.3 Scheduling and Resource Multiplexing

For in-band IAB operations, [39] prioritizes half-duplex operations to multiplex access and backhaul traffic within the same frequency band, although studies on full-duplex solutions are not excluded. Therefore, the radio resources must be orthogonally partitioned between access and backhaul, either in time (Time Division Multiplexing (TDM), which is the preferred solution in [39]), in frequency (Frequency Division Multiplexing (FDM)), or in space (Space Division Multiplexing (SDM)), using centralized or decentralized dynamic scheduling across the IAB-nodes and the IAB-donor.

When considering operations at mmWave frequencies, most of the literature suggests that the systems will operate in a Time Division Duplexing (TDD) fashion [89, 90]. This choice is mainly driven by the stringent latency requirements which the next generation of mobile networks will be required to support, and by the usage of analog or hybrid beamforming. The usage of Frequency Division Duplexing (FDD), in conjunction with the presence of large chunks of bandwidth, would lead to severe resource under-utilization and make channel estimation more difficult. Coupled with mmWaves directionality, this means that both self- and inter-cell interference are limited, as reported by [91].

Furthermore, at any given time instant, each node of the IAB network cannot be simultaneously involved in more than one transmission or reception. In particular, IAB-nodes cannot schedule time and frequency resources which are already allocated by their parent for backhaul communications which involve them. Moreover, the backhaul links of a given gNB might also carry data which is destined to (and/or generated by) UEs which are connected to different base stations. As a consequence, an IAB network exhibits a marked and peculiar inter-dependence between the resource allocations of the various base stations.

Finally, despite the half-duplex constraint, the IAB network is required to address the traffic requirements of all the UEs. Consequently, the resources should be allocated fairly, accounting for channel measurements and topology-related information shared by the IAB-nodes. Furthermore, both hop-by-hop and end-to-end flow control mechanisms should mitigate the congestion which might arise on intermediate hops with poor propagation conditions.

3.3 End-to-end Evaluation of IAB

We evaluated the end-to-end performance of an IAB mmWave network using the open source simulator described in [92], which implements the full stack of a cellular network and the 3GPP channel model for mmWave frequencies, and supports directional transmissions with beamforming. Unlike traditional performance analyses, e.g., [93, 94, 95], focused on PHY or MAC layers, we also consider the impact of upper layers, thereby providing a more comprehensive system-level analysis. Indeed, the integration with ns-3 makes it possible to study end-to-end scenarios with the TCP/IP stack [36] and realistic applications, such as the 3GPP

HyperText Transfer Protocol (HTTP) model. The acknowledged mode of RLC is used, thus providing additional retransmissions, besides those at the MAC layer.

In the Monte Carlo evaluation, the base stations have a height of 10 m and are randomly deployed in each simulation run following a Poisson Point Process (PPP) with density λ BS/km^2 inside a square area with side length 550 m. A fraction $0 \leq p \leq 1$ of the N base stations have wired backhaul connections (i.e., are IAB-donors), while the others (i.e., the IAB-nodes) are wirelessly connected to the IAB-donors, possibly over multiple hops. The network implements in-band backhaul, at 28 GHz. The access and backhaul resources are allocated using a dynamic TDM scheme [92], where a distributed scheduling process assigns the resources to UEs or downstream IAB-nodes in each subframe. Before this happens, parents inform downstream IAB-nodes of their scheduling decisions for the backhaul link, which are thus performed a number of subframes in advance, as proposed in [92]. We consider uniform rectangular antenna arrays in the base stations and UEs, with 64 and 16 elements, respectively, and the beamforming model described in [46]. Both the base stations and the UEs have a transmission power of 30 dBm and a receiver noise figure of 5 dB, and a bandwidth of 1 GHz. The UEs are dropped with a PPP of density $\lambda_u = 10\lambda$ UE/km^2 inside the deployment area and have a random height between 1.6 and 1.75 m, although we only evaluate the performance of the subset of UEs connected to a target base station, which is either the first gNB deployed in a baseline scenario where all nodes have a wired connection to the core network, or the first IAB-node that performs the initial access with IAB.

Backhaul path selection policies. Figure 3.3 analyzes the impact of different backhaul path selection policies in an IAB setup. As introduced in Section 3.2.2, IAB-nodes use a path selection procedure to find the route towards an IAB-donor, possibly through multiple hops. A previous work [94], investigated two different policies for backhaul traffic forwarding: (i) *Highest-quality-first (HQF)*, which selects the gNB with the highest SNR as a parent; and (ii) *Wired-first (WF)*, which chooses a direct link to the IAB-donor with the best signal above a minimum threshold (5 dB), even if an IAB-node with better channel quality is available. HQF facilitates a best-quality wireless backhaul connection in the first hop but increases the number of hops to an IAB-donor. The second approach, while minimizing the number of end-to-end hops, may choose backhaul

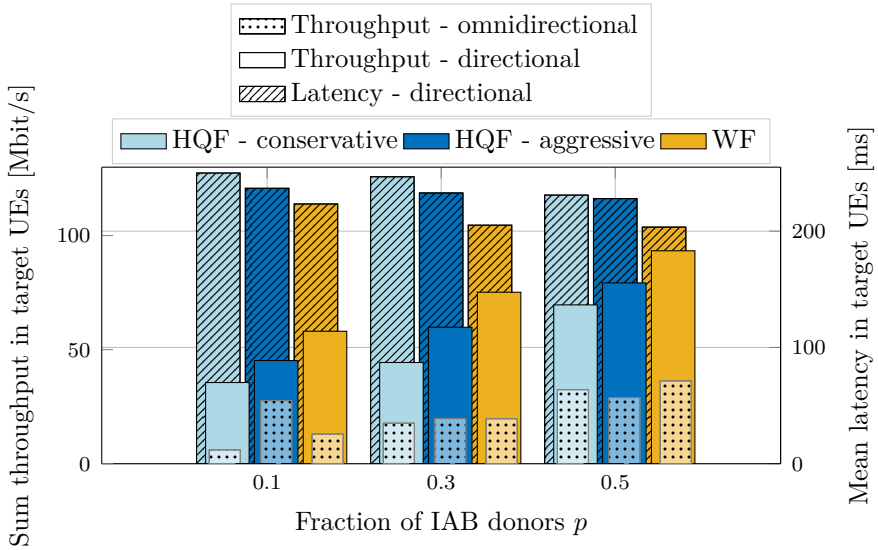
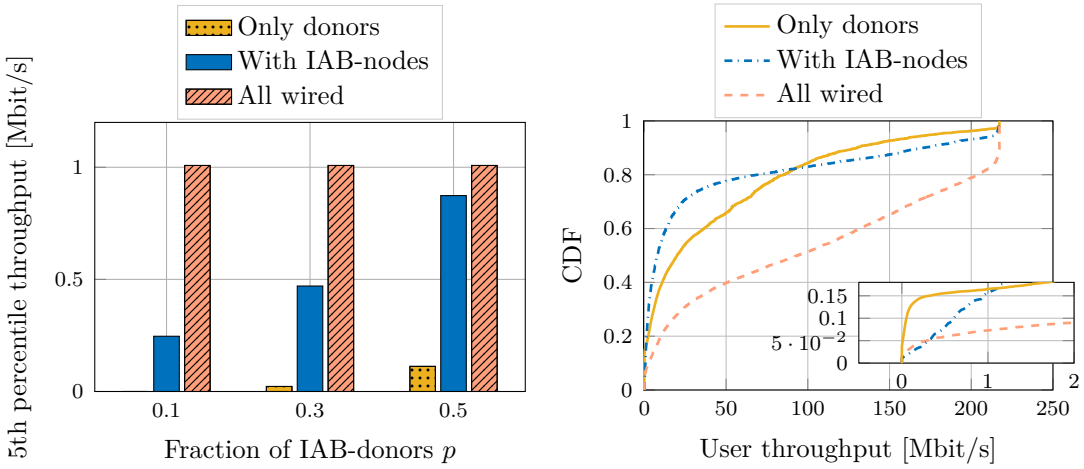


Fig. 3.3: Throughput (left y-axis) and latency (right y-axis) of IAB path selection policies varying the percentage of IAB-donors p for a density of 45 gNB/km^2 and a constant bitrate traffic.

links with poorer channel quality. The HQF policy may also leverage a function that biases the link selection towards gNBs with wired backhaul to decrease the number of hops to the core network. The bias depends on the number of hops from the IAB-node to the candidate parent it is trying to connect to [94]. Conservative and aggressive bias functions can be designed (aggressive HQF policies will progressively operate like WF policies).

The performance evaluation of [94] only considered physical layer metrics, which do not necessarily represent the quality of experience of the UE. In Figure 3.3, instead, we report end-to-end metrics, i.e., the throughput and latency at the application layer as a function of the policies and of different network parameters. These results show that the WF approach is preferable, with lower end-to-end latency and higher total throughput compared to the other policies. This is because it minimizes the number of hops to an IAB-donor: in the simulated scenarios, when $p = 0.3$, the average number of hops required to connect an IAB-node to a donor is 1.06 for WF, compared to 2.33 for conservative HQF. As a result, both the overhead and the congestion at intermediate nodes are reduced.

We also compare the throughput for the default setup (with directional transmissions) with that for omnidirectional transmissions, represented by the narrow bars in Figure 3.3. As expected, the system achieves lower throughput with re-



(a) Fifth percentile throughput.

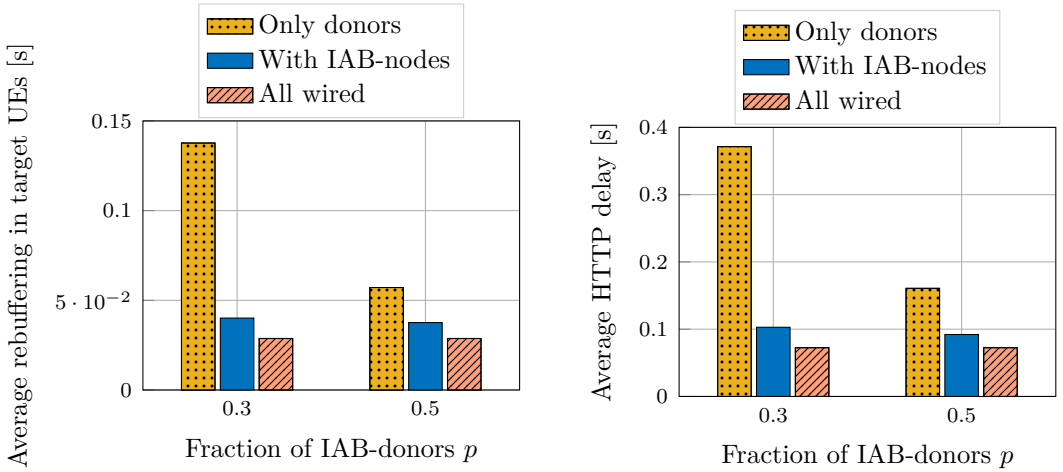
(b) Throughput CDF for $p = 0.3$.

Fig. 3.4: Fifth percentile and CDF of the throughput for the UEs of a target IAB-node, varying the percentage of IAB-donors p and the deployment strategies, for a density of 45 gNB/km^2 and UDP source traffic.

spect to the directional case, due to reduced overall link budget and increased inter-node interference, thereby pointing towards the need to perform beamforming.

IAB deployment scenarios. We tested three different deployment scenarios. The best case is when all the N base stations in the network are equipped with a wired connection to the core network (i.e., the *all wired* scenario). This represents the most expensive solution, in terms of density of fiber drops, but allocates the whole bandwidth to access traffic. With the *IAB-nodes* option, pN base stations are IAB-donors, i.e., have a wired connection, and $(1-p)N$ have wireless backhaul. Finally, the baseline is what 3GPP considers for comparisons with IAB solutions, described in [39], i.e., a deployment with only pN wired base stations and no IAB-nodes (the *only donors* configuration). In all investigated scenarios, IAB nodes use the WF policy to forward their backhaul traffic to the core network: as demonstrated in the previous paragraph, this approach delivers better end-to-end latency and throughput compared to other path selection techniques.

UDP user traffic. Figure 3.4 considers an IAB network where UEs download content from a remote server with a constant bitrate of 220 Mbps, using UDP as the transport protocol, thus with a full buffer source traffic model. Each end-to-end flow does not self-regulate to the actual network conditions, thus congestion



(a) Average rebuffering for DASH clients.

(b) Average delay to retrieve an HTTP web page.

Fig. 3.5: Performance for UEs in a target IAB-node, with different applications, for a density of 30 gNB/km^2.

arises. This experiment tests the performance of an IAB setup in a saturation regime, where the access and backhaul links are constantly used. As expected, the best performance is provided by the all wired configuration, as it provides the same access point density of the IAB setup, but avoids the multiplexing of resources between access and backhaul. On the other hand, it is possible to identify two advantages and one drawback of the IAB configuration with respect to the only donors one. IAB nodes, in fact, make it possible for the worst UEs and for the UEs with the best IAB-donor channel quality to experience a higher throughput, as shown in Figure 3.4. In the first case, the fifth percentile throughput plot in Figure 3.4a demonstrates that, for $p = 0.5$ (i.e., with one relay for each IAB-donor, on average), IAB has 7.8 times higher fifth percentile throughput than the only donors configuration, and only 13% less than the all wired setup. In the second case, Figure 3.4b shows that the usage of IAB-nodes likely offloads the worst UEs from the IAB-donors, which can allocate more resources to UEs with the best IAB-donor channel quality, thereby enabling a higher throughput. For all the other UEs, the IAB solution yields a worse performance, as they are throttled by the round robin scheduler at the donors and have a smaller throughput than with the only donors setup.

DASH, HTTP user traffic. We also consider a more common scenario, where

the UEs either stream video using Dynamic Adaptive Streaming over HTTP (DASH) [96] or access web pages using HTTP from a remote server. This source traffic is asynchronous and bursty and, in the DASH case, the flow adapts itself to the varying capacity offered by the network. Therefore, the network is not as stressed as in the previous experiment, and in this case the advantage of IAB is more visible, with the performance of IAB closer to that of the all wired deployment. Indeed, the asynchronous and independent nature of the traffic at each UE provides greater multiplexing gains, and the higher gNBs density with respect to the only donors case improves the average channel quality.

Figure 3.5a reports the average duration of a rebuffering event for a DASH stream, for all UEs of a target base station. The rebuffering happens when the DASH framework does not adapt fast enough to the network conditions, or if the capacity is not sufficient to sustain even the minimum video quality available in the DASH server. The only donors setup has the worst performance, with a 5 and 2 times higher rebuffering than the all wired configuration, for $p = 0.3$ and 0.5, respectively. The IAB deployment, instead, degrades the performance of the all wired one only by 1.4 and 1.3 times, for $p = 0.3$ and 0.5, respectively. Likewise, Figure 3.5b shows the average time it takes to completely download a web page, from the first request of the client to the reception of the last object, and the trend is similar to that of the DASH rebuffering. Finally, for this kind of traffic, the improvement introduced by the densification of IAB-donors (i.e., by increasing p from 0.3 to 0.5) is less marked than with the constant bitrate traffic shown in Figure 3.4.

3.4 Potentials and Challenges of IAB

Section 3.3 highlights that IAB presents both benefits and limitations with respect to deployments where the radio resources are used only for the access. These challenges represent promising research directions for self-configuring, easy-to-deploy and high-performance IAB networks, which provide a cost-effective solution for an initial ultra-dense NR deployment at mmWaves.

IAB architecture and application IAB presents lower deployment costs and complexity compared to the all wired setup. However, splitting the available

resources between access and backhaul traffic results in worse network performance in the presence of saturated traffic. With DASH and HTTP, instead, the traffic source is asynchronous and bursty and, consequently, the IAB performance approaches that of the all wired case. Moreover, the main advantages of IAB, when compared to the only donors setup, come from an improved channel quality of cell edge users, on average, which consequently improves the area spectral efficiency.

IAB DEPLOYMENT The interaction of different layers of the protocol stack is a design challenge for IAB. QoS should be enforced in single- and multi-hop scenarios, with IAB traffic flows for different end-to-end applications safely coexisting. Additionally, the admittance of new bearers should account for the multiplexing of resources between the access and the backahul, to avoid overbooking the available resources and introducing congestion. Figure 3.4b shows that this may indeed result in a user experience degradation. Similarly, overloading some IAB-donors or excessively increasing the number of hops should be avoided.

PATH SELECTION POLICIES The reduction of the number of relay operations (i.e., through the WF approach) limits the overhead and congestion at intermediate IAB-nodes, with improved end-to-end latency and throughput. However, the design of more efficient path selection strategies, robust to network topology changes and end terminals' mobility, is a research challenge deserving further investigation.

ANTENNA ARCHITECTURE A large number of antennas enable narrow beams, and high received power and throughput. The beamwidth, however, has an inverse relationship with how many directions to scan during the network setup phase, i.e., when the IAB-nodes perform initial access to their IAB parents. Narrower beams with analog beamforming and sequential scans can indeed increase the initial access delay. The usage of hybrid or digital beamforming can reduce this latency, and also benefit the data plane, by avoiding time or frequency multiplexing in favor of, instead, spatial multiplexing.

Scheduling and retransmissions Most system-level challenges are related to the design of ad hoc scheduling procedures at the MAC layer, to efficiently split the resources between the access and the backhaul and to provide interference management. Moreover, cross-layer effects may emerge from retransmissions at multiple layers, thus the configuration of RLC and transport layer timers needs to account for the additional delays related to the multi hop retransmissions and the packet reordering at the receiver.

Economic benefits Deploying gNBs without the need for fiber connectivity makes the IAB technology attractive and cost-effective for operators that want to improve QoS through infrastructure densification. The report [38] shows that the massive deployment of low cost small cells with wireless backhauling capabilities enables a higher capacity and a reduction of the cost per bit. In this context, wireless backhaul is a key element, because it facilitates the site installation, allows the deployment of cells even where fiber may not be available, and is cheaper to maintain. [37] reports that, for a single sector LTE small cell, the capex of wireless backhaul ($2500) is higher with respect to the wireline option ($1000), but the opex is 82% lower ($1800 vs. $10000 per year), hence the additional part of the initial investment can be quickly recovered. Although these results refer to LTE deployments, we expect that a similar trend will hold for 5G deployments, given that the cost of renting fiber will be comparable. Additionally, compared to other relaying solutions, mmWave IAB offers further advantages, including the possibility to multiplex the access and backhaul data within the same frequency band, thereby removing the need for additional hardware and/or spectrum license costs.

3.5 Resource Management Framework for IAB

In this section, we tackle the access and backhaul partitioning problem by proposing an optimal, semi-centralized resource allocation scheme for 3GPP IAB networks, based on the MWM problem on graphs. It receives periodic L1 and/or L3 measurements from the nodes of the IAB deployment, a possibility which is explicitly mentioned by 3GPP in [39, Section 7.3.3], constructs a spanning tree that represents the deployment, and uses a simplified, low-complexity version of the MWM to partition the links between access and backhaul. After a feedback step, each node can then schedule the resources at the subframe level among the connected devices.

To the best of our knowledge, this is the first MWM-based resource allocation framework for 3GPP IAB networks at mmWaves. As such, it exhibits the following benefits:

(i) no constraints on the number of hops in the IAB network are introduced, and, more in general, it is 3GPP-compliant;

(ii) a global optimum is computed;

(iii) generic network utility functions can be used;

(iv) it features a computational complexity which is linear in the number of gNBs which are connected to the same IAB-donor; and

(v) a very limited communication overhead is required.

In particular, the flexibility makes it possible to easily adapt the resource allocation strategy to different requirements, use cases, and classes of traffic for 5G networks. We achieve this by developing a generic optimization algorithm, which identifies with a configurable periodicity the access and backhaul partition that optimizes a certain utility function. The selection of the utility function prioritizes the optimization of different metrics, e.g., throughput or latency, which in turn can be mapped to different classes of traffic.

Moreover, to achieve the compliance with the 3GPP IAB specifications, the resource allocation framework relies only on information that can be actually exchanged and reported in a 3GPP deployment. Nevertheless, our solution can

be easily extended to consider other types of feedback information. Finally, the algorithm operates with a low complexity, i.e., we propose a version of the MWM algorithm that can be applied on spanning trees with linear complexity in the number of nodes in the network infrastructure, and demonstrate its equivalence to the generic (and more complex) MWM. Additionally, the proposed framework also relies on a feedback exchange that is linear in the number of base stations, and is thus decoupled from the number of users. Along this line, the semi-centralized nature of the proposed solution combines the benefit of a centralized point of view for the allocation of inter-dependent IAB links and a limited complexity.

3.5.1 State of the Art

This section reviews relevant research on resource allocation in a multi-hop wireless network, deployed through either IAB or other wireless mesh solutions [97].

The literature adopts different approaches to model and solve the resource allocation problem. The first, discussed in [91, 98, 99, 100, 101, 102, 103] is based on conventional optimization techniques. Specifically, the authors of [91] present a simple and thus tractable system model and find the minimal number of gNBs featuring a wired backhaul that are needed to sustain a given traffic load. Their work is further extended in [98], which provides an analysis of the performance benefits introduced by additional, fiber-less gNBs. In [99], the mobile network is modeled as a noise-limited, k-ring deployment. Such model is then used to obtain closed-form expressions for the max-min rates achieved by UEs in the network. Moreover, [100] proposes a system model which leads to an NP-hard optimization problem, even though it considers single-hop backhaul networks only, and uses deep Reinforcement Learning (RL) to reduce its computational complexity. In [101], the joint routing and resource allocation problem is tackled via a Linear Programming (LP) technique. Notably, this work assumes that data can be transmitted (received) toward (from) multiple nodes at the same time. Similarly, the authors of [102] formulate a TDD, multi-hop resource allocation optimization problem which leverages the directionality of mmWave antennas, albeit in the context of Wireless Personal Area Networks (WPANs). Since such problem is also NP-hard, a sub-optimal solution is found. Finally, [103] focuses on joint link scheduling, routing and power allocation in multi-hop wireless networks.

As in previous cases the obtained optimization problem is not tractable: in this instance such obstacle is overcome by studying the dual problem via an iterative approach.

The second approach relies on stochastic geometry to model IAB networks [104, 105]. Specifically, [104] determines the rate coverage probability of IAB networks and compares different access/backhaul resource partitioning strategies. Similarly, [105] provides a comparison of orthogonal and integrated resource allocation policies, although limited to single-hop wireless networks.

Another significant body of literature leverages Multi-Connectivitys (MCs) to study IAB networks; some of these works can be interpreted as a direct application of such theory [106, 107], while others [108, 109, 110, 111] exploit a more complex framework. The papers which belong to the former class are based on the pioneering work of [112], which inspects the stability of generic multi-hop wireless networks and formulates a throughput-maximizing algorithm known as *back-pressure*. In particular, [106] focuses on the optimization of the timely throughput, i.e., takes into account that packets usually have an arrival deadline. Such problem is then addressed by formulating a Markov Decision Process (MDP), leading to a distributed resource allocation algorithm. Similarly, [107] proposes an algorithm that also targets throughput optimality but, contrary to the back-pressure algorithm, manages to avoid the need for per-flow information. On the other hand, the body of literature which belongs to the latter class uses the MC-derived Network Utility Maximization (NUM) framework first introduced in [113] and [114]. Specifically, the authors of [108] focus on satisfying the URLLC QoS requirements by jointly optimizing routing and resource allocation. Then, the problem is solved using both convex optimization and RL techniques. In [109], an in-depth analysis of a mmWave, multi-hop wireless system is presented, proposing and comparing three different interference frameworks, under the assumption of a dynamic TDD system. This work is extended in [110] and [111], which consider respectively a Spatial Division Multiple Access (SDMA) and a Multi-User (MU)-MIMO capable system.

Finally, only a small portion of the literature [92, 94, 115] analyzes the end-to-end performance of IAB networks. Specifically, the authors of [92] extend the ns-3 mmWave module, introducing realistic IAB functionalities which are then used to characterize the benefit of deploying wireless relays in mmWave networks. Their

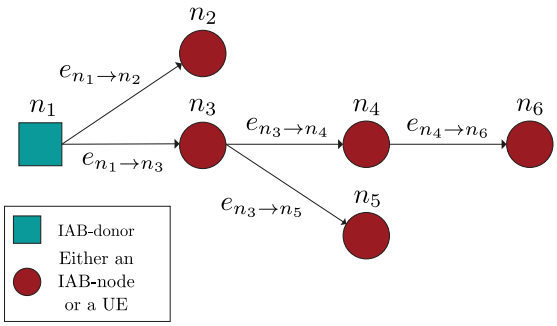


Fig. 3.6: System model notation.

work is extended in [94], where path selection policies are formulated and their impact on the system performance is inspected. A further end-to-end analysis of IAB networks is carried out in [115], providing insights into the potentials of this technology and the related open research challenges.

In conclusion, the literature exhibits the presence of algorithms relying on a varying degree of assumptions on the network topology and of knowledge about the system. Furthermore, most of the aforementioned studies lack an end-to-end, full-stack system-level analysis of the proposed solution. Conversely, we propose a semi-centralized resource allocation scheme, which also has a low complexity, both computationally and in terms of required feedback. Moreover, we provide considerations on how our proposed solution can be implemented and deployed in standard-compliant 3GPP IAB networks, and compare such solution to the state of the art with an end-to-end, realistic performance analysis.

3.5.2 System Model

A generic IAB network can be modeled as a directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where the set of nodes $\mathcal{N} \triangleq \{n_1, n_2, \dots n_{|\mathcal{N}|}\}$ comprises the IAB-donor, the various IAB-nodes and the UEs. Accordingly, the set of directed edges $\mathcal{E} \triangleq \{e_1, e_2, \dots e_{|\mathcal{E}|}\} \equiv \{e_{n_j \to n_k}\}_{j,k}$, where the edge $e_{n_j \to n_k}$ originates at the parent node n_j and terminates at the children n_k, comprises all the active cell attachments, either of mobile terminals to a gNB or from IAB-nodes towards their parent node. Since our goal is to study backhaul/access resource partitioning policies, this generic model can be actually simplified: in fact, all the UEs connected to a given gNB can be represented by a single node in $\mathcal{G}$ without any loss of generality. Similarly,

the same holds true for their links toward the serving gNB, which can then be represented by a single edge. Furthermore, this work focuses on ST topologies only.

We define as *feasible schedule* any set of links $\mathcal{E}' \subseteq \mathcal{E}$ such that none of them share a common vertex, i.e., $\forall\, e_{n_j \to n_k} \neq e_{n_l \to n_m} \in \mathcal{E}'$ it holds that $n_j \neq n_m$ and $n_l \neq n_k$. Let then f_u be a utility *additive map*, namely, a function such that the overall utility experienced by the system when scheduling edges e_1 and e_2 satisfies $f_u(e_1, e_2) = f_u(e_1) + f_u(e_2)$. Let also $\mathcal{W} \triangleq \{w_1,\, w_2,\, \dots\, w_{|\mathcal{E}|}\}$ be the set of positive weights whose generic entry w_j represents the utility which is obtained when scheduling the j-th edge, namely, $w_j \triangleq f_u(e_j)$. Then, the overall utility of the system is $\mathcal{U} \triangleq \sum_{e_k \in \mathcal{E}'} f_u(e_k) = \sum_{e_k \in \mathcal{E}'} w_k$. The goal is to find the feasible set $\mathcal{E}'^*$ which maximizes the overall utility, i.e., $\underset{\mathcal{E}'}{\operatorname{argmax}}\, \mathcal{U}$. In computer science, this task is typically referred to as the *Maximum Weighted Matching* problem [116].

Finding the MWM of a given graph, in the general case, is not trivial from a computational point of view. In fact, the fastest known MWM algorithm for generic graphs has a complexity of $\mathcal{O}(|V||E| + |V|^2 \log |V|)$ [117], posing serious limitations to the suitability of such algorithm to 5G and beyond networks, which target a connection density of 1 million devices per km^2. However, we argue that under the aforementioned assumptions on the system model, which restrict the network to an ST topology, it is possible to design an MWM-based semi-centralized resource partitioning framework which exhibits linear complexity with respect to the network size and which, as a result, is able to satisfy the scalability requirements highlighted by 3GPP in [39]. Nevertheless, the proposed framework can be easily extended to the case of a DAG IAB network. In such regard, a sub-optimal strategy is to periodically discard, during each centralized allocation, the redundant edges of each node. In such a way, the input which is fed to the `T-MWM` algorithm is, effectively, an ST. A second, optimal extension can be obtained by computing at the controller the MWM of the network via a generic MWM algorithm, instead of using the ST-specific `T-MWM` as in the proposed framework. However, this strategy would feature a higher computational complexity.

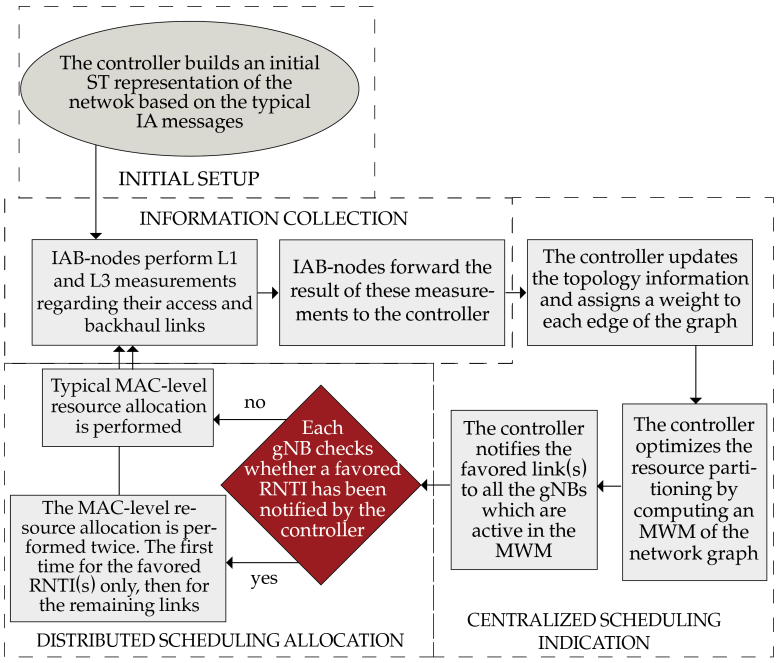


Fig. 3.7: High level diagram of the proposed MWM-based framework.

3.5.3 Semi-centralized resource allocation scheme for IAB networks

In the following, we present an MWM algorithm for ST topologies, an efficient and MWM-based semi-centralized resource partitioning framework for IAB networks and some considerations about its implementation. Specifically, the proposed scheme collects at a controller installed on the IAB-donor L1 and/or L3 measurements from the various gNBs. Then, it uses such information to build a weighted ST which represents the IAB network. In particular, the network topology is inferred by examining the incoming parent-child associations. The edge weights are also computed from the received measurements, based on the specific policy (hence, of target Key Performance Indicators (KPIs)) of choice. Finally, the resource partitioning is optimized by computing an MWM of the network and then prioritizing the links which comprise it. A high level diagram of the whole scheme is provided in Figure 3.7.

MWM for ST graphs

We present an algorithm, hereby called `T-MWM`, which computes the MWM of an ST in linear time. In particular, `T-MWM` is a bottom-up algorithm which, upon receiving as input a weighted ST $\mathcal{G}$ described by its edge map $\mathbf{E}$ and the corresponding weight map $\mathbf{W}$, produces as output a set of active edges $\mathbf{E}^*$ which are an MWM of $\mathcal{G}$. That is to say, $\mathbf{E}^*$ is a matching of $\mathcal{G}$ which yields the globally maximum utility. Furthermore, $\mathbf{E}$ is from now on assumed to exhibit the following invariant: each IAB parent precedes its children in the map, hence avoiding the need for a recursion. This is automatically obtained as each IAB child connects after its parent, and is thus added to the map in a subsequent position. Nevertheless, this assumption can be easily relaxed by reformulating the proposed dynamic programming `T-MWM` algorithm with a top-down approach.

The proposed algorithm is designed starting from the observation that, given the generic node $n_k \in \mathcal{G}$ and a matching $\bar{\mathbf{E}}$ of $\mathcal{G}$, we can identify the following mutually exclusive and collectively exhaustive cases: $\bar{\mathbf{E}}$ can contain either one or zero edges which originate from n_k. Based on this fact, we then discern the optimal utilities which can be obtained in each of these cases. Specifically, we define the maximum utilities yielded by a matching of n_k's sub-tree which either contains a link originating from n_k or not as $\mathbf{F}(n_k)$ and $\mathbf{G}(n_k)$, respectively. Then, as can be seen in Alg. 3.1, the `T-MWM` algorithm basically consists in two traversals of the network graph. During the first one we compute the $\mathbf{G}$ and $\mathbf{F}$ functions for all the nodes in $\mathcal{G}$ using the recursive formulas provided by Lemma 1. Finally, during the second traversal, this knowledge is used for computing an MWM of the network; the correctness of this last phase is proved by Lemma 2.

Lemma 1. *Given an ST $\mathcal{G}$, consider its generic internal node n_k. Let then $\mathbf{F}(n_k)$ be the maximum utility yielded by a matching of n_k's sub-tree which activates a link originating from n_k, and $\mathbf{G}(n_k)$, conversely, the utility provided when such matching contains no links which feature n_k as parent. Then, we have that:*

$$\begin{cases} \mathbf{G}(n_k) = \sum\limits_{\{n_j\}_k} \max\{\mathbf{F}(n_j), \mathbf{G}(n_j)\} \\ \mathbf{F}(n_k) = \mathbf{G}(n_k) + \max\limits_{\{n_j\}_k}\{\mathbf{W}(e_{n_k \to n_j}) \\ \qquad\qquad - [\mathbf{F}(n_j) - \mathbf{G}(n_j)]^+\} \end{cases}$$

Algorithm 3.1 Tree-Maximum Weighted Matching

Input: A weighted ST $\mathcal{G}$ encoded by a map $\mathbf{E}$, which associates each node in $\mathcal{G}$ to its edges, and the corresponding weights map $\mathbf{W}$.

Output: An MWM $\mathbf{E}^*$ of $\mathcal{G}$.

1: **procedure** `T-MWM`($\mathbf{E}, \mathbf{W}$)
2: $\mathbf{F} \leftarrow \mathbf{0}$; $\mathbf{G} \leftarrow \mathbf{0}$ ▷ Initialize the utility vectors to zero vectors
3: $\mathbf{E}^* \leftarrow \{\}$ ▷ Initialize the set of active edges as empty
4: **for** each internal node $n_k \in \mathbf{E}$ ▷ In ascending order w.r.t. to their depth in $\mathcal{G}$
5: $maxUtil \leftarrow -\infty$; $\mathbf{maxUtilChild}(n_k) \leftarrow \{\}$
6: **for** each edge $e_{n_k \to n_j} \in \mathbf{E}(n_k)$ ▷ Iterate over its edges
7: $\mathbf{G}(n_k) \leftarrow \mathbf{G}(n_k) + \max\{\mathbf{F}(n_j), \mathbf{G}(n_j)\}$
8: $currUtil \leftarrow \mathbf{W}(e_{n_k \to n_j}) - [\mathbf{F}(n_j) - \mathbf{G}(n_j)]^+$
9: **if** $currUtil > maxUtil$
10: $maxUtil \leftarrow currUtil$; $\mathbf{maxUtilChild}(n_k) \leftarrow n_j$
11: $\mathbf{F}(n_k) \leftarrow \mathbf{G}(n_k) + maxUtil$
12: **for** each internal node $n_k \in \mathbf{E}$ ▷ In ascending order w.r.t. to their depth in $\mathcal{G}$
13: **if** $\mathbf{F}(n_k) \geq \mathbf{G}(n_k)$
14: $\mathbf{E}^* \leftarrow \mathbf{E}^* \cup e_{n_k \to \mathbf{maxUtilChild}(n_k)}$
15: $\mathbf{F}(\mathbf{maxUtilChild}(n_k)) \leftarrow -\infty$ ▷ Ensure child does not get activated multiple times
16: **return** $\mathbf{E}^*$

where the set $\{n_j\}_k$ *comprises all the children of* n_k *and* $[x]^+ = \max\{x, 0\}$ *is the positive part of* x*. Conversely, for leaf nodes* n_l*,* $\mathbf{F}(n_l) \equiv \mathbf{G}(n_l) \equiv 0$*.*

Proof. This lemma can be proved by induction over the height h_k of the sub-tree corresponding to node n_k. The base case is $h_k = 0$, i.e., when n_k is a leaf node; in this case, trivially, both $\mathbf{F}(n_k)$ and $\mathbf{G}(n_k)$ are zero since no links exhibit n_k as parent node and the sub-tree of $\mathcal{G}$ which originates in n_k consists of n_k only, respectively.

Then, assume that n_k's sub-tree exhibits a generic height $h_k > 0$, and that the above formulas hold for each of its children sub-trees, which exhibit a height $h_j < h_k$. If we do not activate any edge which originates from n_k, then no added constraints are introduced concerning the edges which can be activated in its children sub-trees. Therefore, $\mathbf{G}(n_k)$ is simply the sum of the utilities achieved by any MWM computed on its children sub-trees, i.e., $\mathbf{G}(n_k) = \sum_{\{n_j\}_k} \max\{\mathbf{F}(n_j), \mathbf{G}(n_j)\}$. The remaining option is to activate exactly one edge, hereby called $e_{n_k \to n_m}$, which originates from n_k. In this case, no additional edges which feature n_m as parent can be added to the matching. As a consequence, the contribution of n_m's sub-tree on $\mathbf{F}(n_k)$ is $\mathbf{G}(n_m)$. Conversely, no additional constraints are introduced regarding the other nodes. It follows that the utility obtained in this instance is:

$$\sum_{\{n_j \neq n_m\}_k} \max\{\mathbf{F}(n_j), \mathbf{G}(n_j)\} + \mathbf{W}(e_{n_k \to n_m}) + \mathbf{G}(n_m)$$

and can be rewritten as:

$$\mathbf{G}(n_k) + \mathbf{W}(e_{n_k \to n_m}) - [\mathbf{F}(n_m) - \mathbf{G}(n_m)]^+$$

Finally, such utility is clearly maximized when n_m is chosen as $\underset{\{n_j\}_k}{\operatorname{argmax}} \{\mathbf{W}(e_{n_k \to n_j}) - [\mathbf{F}(n_j) - \mathbf{G}(n_j)]^+\}$, yielding:

$$\mathbf{F}(n_k) = \mathbf{G}(n_k) + \max_{\{n_j\}_k} \{\mathbf{W}(e_{n_k \to n_j}) - [\mathbf{F}(n_j) - \mathbf{G}(n_j)]^+\} \qquad \square$$

Lemma 2. *Given an ST* $\mathcal{G}$ *of root* n_r *and the* $\mathbf{F}$ *and* $\mathbf{G}$ *functions computed as per Lemma 1, an MWM* $\mathbf{E}^*$ *of* $\mathcal{G}$ *can be computed by performing the following procedure:*

1. *If* $\mathbf{F}(n_r) \geq \mathbf{G}(n_r)$*, add to* $\mathbf{E}^*$ *the edge from* n_r *to* n_m*, where the latter is defined as* $n_m \triangleq \underset{\{n_j\}_r}{\operatorname{argmax}} \{\mathbf{W}(e_{n_r \to n_j}) - [\mathbf{F}(n_j) - \mathbf{G}(n_j)]^+\}$*. Then, repeat recursively on all the sub-trees corresponding to* n_r*'s children* $\{n_j\}_r \,|\, n_j \neq n_m$ *and on the children of* n_m *itself.*

2. *If* $\mathbf{F}(n_r) < \mathbf{G}(n_r)$*, repeat recursively on all the sub-trees corresponding to* n_r*'s children.*

Proof. The above procedure always yields a feasible activation, i.e., a matching of $\mathcal{G}$. In particular, in either options we never recurse on a node which has already been activated, hence no pair of edges $\in \mathbf{E}^*$ can share any vertices. Furthermore, due to the properties of $\mathbf{F}$ and $\mathbf{G}$, whenever $\mathbf{F}(n_r) \geq \mathbf{G}(n_r)$ a matching yielding maximal utility can be obtained by activating the edge $e_{n_r \to n_m}$, where $n_m \triangleq \underset{\{n_j\}_r}{\operatorname{argmax}} \{\mathbf{W}(e_{n_r \to n_j}) - [\mathbf{F}(n_j) - \mathbf{G}(n_j)]^+\}$. Since the procedure is then recursively repeated on n_r's children and the validity of $\mathbf{F}$ and $\mathbf{G}$'s properties holds for each sub-tree in $\mathcal{G}$, the set of edges $\mathbf{E}^*$ produced by the above procedure comprises a *maximal* matching, i.e., it yields the maximum possible utility among all the feasible schedules. □

Regarding the computational complexity of the proposed algorithm, it can be observed that during the first phase the main loop effectively scans each edge of $\mathcal{G}$, hence exhibiting a complexity $\mathcal{O}(|\mathbf{E}|)$. Moreover, the second phase of `T-MWM` has complexity $\mathcal{O}(|\mathbf{V}|)$, since it loops through all the network nodes. Therefore, we can conclude that the overall asymptotic complexity of the algorithm is $\mathcal{O}(|\mathbf{V}| + |\mathbf{E}|)$, or, equivalently, $\mathcal{O}(|\mathbf{V}|)$ since in an ST the number of edges equals $|\mathbf{V}| - 1$.

Semi-centralized resource partitioning scheme

Based on the system model introduced in Section 3.5.2, and the `T-MWM` algorithm, we present a generic optimization framework which partially centralizes the backhaul/access resource partitioning process, in compliance with the guidelines of [39]. The goal of this framework is to aid the distributed schedulers, adapting the number of OFDM symbols allocated to the backhaul and access interfaces to the phenomena which exhibit a sufficiently slow evolution over time, i.e., large scale fading and local congestion. This optimization is undertaken with

respect to a generic additive utility function f_u. An IAB network of arbitrary size is considered, composed of a single IAB-donor, multiple IAB-nodes and a (possibly time-varying) number of UEs which connect to both types of gNBs. Furthermore, assume that a central controller is installed on the IAB-donor.

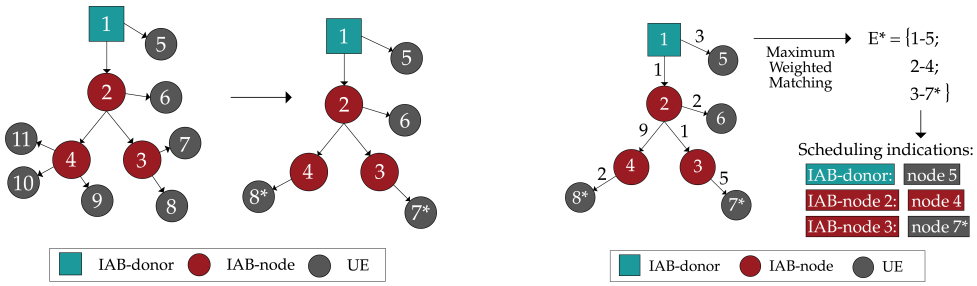
The proposed framework can be subdivided into the following phases, which are periodically repeated every T_{alloc} subframes:

1. **Initial setup**. This step, which is depicted in Figure 3.8a, consists in the computation of the simplified IAB network graph $\mathcal{G} \equiv \{\mathcal{V}, \mathcal{E}\}$. Specifically, after this phase $\mathcal{V}$ comprises the donor and the various IAB-nodes. Accordingly, $\mathcal{E}$ contains their active cell associations.

2. **Information collection**. During this phase, the various IAB-nodes send to the central controller a pre-established set of information for each of their children in $\mathcal{G}$. For instance, this feedback may consist of their congestion status and/or information regarding their channel quality. To such end, our implementation uses modified versions of pre-existing NR Release 16 Control Elements (CEs), as strongly recommended in the IAB SI [39]. However, the scheme does not actually impose any limitations in such regard.

3. **Centralized scheduling indication**. Upon reception of the feedback information, the central controller updates $\mathcal{G}$ by inspecting the received node-parent associations. Then, the set of weights $\mathcal{W}$ is calculated and an MWM of $\mathcal{G}$ is computed, using the `T-MWM` algorithm. The output of this procedure is the activation set $\mathbf{E}^*$, which yields a globally optimum solution with respect to the chosen utility function. Subsequently, $\mathbf{E}^*$ is used as to create a set of *favored* downstream nodes, i.e., of children which will be served with the highest priority by their parent, as depicted in Figure 3.8b. Finally, these scheduling indications are forwarded to the various IAB-nodes which act as parents in the edges of $\mathbf{E}^*$.

4. **Distributed scheduling allocation**. During this phase, the various IAB-nodes make use of the indications received by the central controller, if available, in order to perform the actual scheduling (which is, therefore, predominantly distributed). Specifically, the favored nodes are served with the highest priority, while the remaining downstream nodes are scheduled



(a) The original topology, exhibiting the actual cell attachments, is depicted on the left. Conversely, the reduced one is shown on the right.

(b) Computation of the MWM and of the corresponding scheduling indications.

Fig. 3.8: High level scheme of the initial setup and centralized scheduling indication phases.

if and only if the resource allocation of the former does not exhaust the available OFDM symbols.

It is important to note that since $\mathcal{G}$ contains only the IAB-nodes, the donor and at most one "representative" UE per gNB, the proposed scheme effectively performs only the backhaul/access resource partitioning in a centralized manner. On the other hand, the actual MAC-level scheduling is still undertaken in a distributed fashion, albeit leveraging the indications produced by the central controller. The major advantages which this two-tier design exhibits, compared to a completely centralized solution, are the presence of a relatively light signaling overhead and the ability to promptly react to fast channel variations, for instance caused by small scale fading.

Implementation of semi-centralized allocation schemes in mmWave IAB networks

The remainder of this section discusses how the proposed scheme can be implemented in IAB deployments, with reference to how the 3GPP specifications can support it. Moreover, an in-depth analysis of the framework's communication overhead and computational complexity is provided. To such end, let $\mathcal{G} = \{V, E\}$ be the reduced network graph, computed as per 3.5.2, and, conversely, let $\bar{\mathcal{G}} = \{\bar{V}, \bar{E}\}$ comprise all the nodes in the IAB network.

In general, the resource allocation framework requires (i) a central controller, which is installed on the IAB-donor, or could be deployed in a RAN Intelligent

Controller (RIC) following the O-RAN architecture [118]; and (ii) a scheduler which exchanges resource coordination information with the former. In particular, and referring to the aforementioned phases of the proposed scheme, the following additional considerations can be made.

Initial setup During this phase, which takes place when the IAB-nodes perform their first connection to the network, the controller acquires preliminary topology information by leveraging the configuration messages which are already exchanged during the typical Rel. 16 Initial Access (IA) procedure [39, Section 9.6]. Therefore, no additional overhead is introduced. Specifically, a map which associates each IAB-node in the network to a list of its edges, identified by global identifiers (which from now on will be referred to as "IDs"), is computed. As a consequence, $\mathcal{O}(|V|)$ insertions in a sorted map are performed and this one-time setup exhibits a computational complexity of $\mathcal{O}(|V|\log(|V|))$.

Information collection The generation of the feedback information is performed in a distributed manner by the gNBs. To such end, the current implementation features the forwarding of information on the channel quality and buffer status, in the form of Channel Quality Informations (CQIs) and Buffer Status Reports (BSRs) respectively. This choice is driven by both the will to maximize the re-utilization of the NR Rel. 16 specifications and the goal of making use of MAC-level CEs only, hence avoiding the introduction of any constraint regarding the supported IAB-relaying architecture. In particular, the CQI and BSR information is generated by analyzing the corresponding CEs, which are already received by the scheduler of each gNB, and checking whether the source Radio Network Temporary Identifier (RNTI) belongs to an IAB-node or to a UE. In the first case, the corresponding ID is retrieved and an entry carrying such identifier along with its CQI/BSR value is generated. The feedback information concerning the UEs, instead, is averaged in the case of the CQIs and added up for the BSRs, to obtain a single value for each gNBs.

Referring to the 3GPP specifications of [119], the buffers occupancy can then be forwarded to the IAB-donor by introducing a Short BSR, which carries a single Logical Channel Group (LCG) ID and its respective buffer size. This is motivated by the fact that we do not keep track of per-flow information, i.e., we aggregate

all the different RLC bearers into a single measurement report. Similarly, the channel qualities can be reported by the various IAB-nodes via an additional CQI-only Channel State Information (CSI) report, based on a Wideband (WB) measurement. Therefore, we can upper bound the size of these CEs as 11 [119] and 7 bits [120] respectively. Regarding the computational complexity, in this phase we generate, at each gNB, one CQI and one BSR for each backhaul link, and (possibly) compute one cumulative CQI and BSR for the UEs. Therefore, the asymptotic complexity of this phase can be identified as $\mathcal{O}(|V|)$.

Centralized scheduling indication During this phase, the controller makes use of the feedback received from the gNBs to update the topology information, compute the weights of the various network links and generate the centralized scheduling indications.

Regarding the former, no additional control information is required. In fact, the periodic feedback received from the various IAB-nodes, which carries a list of ID-value pairs, can be used for this purpose. In particular, the controller checks the child-parent associations for discrepancies with its local knowledge, and, if so, updates the stored associations. Discrepancies can arise under two circumstances: the connection of the first UE to an IAB-node and the handover to a different parent of any IAB-node. In the first case, just the corresponding "cumulative access node" needs to be added to the aforementioned map. On the other hand, whenever a backhaul link changes, the topological information for the whole subtending tree must be updated. Since in the worst case this might require an update of the whole map, the asymptotic complexity of the topology information update is $\mathcal{O}(|V|)$. Thanks to this periodic update, our framework is robust with respect to Radio Link Failures (RLFs) and handovers, which may occur due to blockages or mobility of UEs and, possibly, gNBs.

For the computation of the weights for the MWM problem, we propose the following policies:

1. **Max Sum-Rate (MSR)**. This policy maximizes the overall PHY-layer throughput, i.e., the utility function is

$$f_u^{\mathrm{MSR}} \triangleq \sum_{e_{i\to k}\in\mathbf{E}^*} c_{i,\,k},$$

and the weight assigned to the edge from node i to node k is $w_{i,k} \triangleq c_{i,k}$, where $c_{i,k}$ is the capacity of the link $e_{i \to k}$.

2. **Backlog Avoidance (BA)**. This resource partitioning strategy aims at avoiding congestion. Therefore, the system utility is:

$$f_u^{\mathrm{BA}} \triangleq \sum_{e_{i \to k} \in \mathbf{E}^*} q_{i,k},$$

where the weight $w_{i,k}$ is $q_{i,k}$, namely, the amount of buffered data which would reach its next hop in the IAB network by crossing the link $e_{i \to k}$.

3. **Max-Rate Backlog Avoidance (MRBA)**. This represents the most balanced option among the three, since it exploits favorable channel conditions while also preventing network congestion and favoring network fairness. The weight assigned to link $e_{i \to k}$ is:

$$w_{i,k} \triangleq c_{i,k} + \eta \cdot q_{i,k} \cdot \left(\frac{\mu}{\mu_{thr}} \right)^k,$$

where η, μ_{thr} and k are arbitrary parameters and μ represents the number of subframes which have elapsed since the last time edge $e_{i \to k}$ has been marked as favored.

Regardless of the specific policy used, the computation of the weights exhibits a complexity which is linear in the number of edges $|E|$.

Once the weights are computed, the controller obtains an MWM of the network via an implementation of the aforementioned `T-MWM`. The algorithm outputs the activation set $\mathbf{E}^*$, i.e., a map associating the ID of the parent gNBs to the one of their favored downstream node. Moreover, $\mathbf{E}^*$ is also used by the controller in order to keep track of which link has not been favored and for how long; this information may then be used to introduce a weight prediction mechanism, improving the robustness of the scheme with respect to the information collection period. In terms of overhead, the reporting of $\mathbf{E}^*$ to the gNBs would feature as payload just one C-RNTI per IAB-node (at most, since some nodes might not receive any whenever they are not active in the specific MWM solution). In fact, by exploiting the Backhaul Adaptation Protocol (BAP), we can encapsulate this

payload as part of a BAP message, while the destination node is already included as part of the BAP header in the "BAP destination" field. Therefore, the payload size of the scheduling indications is 16 bits.

Finally, based on the previous considerations and the analysis of Section 3.5.3, the overall complexity of this phase is $\mathcal{O}(|E| + |E| + |V|)$, which, when considering ST topologies, is equivalent to $\mathcal{O}(|V|)$.

Distributed scheduling allocation The last phase of the resource allocation procedure consists in the distributed MAC-level scheduling. Before assigning the available resources, the various schedulers check whether any indication has been received from the controller. Based on this condition, the buffer occupancy information is then split into two groups. The first contains the BSRs related to the favored RNTI (if any), with the caveat that if the latter indicates the cumulative access link, then this set contains the BSRs of all the UEs attached to the host gNB, while the other comprises the remaining control information. The resource allocation process is then undertaken twice: first considering the set of favored BSRs only, then the remainder of these CEs. Thanks to this repeated allocation, the favored link(s) is (are) scheduled with the highest priority, while the rest of the network only gets the remaining resources. In such a way, the information received by the controller is actually used as an *indication* and not as the eventual *resource allocation*. For instance, the gNBs are free to override these indications whenever the buffer of the favored child is actually empty, due to discrepancies between its actual status and the related information available to the controller. Moreover, the actual Downlink Control Informations (DCIs) can then be generated by the various gNBs themselves (instead of being generated only by the controller and then forwarded to the IAB-nodes), hence making use of the most updated information on the channel quality and buffer status as well. In fact, the exchange of information between the IAB-nodes and the IAB-donor introduces an inevitable delay, proportional to their distance in terms of wireless hops, between the generation of the control information at a given node (BSRs and/or CQIs) and the reception of the corresponding scheduling indications computed by the controller. Thanks to the aforementioned architecture, we limit quite significantly the performance degradation caused by these possible discrepancies between the actual nodes statuses and the (slightly outdated) information

which the controller holds about them.

The computational complexity of this last phase is different from the baseline, since it requires an additional MAC-level resource allocation. However, the specific impact of this modification is difficult to determine, since the choice of the scheduling algorithm is not part of the NR specifications. Anyhow, it is reasonable to assume that such algorithm exhibits an asymptotic complexity which is at least linear in the number of users N to be scheduled, i.e., the number of computational steps is $\mathcal{O}(N^\alpha) \mid N \in \mathbb{N}^+; \alpha \in \mathbb{R}, \alpha \geq 1$. Furthermore, it can be observed that in our framework, the two allocations receive as input disjoint subsets of the links; let $J, K \mid J, K \in \mathbb{N}; J + K = N$ be their respective sizes. Therefore, the number of operations required for the scheduling can be estimated as $\mathcal{O}((J+K)^\alpha)$ for the typical network operation and $\mathcal{O}(J^\alpha + K^\alpha)$ when using our framework. Since the following holds:

$$(J+K)^\alpha \geq J^\alpha + K^\alpha \;\; \forall \alpha \in \mathbb{N}^+$$

we can claim that, under the aforementioned assumptions, the last phase of the proposed framework introduces no computational overhead with respect to the typical network operation.

In addition to the previous considerations, we also need to take into account that, if no modifications to the Rel. 16 NR specifications are introduced, a set of MAC and BAP headers would also be added to the aforementioned payload estimates; their respective sizes can be estimated as 16 [119] and 46 [121] bits, respectively. Accordingly, the worst-case *overall* network overhead can be estimated as follows. During phase 2, for each backhaul link in the network and towards the controller, up to two BSRs and CQIs are exchanged, originating from the link's parent and child, respectively. Moreover, for each IAB-node in the network, one BSR and one CQI are exchanged for the (possible) "cumulative" access link. Then, in phase 3 the controller sends up to one scheduling indication per IAB-node. Letting then N be the number of IAB-nodes which are connected to the same IAB-donor, the communication overhead can be upper bounded by $N \cdot (2+1) \cdot (65+69) = 402 \cdot N$ [bits] in the UL and $76 \cdot N$ [bits] in the DL.

Notably, the 3GPP also considered the possibility of realizing heterogeneous IAB deployments [39], in which IAB-nodes hold an additional connection with a

macro cell (ideally co-located with the IAB-donor) to handle the control plane. In this context, our framework can be enhanced by carrying feedback information (i.e., CQIs and BSRs) and scheduling indications over the additional connection, reducing the overhead and avoiding the need to travel through multiple hops before reaching the IAB-donor.

We implemented the proposed resource allocation scheme in the popular open source simulator ns-3, exploiting the mmWave module [46] and its IAB extension [92], to characterize the system-level performance of the proposed solution with realistic protocol stacks, scenarios, and user applications.

The ns-3 mmWave module is based on [122] and features highly customizable PHY and MAC layer implementations, with an NR-like flexible OFDM numerology and frame structure. It also includes accurate interference and error models, as well as a detailed channel model, which is compliant with the 3GPP specifications [123] and accounts for large and small scale fading phenomena, as well as for interference. Additionally, the IAB module [92] models wireless relaying functionalities which mimic the specifications presented in [39]. Specifically, this module supports both single- and multi-hop deployment scenarios, auto-configuration (within the network) of the IAB-nodes and a detailed 3GPP protocol stack, allowing wireless researchers to perform system-level analyses of IAB systems in ns-3.

It is of particular relevance to understand how the scheduling operations are implemented in the IAB module, since they offer not only the baseline for the proposed scheme, but also valid guidelines for real-world deployments. The current ns-3 IAB schedulers exhibit a Time Division Multiple Access (TDMA)-based multiplexing between the access and backhaul interfaces. Moreover, scheduling decisions are undertaken in a distributed manner across the IAB network, i.e., each gNB allocates the resources which its access interface offers (to both UEs and IAB-nodes) independently of the other gNBs in the network. In fact, in an IAB network these scheduling decisions are *almost* independent of one another: if a parent node schedules the backhaul interface of a downstream node, clearly the latter will be constrained in its own scheduling decisions, as it will not be allowed to allocate the time resources which have already been scheduled for backhaul transmissions by its parent. Therefore, in a tree-based, multi-hop wireless network the various gNBs need to know in advance the scheduling decisions

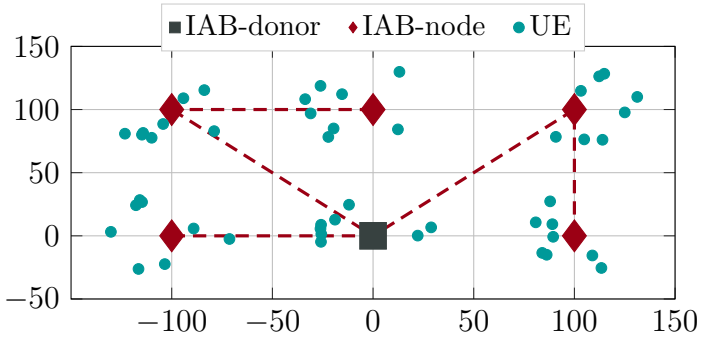


Fig. 3.9: A realization of the simulation scenario; the dotted lines represent the cell-attachments of the IAB-nodes.

performed by their upstream nodes: to solve this problem, the authors of the IAB module for ns-3 introduced a "*look-ahead backhaul-aware scheduling mechanism*" [92]. Such mechanism features an exchange of DCI between the access and backhaul interfaces: in such a way, any time resources already scheduled by the parent for backhaul communications can be marked as such by the corresponding downstream node, preventing any overlap with other transmissions. Furthermore, the *look-ahead* mechanism requires the schedulers of the various gNBs to commit to their resource allocation for a given time T at a time $T - k$, where $k - 1$ is the maximum distance (in terms of wireless hops) of any node from the donor. In such a way, the DCIs will have time to propagate across the IAB network and reach the farthest node at time $T - 1$, thus allowing its scheduler to perform the resource allocation process at least one radio subframe in advance.

Performance Evaluation

The purpose of these simulations is to understand the performance of the proposed resource partitioning framework in the context of its target deployment, i.e., a multi-hop IAB network. As a consequence, the reference scenario consists of a dense urban deployment with a single IAB-donor and multiple IAB-nodes, as depicted in Figure 3.9. In particular, the various gNBs are distributed along an urban grid where the donor is located at the origin while the IAB-nodes are deployed along the street intersections, with a minimum inter-site distance of 100 m. The IAB-nodes attachments are computed using the so-called *HQF* policy presented in [94]; however, this choice does not introduce any loss of generality

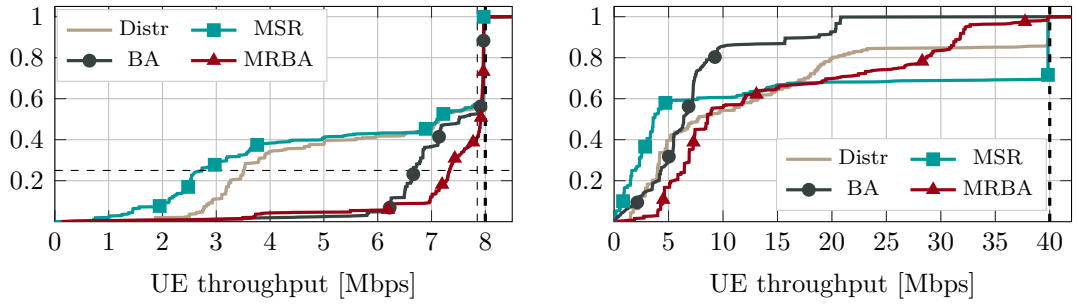
Table 3.1: Simulation parameters

Parameter	Value
Number of runs N_{runs}	25
Simulation time T_{sim}	3 s
MWM period T_{alloc}	$\{1, 2, 4\}$ subframes
Layer 4 protocol	{UDP, TCP}
UDP packet size s_{UDP}	$\{50, 100, 200, 500\}$ B
Weight policy f_u	{MSR, BA, MRBA}

since such parameter is fixed for all the runs. A given number of UEs are deployed within the surroundings of these base stations, with an initial position which is randomly sampled from circles of radius ρ and whose centers are the various gNBs. A summary of the simulation parameters in provided in Table 3.1.

Both the IAB-donor and the IAB-nodes are equipped with a phased array featuring 64 antenna elements, and transmit with a power of 33 dBm; conversely, UEs are equipped with 16 antenna elements and their transmission power is restricted to 23 dBm. Notably, the presence of additional antenna elements at the gNBs is a key (but reasonable) assumption, as it allows base stations to achieve a high beamforming gain. In turn, it is possible to achieve a high capacity, which is fundamental to avoid performance bottlenecks, given the absence of a fiber backhaul. The UEs download data which originates from sources that are installed on a remote host; both UDP and the Transmission Control Protocol (TCP) are used. For the UDP simulations, the rate of the sources is varied from 4 to 40 Mbps to introduce different degrees of saturation in the network. Therefore, in these simulations only DL traffic is considered. Finally, the performance of the proposed policies is hereby compared with the baseline of [92], indicated as "Distr" by examining end-to-end throughput, latency, and a network congestion metric.

Throughput The first metric which is inspected in this analysis is the end-to-end throughput at the application layer. As a consequence, only the packets which are correctly received at the uppermost layer of the destination node in the network are taken into account. In particular, for each UE and each simulation



(a) s_{UDP} = 100 B, i.e., UDP rate of 8 Mbps. **(b)** s_{UDP} = 500 B, i.e., UDP rate of 40 Mbps.

Fig. 3.10: Per-UE end-to-end throughput Empirical Cumulative Distribution Functions (ECDFs). The thick dashed line represents the rate of the UDP sources.

run, the long-term average throughput is computed as follows:

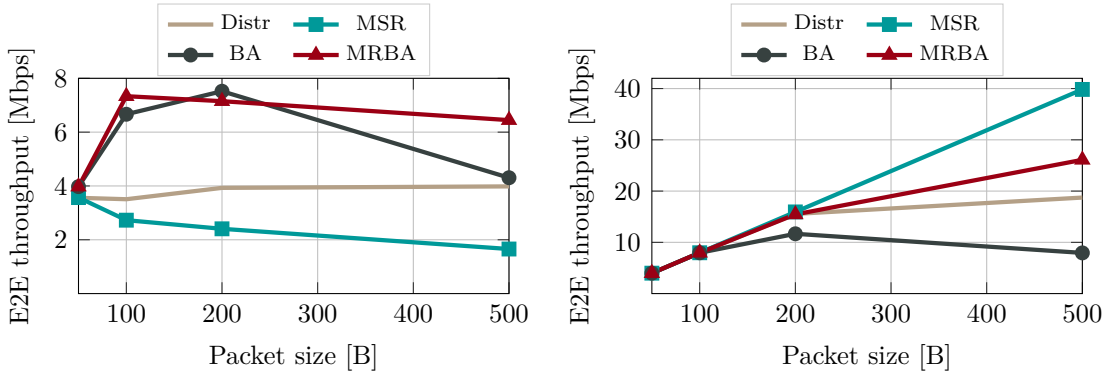
$$S_{k,n}^{\text{APP}} \triangleq \frac{B(T_{sim}, k, n)}{T_{sim}}$$

where $B(T, k, n)$ is the cumulative number of bits received up to time T by the k-th UE, during the n-th simulation run. Then, the distribution of $\mathbf{S}^{\text{APP}}$, namely, the vector containing the collection of the $S_{k,n}^{\text{APP}}$ values across the different runs and UEs, is analyzed.

Figures 3.10a and 3.10b report the ECDF of $\mathbf{S}^{\text{APP}}$, for a UDP packet size of 100 and 500 bytes, respectively, and the policies introduced in Section 3.5.3. In the former, we can notice that the introduction of the semi-centralized framework increases by up to 15% the percentage of UEs whose throughput almost matches the rate of the UDP sources, i.e., achieving approximately 7.9 Mbps. Moreover, by focusing on the leftmost portion of Figure 3.10a we can observe another interesting result, concerning the throughput experienced by the UEs which do not fulfill their QoS requirements. In fact, with respect to the first quartile the distributed scheduler and the MSR policy achieve the worst performance. On the other hand, the MRBA and BA policies significantly improve these results, even though the extent of such improvements varies quite dramatically across the two.

In particular, compared with the distributed case the BA and MRBA policies introduce a 2- and 3-fold increase of the worst case throughput, respectively, coupled with a significantly lower variance in both cases.

These results can be explained as follows: since a UDP packet size of 100 bytes does not saturate the capacity of the access links, the main performance



(a) First quartile.

(b) Third quartile.

Fig. 3.11: End-to-end throughput quartiles, for $s_{UDP} \in \{50, 100, 200, 500\}$ B.

bottleneck of this configuration is represented by the buffering of the aggregated traffic on the intermediate backhaul links. Therefore, the MSR policy provides no improvements compared to the performance of the distributed scheduler, since it simply favors the links with a higher SINR. Conversely, the prioritization of the most congested links introduced by the other two strategies successfully tackles the former problem. In particular, the BA policy exhibits the highest worst-case throughput, while also satisfying the QoS requirements of approximately 40% of the UEs. Moreover, the bias towards high SINR channels introduced by the MRBA strategy further improves the higher percentiles, compared to the BA policy, and dramatically outperforms MSR and the baseline across all percentiles.

By increasing the UDP packet size to 500 bytes, the network becomes noticeably saturated, as depicted by Figure 3.10b; in fact, in this instance only a minority of the UEs achieve a throughput which is comparable to the source rate. With this configuration, the BA strategy achieves the worst performance, providing a significantly lower throughput across most percentiles. On the other hand, both remaining strategies introduce significant improvements, although with different trade-offs. In particular, compared to the distributed case, the MSR policy exhibits an increase of approximately 20% of the number of UEs which satisfy their QoS requirements, albeit at the cost of worse lower percentiles. The MRBA, conversely, introduces performance benefits which mostly affect the bottom percentiles only. However, with this strategy only a limited portion of the UEs

achieves the target throughput of 40 Mbps. As a consequence, we can conclude that with the configuration depicted in Figure 3.10b the network approaches the capacity of the mmWave channels. Therefore, buffering phenomena are likely occurring at each intermediate IAB-node. Moreover, we can say that in a saturated network the congestion is so severe that prioritizing the bottleneck links is not enough: we also need to take into account the channel conditions and prioritize the links which not only are congested, but also have the "biggest chance" of getting rid of the buffered data due to the temporary better channel quality.

Finally, Figure 3.11 presents the first and third quartiles of $\mathbf{S}^{\mathrm{APP}}$ as a function of the UDP packet size s_{UDP}. It can be noted that, with respect to the first quartile, the MRBA outperforms all the other policies by delivering a throughput which is up to 90% higher than the one obtained by the distributed scheduler. On the other hand, Figure 3.11b shows how the best third quartile is achieved by MSR, with up to a 2-fold improvement over the distributed solution. Furthermore, we can observe how the positive impact of the BA strategy is inversely proportional to the saturation in the network. We can then conclude that the bias it introduces loses its effectiveness as the buffering phenomena start to affect the majority of the IAB-nodes.

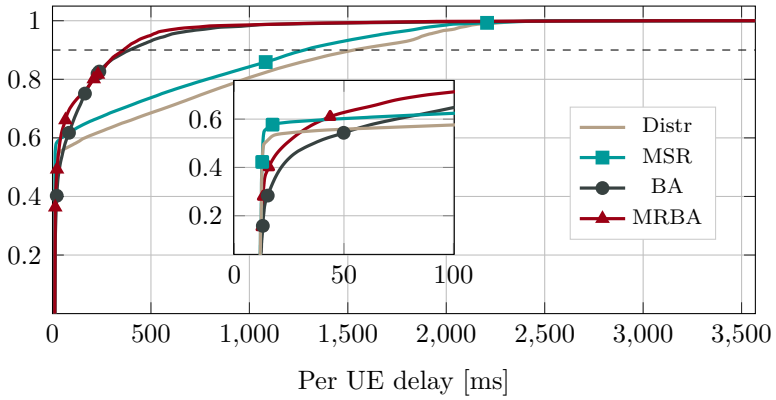
LATENCY Just like the aforementioned metric, latency is measured end-to-end at the application layer. Thanks to this choice, the resulting delay accurately represents the system-level performance, as it includes the latency which is introduced at each hop in the IAB network.

In particular, for each packet correctly received at the uppermost layer of its final destination, the following quantity is traced:

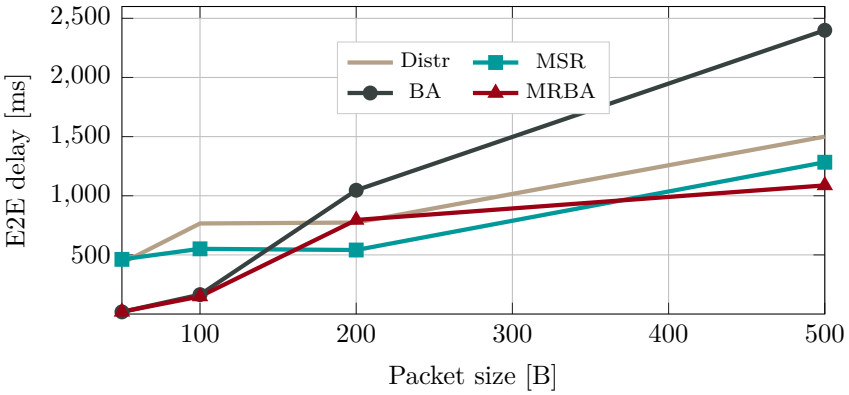
$$D_i^{\mathrm{APP}} \triangleq \sum_{l_k \in \mathcal{E}_i} D_i^{l_k}$$

where $\mathcal{E}_i$ comprises the links in the IAB network that are crossed by the i-th packet, while the term $D_i^{l_k}$ indicates its point-to-point latency over the path link l_k. Finally, these values are collected for each of the various runs into the vector $\mathbf{D}^{\mathrm{APP}}$ and its statistical properties are inspected.

Figure 3.12a shows the empirical ECDF of $\mathbf{D}^{\mathrm{APP}}$ for a packet size of 100 bytes. It can be noticed that, in this case, the 90th percentiles achieved by the BA and



(a) ECDF, for $s_{UDP} = 100$ B.



(b) Third quartile, for $s_{UDP} \in \{50, 100, 200, 500\}$ B.

Fig. 3.12: Per-UE end-to-end delay statistics.

the MRBA policies are approximately 20% smaller than the one obtained by the distributed scheduler. Moreover, these strategies manage to dramatically reduce the number of packets received with extremely high delay, i.e., in the order of seconds, showing the dramatic impact of buffering in the baseline configuration. Conversely, the MSR policy provides the best performance with respect to the best case delay only, although it still outperforms quite significantly the distributed strategy.

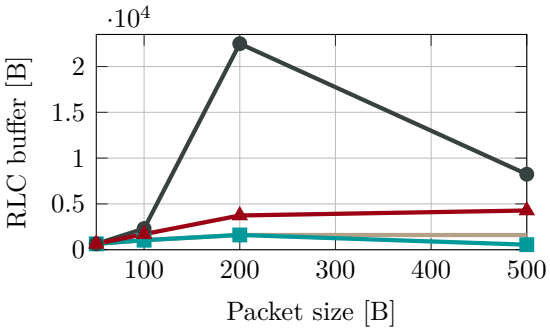
These trends are exacerbated in Figure 3.12b, which shows the third quartile of $\mathbf{D}^{\mathrm{APP}}$ as a function of the UDP packet size s_{UDP}. In fact, we can notice that the effectiveness of the BA policy is inversely proportional to the network saturation, whereas the opposite holds true with respect to the MSR strategy. It follows that, for UDP rates in the order of 5 to 10 Mbps, the network is mainly plagued by local congestion which causes the insurgence of buffering in some of the nodes. Conversely, as the rate of the UDP sources increases the system shifts to a capacity-limited regime, a phenomenon which explains the dominance of the MSR and MRBA policies.

Network congestion Network congestion is measured by collecting, every T_{alloc} subframes, the RLC buffers status of the various nodes into the vector $\mathbf{B}^{\mathrm{RLC}}$. It must be noted that, since RLC Acknowledged Mode (AM) is used, these values will indicate data which is related to both new packets and possible retransmissions.

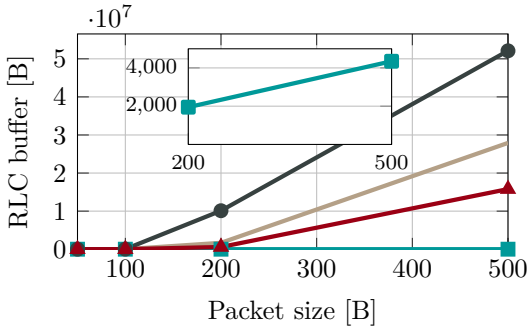
Figs. 3.13b and 3.13c show the median of $\mathbf{B}^{\mathrm{RLC}}$, for traffic flows whose next hop in the network is represented by either UEs or IAB-nodes respectively. Specifically, the BA strategy achieves the worst performance in this metric, leading to unstable systems in the cases of $s_{UDP} = \{200, 500\}$ B. A reason for this behavior can be found in the "locality" of the BA policy criteria and the lack of influence of the past allocations on the weights. These characteristics may lead to favoring the same link repeatedly, hence offering little remedy to the end-to-end congestion.

On the other hand, the buffer occupancy achieved by the MSR strategy reiterates the fact that this policy is progressively more effective as the source rate increases and the system becomes more congested. In particular, a dramatic decrease of up to 4 orders of magnitude is achieved for $s_{UDP} = 500$ B. Finally, when compared to the distributed scheduler, the MRBA policy also achieves a lower

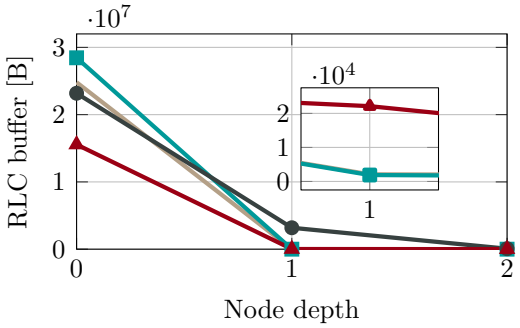
(a) Legend.



(b) Medians, toward UEs.

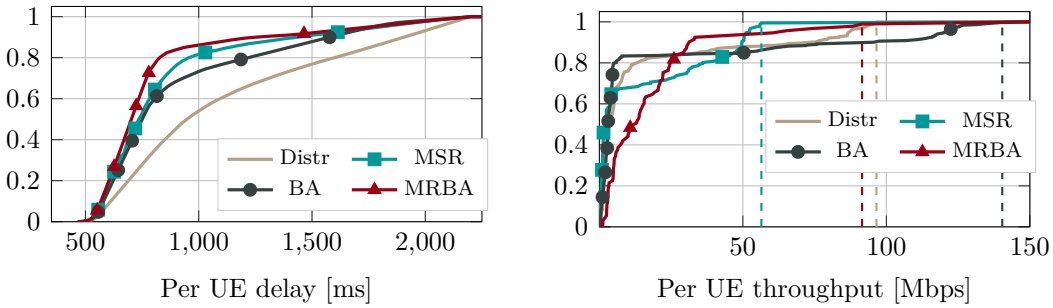


(c) Medians, toward IAB-nodes.



(d) Third quartile vs. depth in the IAB network, for $s_{UDP} = 200$ B.

Fig. 3.13: Buffer occupancy statistics, for $s_{UDP} \in \{50, 100, 200, 500\}$ B.



(a) Delay ECDF.

(b) Throughput ECDF.

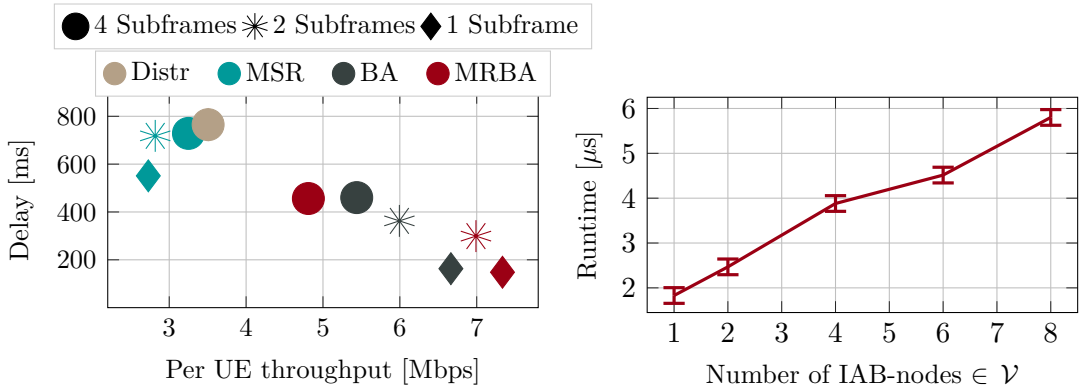
Fig. 3.14: End-to-end delay and throughput statistics, for TCP layer 4 protocol.

median RLC buffer occupancy towards the backhaul links, albeit the difference is less striking than in the case of the MSR policy, and at the cost of slightly more congested UE buffers.

Additionally, Figure 3.13d depicts the third quartiles of $\mathbf{B}^{\mathrm{RLC}}$ as a function of the depth of the corresponding gNB in the IAB network. It is possible to notice that, regardless of the policy in use, the amount of buffering at the various gNBs generally decreases as their distance to the donor increases. This follows from the fact that nodes which have a lower depth exhibit, on average, a bigger subtending tree; therefore the amount of traffic which makes use of their backhaul links is significantly higher.

Performance with TCP traffic This subsection extends the aforementioned analysis by inspecting the performance of the proposed scheme in the case of TCP traffic. Specifically, a TCP full-buffer source model is used, and the various semi-centralized resource allocation policies are compared against the distributed scheduler.

Figure 3.14a shows the ECDF of the end-to-end delay experienced by the successfully received packets. Similarly to the UDP case, the distributed scheduler exhibits the worst performance In fact, the benefits introduced by the semi-centralized policies are noticeable across all percentiles. In particular, with this configuration the MRBA policy provides the best results, followed quite closely by the BA and MSR strategies. Figure 3.14b, which depicts the statistics of the end-to-end throughput achieved by the various UEs, further explains the effect on the system of the various semi-centralized policies. In particular, the BA pol-



(a) Combined per UE end-to-end throughput first quartile and delay third quartile, as a function of the semi-centralized allocation period T_{alloc}.

(b) MWM runtime as a function of the number of IAB-nodes in the network.

Fig. 3.15: Considerations on the formulated assumptions.

icy achieves, approximately, a 45% increase of the peak throughput. Conversely, the MRBA strategy causes a redistribution of the achieved data rate, massively improving the lower quartiles (up to the 80-th), albeit at the expense of the maximum throughput. Finally, MSR also causes a redistribution of the throughput across the different percentiles, but the net benefit is less noticeable.

Therefore, we can conclude that regardless of the specific policies used, the proposed scheme improves the system performance also with this configuration, by limiting the insurgence of local buffering and aiding the end-to-end congestion control mechanism offered by TCP. Furthermore, it can be noted that a prioritization of the most congested links and of the channels featuring a higher quality results in performance benefits in the average case, although it also causes a decrease of the network fairness. On the other hand, the MRBA policy manages to optimize the backhaul/access resource partitioning, while at the same time introducing an increase in the throughput fairness.

FURTHER CONSIDERATIONS It is of particular relevance to analyze the performance of the semi-centralized policies when relaxing the most restrictive hypothesis, i.e., the capability of exchanging feedback information in a timely manner. Such analysis provides also insights regarding the effects of errors and/or crashes in the control messages. Indeed, both control and data channels implement error detection mechanisms, making the likelihood of undetected errors in the feedback

information negligible. As a consequence, the errors would be detected at the receiver, and lost information would be either retransmitted by the source or simply discarded, waiting for the following periodic update; in both cases, the net effect would be a delay in the reception of the message.

To such end, Figure 3.15a shows the performance of the proposed framework as a function of the semi-centralized allocation period T_{alloc}. In particular, each of the depicted points represents the joint end-to-end throughput and delay achieved with the different configurations.

As expected, in general the effectiveness of the various semi-centralized policies progressively deteriorates as the frequency of the scheduling indications decreases. Interestingly, the BA policy exhibits the lowest performance degradation with respect to an increase of the allocation period, which suggests that this phenomenon has a slower evolution over time compared to the one exhibited by the channel quality. Nevertheless, the key takeaway is that all the proposed allocation strategies except MSR outperform the distributed solution, across both metrics. In fact, the latter exhibits the lowest throughput first quartile, but only because it introduces a strong bias towards high SINR channels, as discussed in Section 3.5.3. However, the trend depicted by Figure 3.15a also suggests that there exists a threshold value of T_{alloc} after which the performance of the proposed frameworks brings only marginal performance benefits.

Additionally, the running time of the MWM algorithm presented in Section 3.5.3 was analyzed, in order to understand whether it may partially invalidate the timely feedback assumption. Specifically, Figure 3.15b presents the statistics of the various MWM execution times, obtained on a machine equipped with an i7-6700 4-core processor clocked at 3.4 GHz. The first observation which can be made is that this empirical analysis confirms the previously estimated asymptotic complexity, depicting a running time which exhibits a linear dependence on the number of gNBs in the network. Furthermore, it can be noted that the runtime of the MWM algorithm does not exceed 6 μs, even for a significant number of IAB-nodes connected to the same IAB-donor. As a consequence, we can conclude that the execution times of the semi-centralized allocation process do not pose any threat to the timely feedback assumption, since they are reasonably smaller than the duration of the minimum semi-centralized allocation period.

3.6 Conclusions and Future Work

High-density deployments of 5G mmWave cells require innovative solutions to reduce costs without degrading the end-to-end network performance. IAB has been investigated to relay access traffic to the core network wirelessly, thereby removing the need for fiber backhaul in all the gNBs. In this chapter, we reviewed the latest 3GPP NR Release 16 standardization activities on IAB and evaluated the performance of IAB networks for different applications and traffic types. IAB represents a viable solution to efficiently relay cell-edge traffic, although the benefits decrease for more congested networks. We have also highlighted the limitations of IAB and provided guidelines on how to overcome them.

Moreover, we proposed a semi-centralized resource partitioning scheme for 5G and beyond IAB networks, coupled with a set of allocation policies. We showed that the introduction of this light resource allocation cooperation dramatically improves the end-to-end throughput and delay achieved by the system, preventing (or at least limiting) the insurgence of network congestion on the backhaul links. Specifically, the MRBA policy exhibits the most promising results, offering up to a 3-fold increase in the worst-case throughput and approximately a 30% smaller worst-case latency, compared to the distributed scheduler. On the other hand, the effectiveness of the BA and MSR policies varies quite significantly across the specific system configuration and inspected metric.

We provided considerations on the implementation of a semi-centralized resource allocation controller in real world deployments. In particular, we acknowledged that the proposed scheme relies on the assumption of IAB-nodes being capable of exchanging timely feedback information with the IAB-donor. Even though the amount of signaling data which the proposed solution requires is quite low, and its performance is quite robust with respect to an increase of the central allocation period, we argue that this remains a significant constraint. Moreover, such drawback is exacerbated by the unfavorable mmWave propagation characteristics. As a consequence, we deem that solutions involving a central controller, which rely on the timely exchange of control information with the IAB-donor, are likely to require dedicated control channels, possibly at sub-6 GHz, in order to grant the utmost priority and reliability to the feedback information. Therefore, we can conclude that the aforementioned framework can bring dramatic perfor-

mance benefits to IAB networks, although its introduction in 5G and beyond deployments requires additional research efforts.

For this reason, as part of our future work we plan to design machine-learning algorithms which predict the network evolution at the IAB-donor. This improvement will allow us to relax the timely feedback assumption, by increasing the minimum semi-centralized allocation period which leads to performance benefits over distributed strategies. Moreover, we foresee to implement mechanisms which adapt the parameters of the MRBA policy to the system load and configuration, and additional resource partitioning strategies. Finally, the generalization of the proposed framework to SDMA systems will be studied. The use of this multiple access scheme should significantly improve the performance of mmWave wireless backhauling by introducing the possibility of concurrently serving multiple terminals, provided that they exhibit a sufficient distance among them.

4

Full-stack Evaluation of Hybrid Beamforming in 5G mmWave Networks

4.1 Introduction

mmWave systems make use of antenna arrays to focus the transmit power into the desired direction and then achieve higher antenna gains. This technique, called BF, enables to compensate for the high pathloss experienced at mmWave frequencies. Different BF architectures have been considered in the literature. With analog BF, the transceivers have a single RF chain, and a single beam is generated using analog phase shifters in the N antenna elements of the phase array. More advanced transceivers use hybrid or digital BF architectures, with $K \leq N$ RF chains. While increasing the complexity and power consumption of the device, they enable a finer control on the BF process, which can be based on combined digital and analog processing [124].

Hybrid and digital BF architectures, therefore, are capable of steering multiple beams from a single antenna array, with (possibly) independent data streams*, effectively enabling MU-MIMO operations at mmWaves [125]. As a result, this

*In NR, each component signal of a spatial multiplexing transmission is called a "layer," whereas in the physical layer literature it is often called "stream." We adopt the second term to avoid confusion with the protocol layers.

increases the network spectral efficiency, as different users can be served with SDMA in the same time and frequency resources. HBF solutions, in particular, are considered as a cost- and energy-effective solution for MU-MIMO at mmWaves, and have been practically implemented and deployed in commercial devices [41].

3GPP NR natively supports MU-MIMO transmissions [126]. However, despite the promising features of HBF at mmWaves, the state of the art currently lacks an analysis of how a *physical layer* based on HBF interacts with the *full protocol stack*, from the MAC layer (e.g., for scheduling) to the transport layers and application. Although physical layer performance studies play a critical role, the actual quality perceived by the users can only be measured at the application layer. The ultimate comparison of different solutions must involve higher-layer metrics within a full protocol stack framework. This chapter analyzes the integration of HBF in the protocol stack of 5G and beyond cellular networks, focusing on the interplay between well-established beam design methods and higher layers for different scenarios and applications.

The remainder of the chapter is organized as follows. Section 4.2 summarizes the state of the art on HBF and scheduling for mmWave networks. Section 4.3 discusses HBF and scheduling design in 3GPP NR networks. Section 4.4, then, describes the performance evaluation results, and Section 4.5 concludes the chapter, providing suggestions for future research directions.

4.2 State of the Art

MU-MIMO communications have received considerable interest since the 90s [127], which increased even further as the number of available antennas per device grew, harnessing the advantages of *massive MIMO* [128]. Nowadays, MU-MIMO with very large antenna arrays is an integral part of the 5G NR cellular standard published by the 3GPP [16, 129, 130]. MU-MIMO pilots, channel estimation and feedback procedures in the standard are concisely described in [131]. Beam-management procedures in 5G are covered in [126].

Signal processing techniques for mmWave systems are reviewed in [40, 125]. Concerns about the power consumption in large antenna arrays motivate the use of either HBF architectures [132] or fully digital BF with low-resolution Analog to Digital Converters (ADCs) [133]. We focus on HBF and leaves low-resolution

digital BF for future work. The design of Minimum Mean Squared Error (MMSE) beamforming for single user mmWave HBF under frequency-flat channels is studied in [134, 135, 136, 124]. The extension to OFDM with frequency-selective channels is covered in [137]. The extension to MU-MIMO SDMA has also been well studied in physical layer works such as [138].

There is also significant work that has studied mmWave SDMA scheduling under *physical layer simplifications* that permit the analytical treatment of the scheduling problem in mathematical form. For example, it is common to assume that the number of users is less than or equal to the number of available RF chains, which is rare in a real deployment. A number of models assume some pre-existing Single-User Multiple-Input Multiple-Output (SU-MIMO) beamforming method and find which users can co-exist in the same time slot using sufficiently separated beams [139, 140, 141, 142, 143, 144, 145, 146, 147, 148]. Motivated by the concept of "favorable propagation" in massive MIMO [128], these works lessen the burden on the physical layer and increase the burden on the scheduler. Making the scheduler responsible for avoiding inter-user interference prevents these models from taking full advantage of physical layer MU-MIMO state-of-the-art techniques. Moreover, this reduces the scheduler's flexibility to make decisions based on user traffic needs. Also, it is not certain that practical mmWave HBF can be sufficiently close to asymptotic massive MIMO regimes. Instead, in our model we assume that the scheduler makes decisions based on user traffic needs and channel gains only, as in classic scheduling literature, and trusts that the physical layer will cancel the interference with methods such as [138]. Due to the large scope of the problem, joint beamforming and scheduling is a hard problem. A simplified analysis using a tractable convex lower bound is given in [149]. In comparison, we adopt a full-stack approach to show that the well-known Round Robin (RR) scheduling policy can be combined with well-known MMSE MU-MIMO beamforming, revealing novel issues in the interplay of both schemes that need to be addressed for the system to work. The full-stack approach also makes it possible to understand the effects on the MU-MIMO performance of a number of other NR characteristics such as the frame structure, the DeModulation Reference Signal (DMRS), the retransmissions, different traffic or transport protocols, etc.

Novel applications, protocols and deployments give rise to the need for evaluating the end-to-end performance of 5G mmWave networks [46, 150, 151, 152, 153,

47, 154, 155, 156, 157, 158, 159, 160, 161, 162, 123, 52]. These evaluations must go beyond physical layer capacity evaluations [163, 164, 165, 166]. For example, [153] investigates the behavior of the TCP on top of mmWave links. The existing literature on end-to-end mmWave cellular network performance, however, has mostly focused on the interaction between the higher layers of the protocol stack and a physical layer with analog BF [47, 151]. These works used geometric (i.e., positional) beam designs, and did not support the advanced HBF schemes that are well-known in physical layer literature, nor the MU-MIMO SDMA techniques that support active links between the base stations and multiple users simultaneously. This work introduces the novelty of SDMA and MMSE MU-MIMO beamforming to prior full-stack end-to-end mmWave evaluations, building on the ns-3 mmWave simulation module [46, 150, 151, 152, 153, 47, 154, 155, 156, 157, 158, 159, 160, 161, 162, 123, 52], and provides new key insights on how their interactions with the higher protocol layers affect the performance experienced by the user.

4.3 Full-stack Integration of HBF in mmWave Networks

The 3GPP 5G standard specifies the NR waveform and device requirements [16, 167]. The interface between the waveform and the antenna array is standardized through a series of "antenna ports." The details of BF operations applied to signals in each port are left to the vendor implementation, and constrained only by conformance requirements such as those in [167].

In the NR waveform, complex symbols are mapped in a 3-dimensional OFDM resource grid comprising the OFDM symbol number in time (n), the OFDM subcarrier number in frequency (k), and the "*SDMA stream*" number (ℓ) [16, Table 7.3.1.3-1]. Furthermore, streams can be mapped to more than one antenna port (p) using several precoding configurations [16]. These antenna ports are defined as signal input/output interfaces to the antenna array. The mapping of antenna ports to actual antenna elements may be vendor-specific, but must guarantee that in the same port and within the same "slot" of 14 consecutive OFDM symbols, the channel may be inferred through DMRSs that are orthogonal and specific in each port [129].

In NR limited feedback beamforming [134], the gNB may transmit separate

reference signals using different beams at different times or on different subcarriers. The UE observes the reference signals and reports the "best" beam ID back to the gNB. The channel during a slot of 14 consecutive OFDM symbols can be inferred from a single DMRS, and therefore we can assume that channel coefficient estimation is performed once at the beginning of each transmission. We assume an idealized version of the beam selection and DMRS procedures in which the true best beam and channel values are revealed to the UE and gNB in each allocation. Due to mmWave channel sparsity and recent advances in estimation [168, 169], the rate in real systems would be close to its upper bound obtained by assuming noiseless channel observations. Moreover, it would be possible to further extend our model by assuming noisy channel estimation and quantized feedback, which we leave for future work.

At the NR MAC layer, the scheduler assigns radio Resource Blocks (RBs) in the grid with indices (n, k, ℓ). The broadest time division are *frames* of 10 ms duration. Each frame is divided in 10 *subframes* of 1 ms. These large scale time units are similar to LTE, and can admit either TDD or FDD configurations. However, differently from LTE, slots in any subframe can be labeled as Downlink (DL), Uplink (UL) or *flexible*, where the latter represent an innovation in 5G that permits the scheduler to dynamically change the DL/UL division over consecutive subframes.

In the smaller time scale of the OFDM signal, the dimensions of the time-frequency grid depend on the fundamental *numerology* parameter $\mu \in \{0, 1, 2, 3, 4\}$. OFDM symbols are grouped in *slots* of 14 symbols, such that each subframe has 2^μ slots. The inter-carrier spacing is $\Delta f = 2^\mu \times 15$ kHz, and each OFDM symbol's duration is $\frac{2^{-\mu}}{14}$ ms. The maximum bandwidth without carrier aggregation is 400 MHz. In addition, the smallest scheduling granularity in NR is the "mini-slot," which can be only 2 OFDM symbols long and does not necessarily have to be time-aligned with multiples of the nominal slot start instants. This allows NR schedulers to assign "asynchronous" transmissions that do not start at the same time.

The NR versatile waveform supports different frequencies, from the conventional 700 MHz–6 GHz spectrum up to the 24–70 GHz mmWave spectrum. As a result, some NR options are not useful when operating at mmWaves. One is Orthogonal Frequency Division Multiple Access (OFDMA), since highly direc-

tive mmWave BF typically requires frequency-flat analog operations. This means that the radio device cannot apply different analog beams to different subcarriers of the same OFDM symbol. Therefore, in mmWave all subcarriers k in a port-symbol pair need to be assigned to the same user. As a consequence, scheduling in mmWave NR reduces to a 2-dimension TDMA and SDMA grid (n, ℓ). A second NR option of low interest in mmWave is the use of SDMA transmissions with more than one stream to the same user, since typical mmWave MIMO channel matrices are rank deficient (i.e., the second largest eigenvalue is much smaller than the first) [125]. This would make multiple transmissions with rank ≥ 2 to the same user ineffective, thus SDMA can only be implemented as an MU-MIMO technique, but not as an SU-MIMO technique.

4.3.1 HBF Design

We assume that simultaneous SDMA transmissions are allocated into different SDMA streams ℓ. We assume each user can only receive one stream. The DL normalized physical mmWave channel matrix between the BS with N_t transmit antennas and each UE u with N_r receive antennas, in OFDM symbol n and subcarrier k is $\mathbf{H}_u[n,k] \in \mathbb{C}^{N_t \times N_r}$, satisfying the norm $\|\mathbf{H}_u[n,k]\|_F^2 = N_t N_r$. In DL, the BS selects a wideband analog BF vector for each transmit stream $\mathbf{v}_\ell$ using some BF scheme, and the UE receives with the analog BF vector $\mathbf{w}_u$. Thus the *effective* scalar complex channel between the transmit stream ℓ and the single-stream receive antenna port of the UE is given by

$$g[u, \ell, n, k] = \mathbf{w}_u^T \mathbf{H}_u[n,k] \mathbf{v}_\ell,$$

and the UL channel is computed with the transposed channel matrix and swapping transmitter and receiver beamforming vectors, resulting in the same complex scalar number.

We follow a SINR-based point-to-point link performance model. For each link we obtain an analytical expression for the instantaneous SINR for each subcarrier k at each time instant n. For each allocated segment with a start time n_o and end time n_e, we map the set of all the instantaneous SINRs, $\{SINR[n,k]\}_{n \in \{n_o \ldots n_e\}, k \in \{0 \ldots K-1\}}$, to a single Block Error Rate (BLER) value associated with the entire block, assuming that signals are decoded independently. This technique, known as Effective

SINR Mapping (ESM), is widely used in system-level simulation for 4G LTE [170], 5G mmWave [92], and 5G NR [171]. Each simulated packet transmission is randomly dropped (not passed to the receiver's upper protocols) with a probability equal to its BLER. While ESM is a simplification of real decoding hardware that makes the simulation of a large network tractable, real NR demodulation and decoding may use sophisticated joint decoding such as MU-MIMO sphere decoding [172], combined with the channel codes of NR [16]. To model SDMA with multiple streams, we write the instantaneous DL SINR of user u at time n and in subcarrier k as a function of the effective channel gains as

$$SINR_u^{DL}[n,k] = \frac{P/(KL_u)|g[u,\ell(u),n,k]|^2}{\sum_{u'\neq u} P/(KL_u)|g[u,\ell(u'),n,k]|^2 + \mathrm{OI} + \Delta f N_o}, \tag{4.1}$$

where K is the number of subcarriers, L_u is the pathloss of user u, P/K is the BS transmitted power per stream equally divided among all subcarriers, OI is the "out-of-cell" interference, N_o is the noise PSD and Δf is the inter-carrier spacing. For the discussion of BF design at a single BS, we discuss only the sum of interference over UEs connected to the same BS, producing side lobe interference between any two beams of the same radio device. If there are other BSs in the scenario, their BF design is independent and the out of cell interference OI is considered as a constant in addition to the noise $\Delta f N_o$ by the beamforming algorithm. We note that the side lobe interference terms correspond to "mismatched" beams, i.e., an interfering signal is received by user u through the channel of user u but it was sent by the transmitter stream $\ell(u')$ using a BF vector designed for user u'. This makes the side-lobe-interference power much lower than the desired signal power. Still, this interference term may be stronger than the noise, and the link SINR may be much lower than the SNR.

In UL, side lobe interference is even more severe because the UL SINR expression is in turn

$$SINR_u^{UL}[n,k] = \frac{P/(KL_u)|g[u,\ell(u),n,k]|^2}{\sum_{u'\neq u} P/(KL_{u'})|g[u',\ell(u),n,k]|^2 + \mathrm{OI} + \Delta f N_o}, \tag{4.2}$$

where the stream assigned to user u receives the interference signal transmitted by user u' through the channel of user u' and the receive beamforming vector of stream $\ell(u)$. In other words, pathloss gains in the denominator $L_{u'}^{-1}$ are also

mismatched. In some cases, such as for example if the interferer is much closer to the gNB than the desired transmitter, the pathloss gain of the interferer may be much greater than the pathloss gain of the desired signal L_u^{-1}. Thus, SU-MIMO BF strategies that only focus on maximizing $|g[u, \ell(u), n, k]|^2$ cannot guarantee that the side lobe interference power is lower than the desired signal power.

Under our ESM link model, we considered several classic design methods in the literature to obtain BF vectors to achieve good link SNR or SINR values in Equation (4.1) and Equation (4.2). We select a sample representation of linear algoritms that have an intuitive relation to the ESM model, whereas we remark that the issues we identify regarding their interaction with realistic scheduling and traffic are fully general to any MU-MIMO physical layer and have not been previously reported.

Geometric Beamforming

This was the only technique implemented in the previous version of the mmWave ns-3 module [52] we extended. Geometric BeamForming (GBF) displays worse performance than the other schemes we adopted, as detailed in the next subsection. We denote the antenna array response vector $\mathbf{a}(\theta, \phi)$ as a function that depends on the angles of azimuth (θ) and elevation (ϕ). For example, in a transmit Uniform Planar Array (UPA) with $N_1 \times N_2$ antennas separated by half a wavelength, $\mathbf{a}(\theta, \phi)$ is a vector of size $N_t = N_1 N_2$ whose i-th coefficient we denote as $a_i(\theta, \phi)$, satisfying

$$\begin{aligned} a_i(\theta, \phi) &= e^{-j\pi((i \mod N_1)\sin(\theta) + \lfloor i/N_1 \rfloor \sin(\phi))} \\ &\forall i \in \{0, \dots N_1 N_2 - 1\}. \end{aligned} \tag{4.3}$$

Notably, we can easily adopt the vector $\mathbf{a}(\theta, \phi)^H$ to design a beam that points in the direction (θ, ϕ). In GBF, the vectors are simply selected by pointing the array in the physical direction between the BS position and the UE position. That is, we assume the devices have some location hardware, and their coordinates

(x_{BS}, y_{BS}, z_{BS}) and (x_u, y_u, z_u) are known, producing

$$\begin{aligned}
\theta_D &= \arctan \frac{y_u - y_{BS}}{x_u - x_{BS}} + \pi \mathrm{I}_{(x_u < x_{BS})} \mod 2\pi, \\
\phi_D &= \arctan \frac{z_u - z_{BS}}{\sqrt{(y_u - y_{BS})^2 + (x_u - x_{BS})^2}}, \\
\mathbf{v}_{\ell(u)}^T &= \mathbf{a}(\theta_D, \phi_D)^H, \\
\theta_A &= \theta_D + \pi, \\
\phi_A &= -\phi_D, \\
\mathbf{w}_u^T &= \mathbf{a}(\theta_A, \phi_A)^H,
\end{aligned} \tag{4.4}$$

where $\mathrm{I}_{()}$ is the indicator function, subindex D indicates the angles of departure, subindex A indicates the angles of arrival, and subindex $\ell(u)$ indicates the stream assigned to UE u.

Finally, GBF vectors are analog, so streams are matched one-to-one with array ports in this scheme. We remark that GBF is based only on the array radiation pattern and the angular position of the devices, and does not adapt to changes in the channel matrix. However, the strongest channel gain, associated with the largest singular value and singular vector of $\mathbf{H}_u[n, k]$, could be very different from the geometric direction between the locations, especially in NLOS channels.

Codebook Beamforming

This is a well-known limited-feedback technique in SU-MIMO mmWave BF literature [138]. However, we present the first evaluation of Codebook BeamForming (CBF) in full-stack simulations with SDMA support. We denote a *BF codebook* $\mathcal{B}$ as a small collection of possible BF vectors (either because quantized phase shifters are employed in analog BF or because the feedback is limited to $\log_2|\mathcal{B}|$ bits). The transmitter sends reference signals using all the vectors in $\mathcal{B}_D$, and the receiver tests decoding the reference signals with all vectors in $\mathcal{B}_A$. Finally, the receiver chooses a pair of vectors from each codebook based on the observations. For example, the receiver may choose the beams with the largest observed reference signal power. In our implementation we model this procedure as noiseless

resulting in the max-SNR criterion

$$\mathbf{v}_{\ell(u)}, \mathbf{w}_u = \arg \max_{\mathbf{v}\in\mathcal{B}_D, \mathbf{w}\in\mathcal{B}_A} |\mathbf{w}^T \mathbf{H}_u[n, k_{ref}]\mathbf{v}|^2, \tag{4.5}$$

where k_{ref} is a single subcarrier index where we assume a narrowband reference signal is transmitted. Finally, the receiver would only need to send to the transmitter a beam indicator message describing the index that $\mathbf{v}_{\ell(u)}$ occupies in the look-up table containing $\mathcal{B}_D$.

We implemented the "Discrete Fourier Transform (DFT) codebook," which for the $N_1 \times N_2$ antenna array contains $N_1 N_2$ beams pointed in the directions $\theta \in \{\arcsin(\frac{2n_1}{N_1})\}_{n_1=-\frac{N_1}{2}}^{\frac{N_1}{2}-1}$ and $\phi \in \{\arcsin(\frac{2n_2}{N_2})\}_{n_2=-\frac{N_2}{2}}^{\frac{N_2}{2}-1}$. Due to (4.3), the DFT codebook is exactly the set of columns of a DFT matrix, allowing an efficient implementation using the Fast Fourier Transform (FFT) algorithm. The CBF beams have roughly the same beam width and sidelobes as GBF beams but can only point to a finite set of angles. Despite this, the beamforming gain with CBF in Equation (4.5) is much larger than in GBF for NLOS channels, where the largest singular vector of $\mathbf{H}_u$ is unrelated to the device's physical positions used in Equation (4.4).

CBF vectors are analog and streams are matched one-to-one with array ports, as in GBF. Thanks to the use of a simple codebook-lookup technique, feedback overhead would be very low. A potential drawback is that by using a single-subcarrier reference $|\mathbf{w}_u^T \mathbf{H}_u[n, k_{ref}]\mathbf{v}_\ell|^2$ the selection does not take into account the gains that would be experienced by any other subcarrier $|\mathbf{w}_u^T \mathbf{H}_u[n, k]\mathbf{v}_\ell|^2 \; \forall k \neq k_{ref}$. This means that only the SNR of the reference is maximized while those of other subcarriers are not. Nonetheless, due to the fact that the mmWave channel matrix is rank-deficient and the beams in the codebook are sufficiently wide, the SNR in different subcarriers can be quite similar and this shortcoming is not too severe.

Frequency-Flat MMSE Beamforming

GBF and CBF focus solely on maximizing the effective channel gains between the BS and UE u, denoted as $|g[u, \ell(u), n, k]|^2$. This, in turn, maximizes user u's link SNR, but does not account for interference to and from other UEs.

Therefore, to improve the SINR in (4.1) and (4.2), we introduce a low-complexity, low-dimensional linear matrix mapping between streams and ports, in combination with an auxiliary analog CBF underlying scheme. Let us denote the BF vectors selected using CBF by $\mathbf{w}_u^{CB}$ and $\mathbf{v}_p^{CB}$, and by $g^{CB}[u,p,n,k_{ref}] = (\mathbf{w}_u^{CB})^T\mathbf{H}_u[n,k_{ref}]\mathbf{v}_p^{CB}$ the complex channel coefficient observed between user u and port p at the reference subcarrier k_{ref}.

We assume that first the system conducts a codebook exploration as in CBF and loads the best codebook BF vector for each user u to different antenna ports denoted by $p(u)$. In addition, since the standard states that the channel on each antenna port may be inferred using the different DMRSs, we assume that right after the codebook exploration the BS notifies each user of all the vectors of interest, and the receivers estimate the effective complex scalars $L_u^{-\frac{1}{2}} g^{CB}[u,p(u'),n,k_{ref}]$ for all pairs $(u,p(u'))$ at the beginning of their assigned RB allocation. To report these auxiliary effective channel coefficients back to the BS, since a single reference subcarrier is used, would require $N_u^2 N_{bit}$ bits of feedback, where N_{bit} is the number of bits used to encode each complex number and N_u is the number of simultaneous users. For example with $N_u = 4$ the feedback would be 1024 bits with 32-bit floating point encoding, or 96 bits with a 3-bit quantizer.

To simplify the notation, we assume in this section that the active users are numbered sequentially $u \in \{1 \dots N_u\}$ and that their assigned stream and port numbers are equally sequential, i.e., $\ell(u) = p(u) = u$. We also omit the OFDM symbol index n. Using the auxiliary scalar channel coefficients the BS builds the MU-MIMO reference equivalent channel matrix given in (4.7),

$$\mathbf{G}[k_{ref}] = \begin{pmatrix} L_1^{-\frac{1}{2}} g^{CB}[1,1,k_{ref}] & L_1^{-\frac{1}{2}} g^{CB}[1,2,k_{ref}] & \dots & L_1^{-\frac{1}{2}} g^{CB}[1,N_p,k_{ref}] \\ L_2^{-\frac{1}{2}} g^{CB}[2,1,k_{ref}] & L_2^{-\frac{1}{2}} g^{CB}[2,2,k_{ref}] & \dots & L_2^{-\frac{1}{2}} g^{CB}[2,N_p,k_{ref}] \\ \vdots & \vdots & \ddots & \vdots \\ L_{N_u}^{-\frac{1}{2}} g^{CB}[N_u,1,k_{ref}] & L_{N_u}^{-\frac{1}{2}} g^{CB}[N_u,2,k_{ref}] & \dots & L_{N_u}^{-\frac{1}{2}} g^{CB}[N_u,N_p,k_{ref}] \end{pmatrix} \tag{4.7}$$

where $N_p = N_u$ is the number of analog BF ports, each associated to a single user. Moreover, since $\ell(u) = p(u) = u$, the desired channels are in the main diagonal of this matrix. Finally, on the receiver side the BF vectors remain those of CBF, while on the transmitter side the BS designs the following MMSE DL precoding

matrix matching streams to antenna ports

$$\mathbf{V}_{MMSE}[k_{ref}] = \mathbf{G}^H \left(\mathbf{G}\mathbf{G}^H + \frac{N_o \Delta f K}{P} \mathbf{I} \right)^{-1}. \tag{4.8}$$

Zero-Forcing (ZF) and MMSE precoding are well-known in physical layer MU-MIMO literature[134, 135]. However, like CBF, we present their first full-stack evaluation and a detailed discussion of their interplay with upper protocol layers. We adopt the MMSE technique rather than ZF because, when the noise is weak compared to the transmitted power, then $\frac{N_o \Delta f K}{P} \to 0$, and (4.8) converges to the ZF precoder, i.e., $\mathbf{G}[k_{ref}]\mathbf{V}_{MMSE}[k_{ref}] = \mathbf{I}$, suppressing the interference. In addition, when the noise is strong, i.e., in the limit as $\frac{N_o \Delta f K}{P} \to \infty$, (4.8) converges to the Hermitian (matched filter) which maximizes the SNR. Thus, MMSE balances interference and noise reduction, giving good SINR values for any noise-to-transmitted power ratio ($\frac{N_o \Delta f K}{P}$).

Finally, the *effective* transmit BF vectors at the BS for DL are obtained by first computing

$$\left(\tilde{\mathbf{v}}_1^{MMSE} \dots \tilde{\mathbf{v}}_{N_u}^{MMSE} \right) = \left(\mathbf{v}_1^{CB} \dots \mathbf{v}_{N_u}^{CB} \right) \mathbf{V}_{MMSE}[k_{ref}],$$

and then performing the following transmitted power constraint normalization in each stream:

$$\mathbf{v}_u^{MMSE} = \tilde{\mathbf{v}}_u^{MMSE} / |\tilde{\mathbf{v}}_u^{MMSE}|.$$

Introducing these effective vectors into Equation (4.1) returns the SINR values of the MMSE technique.

For UL, an equivalent hybrid combining at the BS receiver can be formulated by adopting the transpose of the matrices described in this section.

The Frequency-Flat MMSE BeamForming (FMBF) technique relies solely on estimations performed in the reference subcarrier k_{ref} and is still frequency-flat. This introduces only a small amount of additional feedback as only one subcarrier equivalent matrix needs to be reported. The precoding/combining matrix is explicitly designed to improve the SINR in the reference subcarrier and, as a side effect of the mmWave channel sparsity, the SINR can improve also in other subcarriers. However, since FMBF does not take into account the effective channel of the other subcarriers, it does not guarantee complete interference suppression

in all subcarriers. Due to the fact that in NR DMRS the channel measurements must remain unchanged in all OFDM symbols in the same slot and with the same user allocation, we note that the beamforming vectors cannot be changed in the middle of an allocation, and MMSE cannot be used for transmissions that start at different times and become simultaneous later. We will further address this issue on the scheduler design section. Even though MMSE techniques are well-known, to the best of our knowledge this work is the first to report this issue.

Frequency-Selective MMSE Beamforming

The principle of combining different digital precoders in each subcarrier with frequency flat analog beams is well established in mmWave physical layer literature [137]. However, as CBF and FMBF, to the best of our knowledge this work is the first to assess this technique with a full-stack network evaluation.

In DL, we need to assume that after beam codebook exploration is performed, pilot signals are transmitted in all subcarriers and the receivers can report back to the transmitter a large set of effective channel coefficients $\{g^{CB}[u, p(u'), n, k]$ for all pairs $(u, p(u'))$ and subcarrier indices $k\}$. This would require roughly an increase of the feedback by a factor of K compared to FMBF. For example if $K = 100$, $N_u = 4$ and with a coarse quantizer with $N_{bit} = 3$, we could send the resulting 9.6 kbits of feedback in a single OFDM symbol, whereas using high precision complex number encoding with $N_{bit} = 32$ would require more feedback than we can fit in an NR slot. Using the effective channel information the transmitter builds a collection of K different equivalent channel matrices, one for each subcarrier ($\mathbf{G}[k] \forall k \in \{1 \dots K\}$). For each subcarrier k the transmitter designs a different digital precoding matrix

$$\mathbf{V}_{MMSE}[k] = \mathbf{G}^H[k] \left(\mathbf{G}[k]\mathbf{G}[k]^H + \frac{N_o \Delta f K}{P} \mathbf{I} \right)^{-1}.$$

Thus in the Frequency-Selective MMSE BeamForming (SMBF) scheme the precoding matrix that maps antenna ports to streams is different in each subcarrier. Finally, normalization and calculation of effective BF vectors proceeds as in the FMBF case, but with the effective vectors introduced in Equation (4.1) taking different values for each subcarrier index k. For UL, the same considerations

regarding the transpose matrix apply.

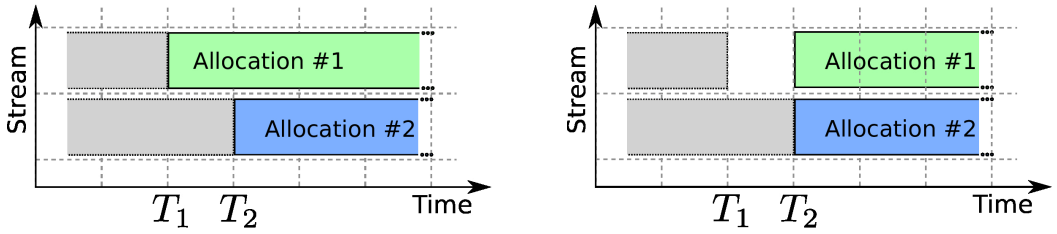
In return for the significant increase in feedback, SMBF guarantees an explicit suppression of the inter-user side lobe interference in all subcarriers.

Future Work

The insights we report are general and related to the interactions between layers rather than to the fine details of one physical layer. Once the need arises, our framework is easily expandable to other existing mmWave physical layer techniques that have not been previously studied from an end-to-end perspective. We leave for future work the MU-MIMO techniques that go beyond linear BF and independent encoding and decoding in each link. In recent years, Non Orthogonal Multiple Access (NOMA) [173] has gained attention in the literature. As mentioned in Section 4.3.1, in UL the BS could decode all incoming transmissions jointly using Sphere Decoding [172] or Successive Interference Cancellation. Conversely, in DL, the BS could jointly encode all transmitted signals using Dirty Paper Coding [174]. In addition, we have assumed fixed equal power allocation in all subcarriers in Equation (4.1) and Equation (4.2), leaving optimal unequal power allocation for future work. We also leave for future work the recent "cell free" technique in which multiple BSs perform joint MU-MIMO transmissions [175]. Moreover, we also leave for future work the study of BF performance degradation due to noise in the beam measurements or in channel estimation, and to quantization in the channel coefficient feedback.

4.3.2 HBF and Scheduling Interaction

We assume that the scheduler allocates transmissions in a 2-dimensional resource grid combining TDMA and SDMA. All subcarriers in the same OFDM symbol are allocated to the same UE due to the fact that the BF system in mmWave is at least partially frequency-flat. The scheduler produces allocation decisions periodically for each slot of 14 symbols. The standard supports flexible configurations for allocating control information, i.e., the Physical Downlink Control Channel (PDCCH), in specific regions of each frame [130]. We assume a periodical control signaling scheme where, for every 14-symbol slot, the first symbol always contains the PDCCH. In the PDCCH, DCI control messages are delivered to all users. In



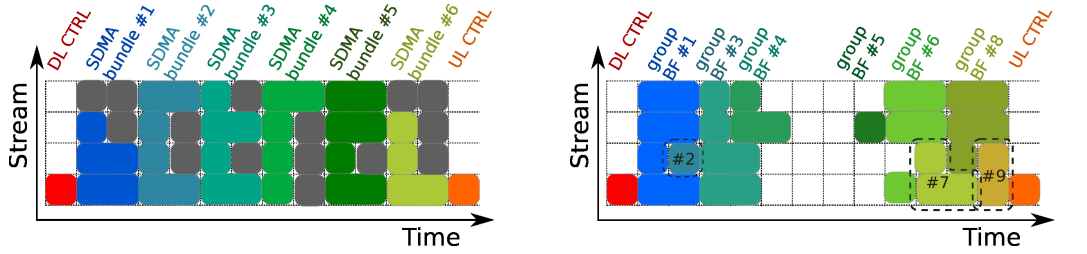
(a) BF conflict with different transmission start times: At T_1 allocation $\#1$ does not observe reference signals from other streams. The BS cannot estimate Equation (4.7) and use MMSE BF, falling back to CBF. When Allocation $\#2$ starts at T_2, the transmission of reference signals by Allocation $\#1$ has already passed, so Allocation $\#2$ must fall back to CBF as well. In addition, Allocation $\#1$ cannot change its BF configuration at T_2 either. After T_2, both allocations experience the interference power of a CBF scenario even though MMSE is supported by all devices.

(b) BF conflict with forced simultaneous transmission start: The transmission in the top stream ends at T_1, but the scheduler leaves a padding symbol without signal and the new allocation starts at T_2. When transmission in the bottom stream ends at T_2, both streams start a new allocation simultaneously. Both allocations can observe each other's reference signals and estimate the off-diagonal coefficients of Equation (4.7). MMSE BF can be employed and interference is reduced. However, the frame resource region corresponding to the time interval $T_2 - T_1$ in the top stream is wasted.

Fig. 4.1: Example of MMSE BF conflict with different transmission start times.

this initial study we assume that control signaling is ideal, control messages are never lost or corrupted, the best beams are always successfully identified without error, and the channel coefficients estimated using DMRSs are noiseless observations of the channel gains. We leave the extension of the results to imperfect control protocols and channel estimation for future work. Symbols 2 to 13 are used for data and marked as "flexible," meaning that they can be employed for DL or UL in any slot and this choice may vary over different slots. Finally, in the 14-th symbol of each slot the UEs transmit UL control information to the BS.

As we assume noiseless channel estimation, we do not model the DMRSs explicitly. Since the smallest scheduling unit is a 2-symbol mini-slot with 1 DMRS symbol [16], in our model we assume that the minimum data allocation unit is reduced to 1 symbol of data transmission. We assume that allocated transmissions on different streams may present different start times (in symbol index units). Since each allocated transmission has only one front-loaded DMRS, the BF configuration of each transmission must be selected at the start of the transmission and cannot vary over the duration of the same transmission. This means that, for a pair of overlapping transmissions that start at different instants, the trans-



(a) PMRS: The resources in gray are padding, to guarantee that simultaneous allocations always start at the same time.

(b) AMRS: The resources in white are available, i.e., if more users were added, these resources could be assigned to them.

Fig. 4.2: Examples of scheduler slot decisions with our two proposals

mission that started first does not have information on the interference to design its BF (Figure 4.1a). Therefore we assume that MMSE precoding/combining can only be applied to groups of allocations that start at the same time (Figure 4.1b). Thus, even though the MMSE technique is well-known in the literature, our work identifies a novel conflict between scheduling constraints and the applicability of the MMSE technique, as illustrated in Figure 4.1. We consider two approaches illustrated in Figure 4.2: the first can make full use of MMSE but imposes additional constraints on the scheduler leading to lower resource efficiency (Figure 4.2a). In the second, we allow the scheduler to freely allocate resources even with different start times among different transmissions, achieving a more efficient frame resource occupation (Figure 4.2b), but causing a fraction of the transmission events to be unable to use MMSE for interference reduction. We call this a "fallback beamforming" event, in which the physical layer supports MMSE but the scheduler causes the use of plain CBF instead.

Padded mmWave Round Robin Scheduler

This scheduler guarantees that possibly overlapping transmissions start at the same time in all streams. To do so, given N_ℓ streams, N_s symbols and N_u total UEs, the scheduler first divides the subframe equally in $N_b = \lceil N_u/N_\ell \rceil$ "SDMA bundles." Each SDMA bundle is defined as a collection of up to N_ℓ concurrent transmissions with the same start time, but allocated to different streams (Figure 4.2a). The bundles are further time-multiplexed using TDMA over the full subframe, where each bundle is exactly $N_a = \lfloor N_s/N_b \rfloor$ symbols long in time. All

streams in the bundle start being transmitted at the same time but may end at different times, according to the amount of data each UE has to transmit. Indeed, within each bundle, and for each stream, one UE is selected. If this UE demands fewer than N_a symbols, then its transmission ends before the end of the bundle, and the remaining symbols are left blank (padding). If $N_u > N_s \times N_\ell$, then some UEs are left unserved and become the first UEs in the list for the next subframe in an RR fashion.

Padded mmWave RR Scheduler (PMRS) follows a TDMA first and SDMA second principle, guaranteeing equal start times for all transmissions in a bundle. This guarantees that MMSE BF is always usable and interference will fully depend on the chosen BF scheme. The padding part in each bundle constitutes wasted symbols, and thus this scheduler may display some inefficiency in resource occupation.

Asynchronous mmWave almost-Round Robin Scheduler

This scheduler, instead, does not waste any symbol in padding: given N_ℓ streams, N_s symbols and N_u total UEs, the scheduler first divides the users into the N_ℓ streams. Thus the streams each serve a different subset of UEs, partitioning the set of all users. The UEs in each single stream are further time-multiplexed using the same default TDMA mmWave RR Scheduler (TMRS) method that was proposed in previous versions of the mmWave ns-3 module [92], without taking into account decisions for other streams. If all UEs demand more resources than available, then each UE receives $\lfloor N_s N_\ell / N_u \rfloor$ symbols. However, UEs that demand fewer resources receive fewer symbols, their allocations end sooner, and the next UE in the same stream begins its allocation immediately after, without any padding. Due to this, the start times of the transmissions in one stream are determined independently of the start times of other streams. After the frame is populated, the physical layer forms "beamforming groups" composed of transmissions that start at the same instant, as shown in Figure 4.2b. Each beamforming group may use MMSE to reduce side lobe interference, however, simultaneous transmissions with different start times such as BF groups #1 and #2 in the figure, cannot design their beams jointly and may interfere significantly with each other during their overlap. The UEs are divided among the streams

using an integer division that assigns $\lceil N_u/N_\ell \rceil$ users to each stream, and as a result the number of UEs per stream may differ by one unit. Finally, UEs that cannot get any resources will be served first in the next slot. Since allocations may have different sizes we call the scheduler "almost-RR."

The Asynchronous mmWave almost-RR Scheduler (AMRS) thus follows an SDMA first and TDMA second principle, guaranteeing that no symbols are wasted. Free symbols may exist when the total demand of all UEs is lower than $N_s N_\ell$, but these symbols are not "wasted as padding" and would be allocated if there were more demand. However, equal start times for all concurrent transmissions are not guaranteed. This means that MMSE BF is not fully used, there are fallback beamforming events, and interference will display a random mixture of the CBF and MMSE physical layer behaviors.

4.4 Performance Evaluation and Tradeoffs

We implemented an MU-MIMO HBF extension for the ns-3 mmWave module introduced in [46]. Earlier releases of ns-3 introduced the ESM physical layer model for LTE [176]. The NYU mmWave single-stream physical layer and channel model was introduced in [177]. We adopted this well established physical layer implementation as a base and modified it to support multiple simultaneous transmitted or received signals in any device and antenna array. Besides the bulk of the multi-stream implementation, we have introduced adjustments that bring our framework closer to the 3GPP 5G NR standards. Instead of the NYU channel model, we adopt the most recent 3GPP channel model implementation in ns-3 presented in Chapter 2. In addition, the OFDM resource grid parameters (bandwidth, subcarrier spacing, symbol duration, and number of slots per frame) reflect those of NR, as described in Section 4.3 and [16].

Our extension of the ns-3 mmWave module adds support for multiple antenna ports with different BF in the antenna arrays. Moreover, the 3GPP channel module has been extended to compute inter-stream side lobe interference according to Equation (4.1) and Equation (4.2). The channel, physical, and RLC implementations have been extended to support multiple SDMA asynchronous streams. The BF strategies described in Section 4.3.1 have been implemented in a plug-and-play, novel and flexible BF API and with separate objects for each BF method.

Finally, we updated the MAC layer to support multiple asynchronous streams, each with different mapping of upper layer PDUs to NR Transport Blocks, Hybrid Automatic Repeat reQuest (HARQ) retransmission process, CQI estimation, and control messages. The scheduler API permits the plug-and-play change of the scheduler module. We implemented the schemes in Section 4.3.2, and maintained backward compatibility. This allows comparison with the prior single-stream scheduling strategies in earlier versions of the ns-3 mmWave module [46]. The publicly available Github repository with the HBF extension can be consulted for additional details[†].

4.4.1 Simulation Scenario

We present different performance results pertaining to different aspects of the SDMA MU-MIMO mmWave system. For all results below, we simulated a random mmWave cellular system with one BS located at the origin of the coordinates (0,0) with a height of 25 m, and 7 UEs located at random positions uniformly distributed in a disc of radius 100 m with a height of 1.6 m. We generate 20 such random deployments and average the results over the random UE locations and channels. We assume that due to the considerable pathloss in mmWave, inter-cell interference is severely attenuated and, therefore, it is sufficient to model only one cell. This is different from prior work on 4G systems, where also a set of "encircling" neighboring cells had to be considered to model interference realistically [178].

We configured the NR OFDM waveform with numerology $\mu = 2$, which corresponds to a subcarrier spacing of 60 kHz. The system operates at 28 GHz central frequency with a bandwidth of 198 MHz divided into 275 RBs, each including 12 subcarriers. There are 4 slots per subframe with duration 250 μs, and the OFDM symbol duration is 17.85 μs including the CPs. We adopt the channel model described in 3GPP TR 38.901 [59, 123] and consider the "Urban Macro" scenario. The radio hardware configuration and other simulation parameters are summarized in Table 4.1.

[†]https://github.com/signetlabdei/ns3-mmwave-hbf

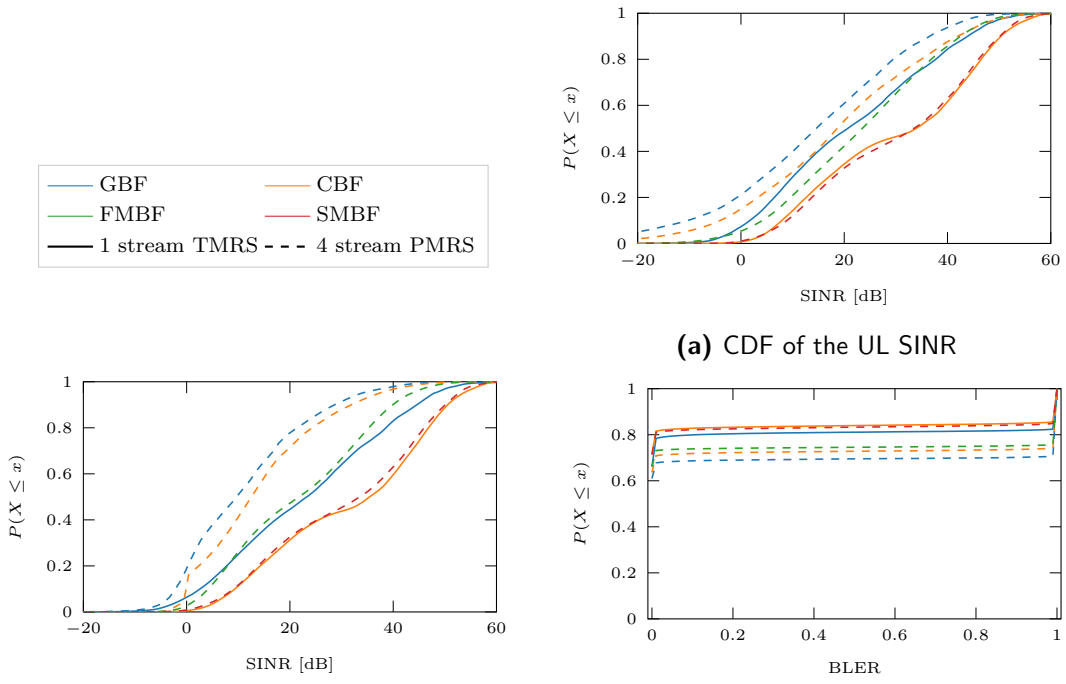
Table 4.1: Simulation parameters

Parameter	Value
Bandwidth B	198 MHz
Frequency f	28 GHz
Numerology μ	2
Subcarrier spacing	60 kHz
Slots per subframe	4
Subframe duration	250 μs
OFDM symbol duration	17.85 μs
Channel scenario	Urban Macro
UE transmit power	30 dbm
UE noise figure	5 dB
UE number of streams	1
UE antenna array configuration	4×4 UPA
BS transmit power	30 dBm
BS noise figure	5 dB
BS number of streams	1 or 4
BS antenna array configuration	8×8 UPA

4.4.2 Comparison of Beamforming Solutions

We compare the BF schemes discussed in Section 4.3.1. To clearly highlight their impact on the physical layer, we use RLC-Unacknowledged Mode (UM) (i.e., without RLC retransmissions), disable the HARQ retransmissions at the MAC layer, and use a low-traffic application in the UEs. This minimizes the difference between the statistics of the SINR and BLER measured at the upper layers and the random distribution that generates these values at the channel model.

The low-rate traffic generator produces a downlink and an uplink packet of 1500 bytes every 1500 μs, for each UE. Roughly speaking, when the MCS coding rate is greater than 3.64 bits per subcarrier, the 3300 subcarriers can carry a full packet in a single OFDM symbol. This means that, for every six slots of 250 μs each, the scheduler receives a demand for at least $\sim$ 14 symbols in the first



(a) CDF of the UL SINR

(b) CDF of the DL SINR

(c) CDF of the UL BLER

Fig. 4.3: Comparison of the different BF schemes.

slot (i.e., each of the 7 UEs requests one downlink and one uplink symbol), and none in the subsequent 5 slots. The scheduler has thus plenty of RBs to satisfy the traffic demand, and, as discussed, there are no retransmissions. This traffic probes the cell regularly and allows to measure the physical layer statistics due to channel conditions and BF schemes by registering the effective SINR of each transmission.

Figure 4.3a represents the received UL SINR CDF for all transmissions in the simulation. We compare 1 stream (solid) and 4 stream (dashed) cases. For the 1 stream case, we use the TMRS without SDMA capabilities of the previous versions of the ns-3 mmWave module, with either GBF or CBF (MMSE BF has no effect in the single stream case, behaving exactly as CBF). For the 4 stream case we consider the PMRS, so that all allocations use the specified BF scheme (i.e., the MMSE schemes never fall back to CBF, as discussed in Section 4.3.1). We compare 4-stream GBF, CBF, FMBF, SMBF. In the 1-stream case there is no inter-beam-interference and the SINR is the same as the SNR, which is better with CBF than with GBF, consistently with our discussion in Section 4.3.1. On

the other hand, for 4 streams using single-stream BF techniques (GBF or CBF), the SINR drops significantly, with frequent -20 dB events. This confirms that the single-stream BF schemes do not perform well and *the use of multi-stream specific BF is necessary.* Adopting the FMBF scheme improves the SINR by a significant margin, but does not fully compensate the interference. As designed, the SMBF scheme does remove almost all interference, and its SINR CDF is nearly identical to that of the single-stream CBF case.

Figure 4.3b represents the received DL effective SINR CDF. The main difference with UL is that in DL -20 dB SINR outages rarely happen with 4-stream CBF. On the other hand, the gap between 4-stream CBF and 4-stream MMSE BF schemes is wider than in UL. Again the SMBF scheme removes virtually all interference and achieves an SINR distribution akin to that of a 1-stream CBF scenario, whereas the FMBF scheme achieves an intermediate SINR improvement.

Finally we depict the instantaneous BLER CDF for all UL transmissions in Figure 4.3c. Generally speaking, the BLER distribution is almost a step function: In each transmission, the CQI feedback is used to select the MCS such that the BLER would be 10^{-2} *if the reported channel stayed the same.* Therefore, we can define a "CQI outage" as the event that, at the moment of transmission, the channel has become much worse compared to when the CQI was reported, and the instantaneous BLER is $\simeq 1$. Figure 4.3c shows that the instantaneous BLER is dominated by such outages, where most transmissions experience either BLER $\leq 10^{-2}$ or BLER $= 1$. The complement of this outage probability corresponds to the height of the flat region of the CDF curves. As we can see, 4-stream GBF and CBF have a much larger outage probability (lower step in the BLER CDF) and result in more severe BLER. Again, we see that SMBF is almost equal to the 1-stream CBF case, and FMBF is between these two cases. We do not depict the DL BLER CDF due to space constraints, as its insights were identical.

In summary, the BF comparison shows that SMBF is necessary in the MU-MIMO 4-stream implementation to ensure that the physical layer achieves the same SINR and BLER as in the 1-stream mmWave system with CBF. Moreover, the main differences between all BF schemes we considered are that GBF performs worse than CBF for any number of streams and that FMBF offers some performance improvement without as much overhead as SMBF.

4.4.3 Cross-layer Beamforming and Scheduling Interactions

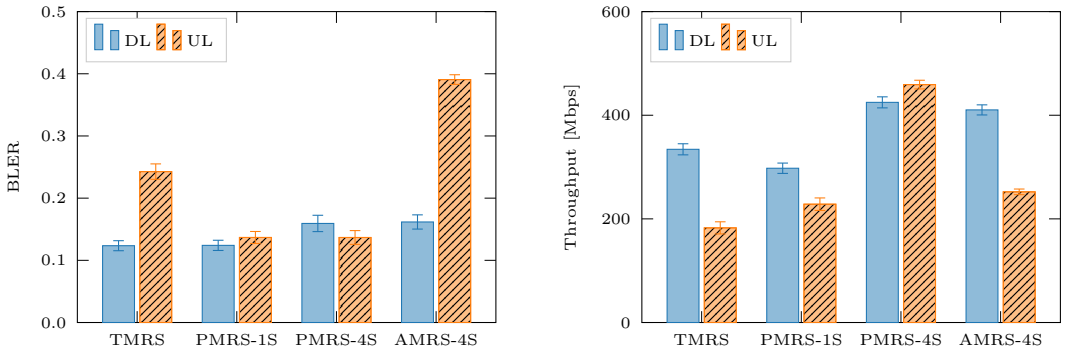
Next, we compare the scheduling algorithms introduced in Section 4.3.2. To highlight the scheduler's behavior with respect to the offered traffic, once again we use RLC-UM and disable the HARQ retransmissions. However, differently from the previous section, we use high-traffic applications in the UEs to emphasize the effect of the scheduler on the system performance.

To adopt the best BF scheme for each number of streams, we consider four scenarios: the TMRS 1-stream scheduler with CBF, our proposed PMRS for both the 1-stream CBF and 4-stream SMBF configurations, and our proposed AMRS for 4 streams with SMBF. PMRS is designed for use with multiple streams, but since it forces all allocations to be of the same size, its behavior when applied to the 1-stream case differs slightly from that of TMRS. Due to this, we include an observation of PMRS in the 1-stream case. AMRS, on the other hand, behaves exactly like TMRS if invoked on a 1-stream frame. AMRS may assign allocations so that two overlapping transmissions do not start at the same time, and in these fallback beamforming events the SMBF scheme behaves like the CBF scheme (see Figure 4.1).

The high-rate traffic generator produces a packet of 1500 bytes every 150 μs in each UE, for both uplink and downlink. For every slot of 250 μs, the scheduler receives a request for at least ~ 23 symbols (more if MCS rate < 3.64 bit/symbol). With 1-stream each slot has 12 data symbols, which are not enough to serve all the demand. In the 4-stream case, there are 12×4 available data symbols per slot, i.e., enough resources if MCS rate > 1.82 bits/symbol.

Figure 4.4a reports the average DL and UL BLER for the four scheduler-BF pairs discussed above. As can be seen, AMRS displays a high UL BLER because it does not fully take advantage of the SMBF technique. The problem is more severe in UL because, as discussed in the previous section, the pathloss leads to more severe SINR drops (outages) in this direction than in DL. The BLER of PMRS with 4 streams is comparable to that of TMRS and PMRS with 1 stream, which is also consistent with the SINR plots discussed in the previous section.

Figure 4.4b depicts the throughput, defined as total data received divided by total simulation duration. Since the nominal application rate is $\frac{7\times1500\times8}{150\times10^{-6}}$ bit/s, the maximum throughput is 560 Mbps. For TMRS, throughput is around 330 Mbps

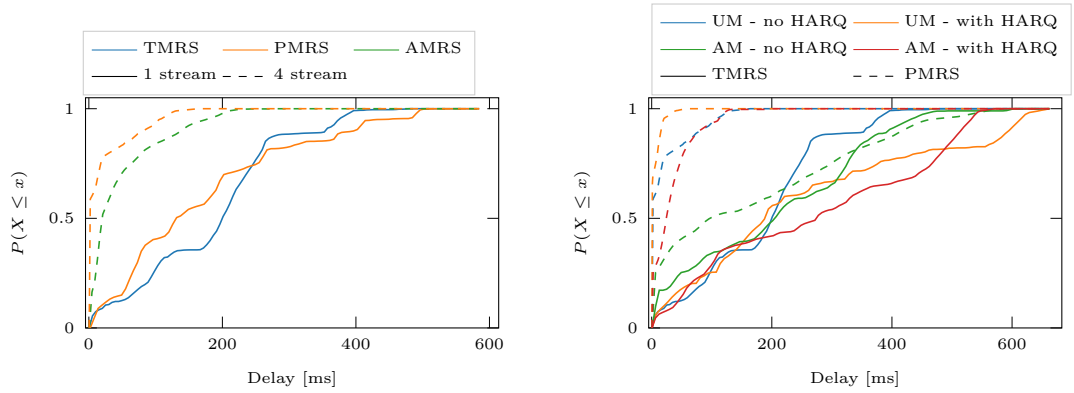


(a) Average BLER **(b)** Throughput

Fig. 4.4: Comparison of the different scheduling strategies.

in DL and 180 Mbps in UL, with significant asymmetry and much lower value than the offered traffic. This is consistent with the fact that the offered traffic greatly exceeds the number of data symbols of the 1-stream frame even with the best MCS. The same holds for 1-stream PMRS with CBF, although its DL/UL traffic is better balanced. With 4 streams, the resources are not saturated, and the throughput with PMRS exceeds 420 Mbps in DL and 450 Mbps in UL. This shows the main advantage of SDMA, i.e., *an increase in the number of available RBs by a factor of N_ℓ allows the network to support much more traffic.* Particularly, our results show 2× more delivered traffic than in the single-stream case. As for AMRS, we see that it also supported over 410 Mbps of DL traffic successfully, but it only delivered around 250 Mbps of UL traffic. This is consistent with the high UL BLER due to SINR outages, because AMRS does not ensure that all streams start their transmissions at the same time, causing fallback beamforming events in SMBF.

This section shows that the interaction between the scheduling and beamforming schemes can be highly complex. While the previous section established that CBF suffers occasional severe outages in a multi-stream setting, this section shows that this can occur in SMBF when using schedulers that do not take into account the properties of the MU-MIMO BF algorithms.



(a) Comparison of Schedulers **(b)** Comparison of Retransmission Schemes

Fig. 4.5: Delay in UL

4.4.4 Delay and Retransmissions

In this section we look at other performance indicators besides capacity. To the best of our knowledge, our model presents the first end-to-end framework capable of studying the relation between SDMA, NR retransmission schemes and delay in MU-MIMO mmWave networks. Delay measurements are taken at the PDCP layer in the 3GPP stack, so undelivered packets do not contribute to delay statistics. On the other hand, retransmission queuing adds to the delay in reliable modes. Thus unreliable transmission modes tend to display shorter delay statistics but with fewer packets successfully arriving. The RLC retransmission mode (i.e., RLC AM) provides reliability on a much larger time scale than the MAC HARQ mechanism. The RLC "reordering timeout" is 10 ms, whereas the HARQ scheme retransmits immediately after a Negative Acknowledgment (NACK) is received. Since UL control information is processed at the end of every slot, the HARQ retransmission time is about 250 μs. For this reason we expect RLC-AM to dominate the increase in delay caused by retransmissions.

Figure 4.5a compares the delay CDFs for different schedulers under the no retransmission configuration (RLC UM without HARQ), with the high traffic UDP application presented in the previous section. We use the CDF instead of bar plots to capture the inverted-L shape of the delay CDFs when most traffic is successfully delivered in the 4-stream configuration. With this, more than 75% of the packets are received in under 20 ms, and more than 90% of the packets

are received under 100 ms. In a deadline constrained application, such as for example video with one frame every 20 ms, this means that 75% of the frames would be received on time and displayed on screen (with no buffering). As for the differences between AMRS and PMRS, the latter indeed guarantees a 10 ms deadline with probability 80% and a 100 ms deadline with probability 95%, which is better than AMRS. The 1-stream schemes, both TMRS and 1-stream PMRS, do not display an inverse-L shaped CDF because the network capacity is exceeded by the applications. Instead, the delay CDF with 1 stream is roughly linear as many packets accumulated long times waiting in queues.

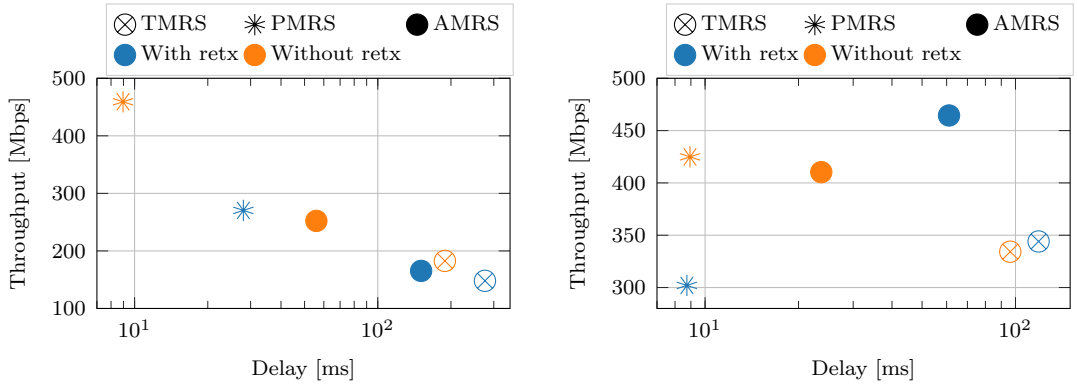
Figure 4.5b displays the delay CDF using all four possible retransmission configurations for TMRS and 4-stream PMRS. Since the offered throughput exceeds the resources of the 1-stream frame, the delays with TMRS present again an almost-linear slope dominated by the queue waiting time. On the other hand the PMRS cases present two-slope inverse-L shapes that are mostly driven by outages and retransmissions. The lowest delay 80%-tile is achieved by the RLC-UM with HARQ PMRS configuration, followed closely by the RLC-UM without HARQ PMRS configuration. This suggests that HARQ retransmissions help improve delay, which suggests that their contribution to improve reliability compensates the small delays incurred by HARQ retransmissions. The CDFs for PMRS with RLC AM with and without HARQ retransmissions exhibit a very dissimilar behavior. Without HARQ, multiple AM retransmissions are needed, where each retransmission adds over 10 ms to the packet delivery delay. On the contrary, with HARQ most retransmissions take place at the MAC layer, with a short round-trip time, and RLC only needs to compensate for occasional HARQ failures. It is noteworthy that the delay CDF for the RLC AM without HARQ padding configuration looks similar to the 1-stream curves, which suggests that RLC retransmission queues are growing without bounds in this scenario. Regarding the differences in behavior between different retransmission configurations for TMRS, it seems that resource occupation dominates the delay since the 1-stream frame capacity is exceeded. That is to say, the RLC UM without HARQ 1-stream configuration does not add any resource demands besides that of the applications, thus alleviating the queues, whereas the RLC AM with HARQ configuration adds resource demands to the scheduler on top of the demands already presented by the fresh packets, making the queues and the delay grow even longer.

The main conclusion of this subsection is that, since the delay is strongly related to resource availability and queuing, the use of SDMA greatly increases the number of available resource blocks, permitting the schedulers to support larger traffic demands with low delay. Among the schedulers, PMRS offers an improved delay profile with respect to AMRS, but both are able to offer under 20 ms delay to a high percentage of the traffic. For intuitive reference, a video at 50 frames per second displays one frame every 20 ms, so this result is of the same order of magnitude as real-time multimedia applications. To introduce reliability, HARQ should be given preference before considering the use of the RLC AM mode, as following the opposite order would cause too many retransmissions and delay at the RLC level.

4.4.5 Throughput vs Delay

This subsection further extends the scheduler comparison by considering joint throughput and delay results using the "full retransmission" scheme, i.e., the RLC AM mode with HARQ, versus the scenario "without retransmissions" consisting in using the RLC UM without HARQ retransmissions. As in the previous section, we consider a high-rate UDP application with 150 μs inter-packet intervals and focus on the delay and throughput. We compare the default TMRS in the ns-3 mmWave module with 1 stream versus our PMRS and AMRS with 4 streams.

Figure 4.6a reports the mean UL delay vs throughput. Each point in the scatter plot corresponds to one possible system configuration, with the best result corresponding to the top left corner, i.e., the highest throughput with the lowest delay. Recalling that the *offered* traffic is $7 \times 1500 \times 8/150 \times 10^{-6} = 560$ Mbps, we note that the 4-stream padding scheduler without retransmissions is able to deliver almost all the traffic. Surprisingly, activating the RLC AM mode with HARQ reduces the throughput, which means that the additional RB demand of the retransmissions overweighs the benefit of increased reliability. Since the offered traffic greatly exceeds the capacity of the 1-stream case, TMRS with 1 stream displays large delays (waiting in queues) and low throughput. This figure focuses on the UL performance, in which AMRS suffers occasional fallback beamforming events, and hence its throughput and delay are worse than with PMRS.



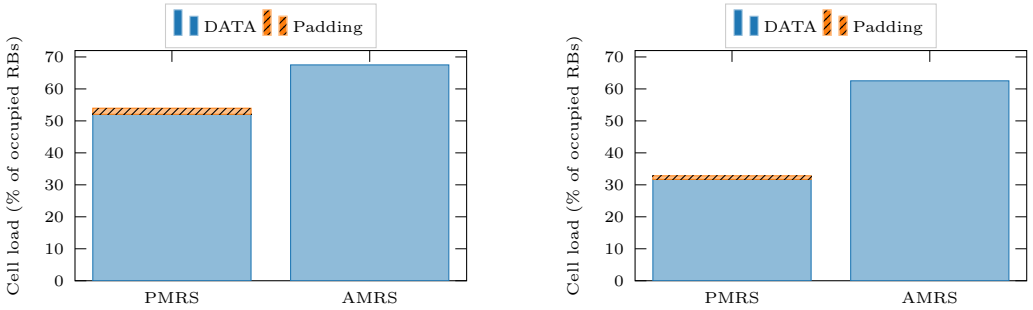
(a) Uplink **(b)** Downlink

Fig. 4.6: Comparison of the different scheduling strategies.

Figure 4.6b reports the same metrics for DL. The major difference with the UL case is that now AMRS performs much better. Indeed, with retransmissions enabled, AMRS displays the highest throughput at the cost of a slightly higher delay (due to the RLC AM retransmission timer). On the other hand, PMRS displays a significant drop in throughput with retransmissions versus the case without them. This suggests that enabling retransmissions may be counter-productive in PMRS. AMRS, on the other hand, takes advantage of the retransmissions to compensate for the occasional drop in SINR due to fallback beamforming. This shows that the interplay between beamforming, scheduling and end-to-end traffic performance is very complex; and that the apparently "worse" beamforming of SMBF+AMRS in the physical layer can potentially be offset by the behavior of the upper layers of the protocol stack.

Figure 4.7 depicts the average cell load, defined as the percentage of the total number of frame RBs that are occupied by transmitted data symbols. For PMRS, we depict the padding symbols as well, where the fraction of remaining useful free symbols would be 100% minus the sum of the cell load and padding. The figures show that the effect of padding is not too severe, and that even with padding PMRS consumed fewer frame resources than AMRS.

The results in this section highlight the importance of a full-stack, end-to-end performance evaluation. Indeed, the evaluation of the BLER and SINR in Section 4.4.3 seemed to suggest that AMRS always performed no better than PMRS. However, in DL, the BLER penalty of AMRS can be compensated using



(a) Without retx. **(b)** With retx.

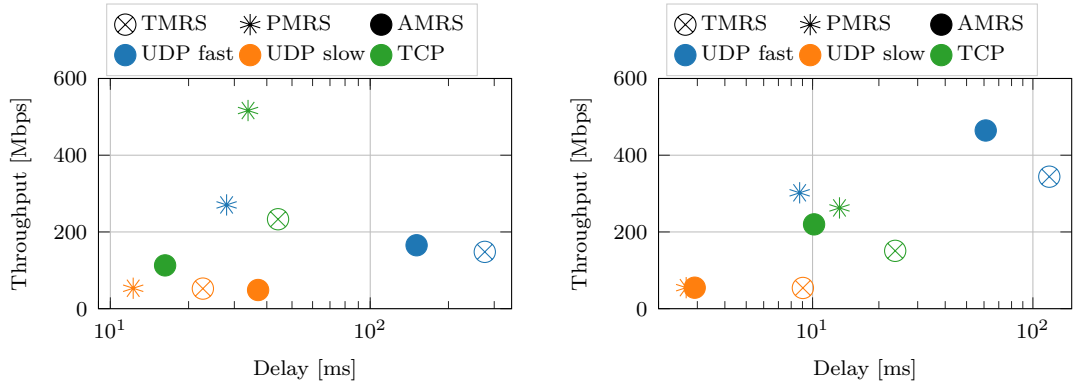
Fig. 4.7: Average cell load (percentage of the total RBs per frame that are occupied)

retransmissions, and, overall, AMRS yields an improved throughput compared to PMRS. Nonetheless, in UL AMRS severely underperforms PMRS. Moreover, the padding overhead in PMRS is tolerable. This suggests that different scheduling principles could be adopted for the two directions.

4.4.6 Performance with Different Traffic Sources

Finally, we compare the system performance under three different applications and transport stream configurations, and investigate the relation between the application traffic and the scheduler. The first two applications are those considered in Section 4.4.2 and Section 4.4.3, i.e., a constant bitrate source that generates a packet of 1500 bytes every 1500 or 150 μs, respectively. In this case, the transport layer is UDP (thus they will be referred to as *UDP slow* and *UDP fast*, respectively). Finally, we also profile the performance with a full buffer application that relies on TCP at the transport layer, to adjust the offered traffic to the maximum supported by the network. We consider retransmissions in the RLC and MAC layers to obtain similar reliability in the applications over UDP as in the application over TCP.

Figure 4.8a represents the UL delay vs throughput for all three applications and all three scheduling solutions. Since in the UDP slow application (in yellow) the offered traffic is much lower than the potential cell capacity, almost all source rate is successfully delivered by all schedulers (about 56 Mbps). In addition, PMRS displays the lowest delay, followed by TMRS, with AMRS offering the worst UL delay. As discussed throughout the prior sections this is because of the



(a) Uplink **(b)** Downlink

Fig. 4.8: Comparison of different applications.

retransmissions caused by the occasional events when AMRS suffers deep SINR outages in UL. In the UDP fast application (in blue) the traffic sources offer 10× more throughput, which is almost fully delivered using PMRS. TMRS and AMRS do not deliver all the UL traffic for different reasons. While in TMRS this is due to the limited resources of the 1-stream frame, in AMRS the reason is the high BLER due to the occasional outages. Since the UDP fast application does not adjust its transmission queues, the delay in these two schedulers increases significantly. Finally, for the TCP application (in green), PMRS offers the best performance achieving about 560 Mbps. TMRS has limited resources in the 1-stream frame and hence the throughput is less than half, but the delay is tolerable under 50 ms. Finally, AMRS achieves a very low rate, which can be explained by the TCP rate control responding too strongly to the occasional SINR outages, which produce packet losses that trigger the TCP congestion control, reducing the transmission window.

Similarly, Figure 4.8b represents the DL delay vs throughput. As in the previous figures, the main difference is that AMRS performs much better in DL than in UL. For the UDP slow application we still see that all the traffic is delivered, but the source rate is small. In DL the AMRS delay is much lower and similar between the two 4-stream schedulers (under 3 ms), whereas the delay of TMRS is a bit higher but still under 10 ms. For the UDP fast application, AMRS turns out to be the best in terms of total DL throughput, albeit with considerable more delay than PMRS. TMRS displays high delay and limited throughput due

to the lack of resources of the 1-stream frame. The throughput with PMRS is about half as much as with AMRS, but with much lower delay. Notably, the throughput-delay behavior of PMRS with UDP fast is similar to that of the TCP application with either 4-stream scheduler. As the TCP rate adaptation reduces the transmission window when certain timers expire, its delay is under 30 ms for all schedulers, but the achieved rate with such delay varies. PMRS offers the best TCP throughput with under 15 ms delay, followed closely by the AMRS scheduler. Finally, TMRS achieves the worst TCP throughput and the highest delay, due to the limited resources of the 1-stream frame.

The main conclusion of this section is that the MU-MIMO system performance depends significantly on the offered traffic. For a lightly loaded cell with fixed traffic, all the configurations discussed offer a satisfactory behavior, whereas strong trade-offs between delay and throughput emerge in an over-loaded cell with fixed traffic. Moreover, the different scheduling algorithms diverge significantly in their response to the over-loaded scenario, with PMRS displaying better delay generally, AMRS displaying more DL rate with some delay increase, and TMRS being overwhelmed by the traffic. Applications on top of TCP are more sophisticated and adapt their rates to the network. In this case the severe delays of the over-loaded scenario are avoided by the rate adjustment, which converges to a significantly larger rate for the 4-stream models compared to the 1-stream baseline. Generally, PMRS offers consistently good performance in both UL and DL, whereas AMRS offers great throughput in DL but has severe shortcomings in UL.

4.5 Conclusions

In this chapter, we have studied the performance of various MU-MIMO HBF implementations for 3GPP NR mmWave cellular systems. We have shown that by supporting multiple transmission streams simultaneously, the system capacity is greatly increased. Moreover, by associating each frequency-flat BF vector to a separate antenna port, the signal processing involving large arrays characteristic of mmWave systems can be addressed using HBF architectures. In addition, by considering a linear matrix mapping logical transmission streams to physical antenna ports, it is possible to leverage well-known MU-MIMO signal processing

techniques in order to alleviate the inter-user side lobe interference and improve the SINR. We have shown that this is indeed necessary, as the SINR would degrade significantly if we merely used separate analog beams for each user without MU-MIMO-aware HBF. With regard to control overhead, we evaluated a frequency-flat MMSE BF scheme with reduced feedback that achieves partial interference removal, and a frequency-selective MMSE BF scheme with significantly more feedback that achieves almost complete interference removal.

Thanks to our full-stack framework, we have revealed a trade-off between retransmissions, MU-MIMO schedulers and BF design. Particularly, due to the characteristics of channel estimation in NR, only coexisting allocations that start at the same time are able to employ MU-MIMO-aware HBF techniques in order to reduce the interference. This raises a conflict between interference mitigation and RB allocation, as some wasteful padding symbols are needed to enforce the constraint that all allocations start at the same time. We have implemented two types of schedulers, one with padding and one that permits asynchronous transmissions and wastes no resources. We have shown that the latter scheduler leads to system performance degradation in UL transmissions but not in DL, since the events with too much interference are only occasional and may be compensated with adequate retransmission schemes, at the expense of some delay increase.

We have also studied the relation between the system throughput and delay performance indicators, the application data rates, and this scheduler-BF trade-off. In general the use of the padding scheduler displayed the most consistent behavior, achieving satisfactory delays with much higher throughput than a baseline 1-stream system in both DL and UL. On the other hand, the asynchronous scheduling approach cannot yet be fully discarded, as it offers even greater throughput in DL scenarios with retransmissions where delay is not a concern.

5

RAN Slicing in mmWave Cellular Networks

5.1 Introduction

5G networks have been designed to provide connectivity for different classes of services, with orthogonal requirements. For example, a packet error rate of 10^{-4} is tolerable in an eMBB system, where the focus is on the high throughput; however, when it comes to industrial real time applications, typical target values for reliability are in the order of 10^{-6}, together with low latency [179, 180]. It follows that the design of new generations of mobile networks should be flexible enough to adapt to the different requirements.

Network slicing, defined by the Next Generation Mobile Networks Alliance (NGNM) as the concept of running multiple indipendent logical networks upon a common physical infrastructure, has been proposed as an enabler of flexible 5G networks [11]. Specifically, a network slice is a self-contained, virtualized and independent end-to-end network that allows operators to execute different deployments in parallel, each based on its own architecture [181]. While there have been several research efforts focused on optimizing slicing operations in wired networks (e.g., in the core of cellular networks), and in traditional, sub-6 GHz wireless networks, the state of the art lacks considerations on how this can be

applied to the radio access of 5G mmWave networks.

In this chapter, we study how to effectively achieve network slicing in mmWave wireless networks, introduce flexible operations and satisfy heterogeneous traffic demands also in these frequency bands. Notably, we focus on how to serve URLLC and eMBB slices that share the same radio access resources, without compromising the quality of service of the users in either of the two.

The proposed slicing framework exploits CA, which enables multi-connectivity at the MAC layer by providing service on multiple links (called Carrier Components (CCs)). With CA, different CCs can use different frequencies, and can be adapted to the channel independently (i.e., they can use different MCSs, and/or retransmission processes), but are usually transmitted by the same base station. CA increases the available datarate for the user, since it aggregates the spectrum across multiple bands, but can also be used for agile interference management [182] and spectrum sharing with unlicensed bands with the LTE-U extension [183]. At mmWave frequencies, CA techniques can be used to combine carriers with very different propagation properties (e.g., 28 and 73 GHz) or in licensed and unlicensed bands [15] in order to improve the reliability of transmission and/or increase the throughput [6].

In our framework, CA is used to (i) distribute the URLLC and eMBB flows among different carriers, which could effectively act as slices; and (ii) provide frequency diversity, e.g., slices that require high reliability could be allocated in lower portions of the spectrum, which benefit from a reduced pathloss. Additionally, we introduce MilliSlice, a cross-carrier packet scheduling policy that dynamically adapts the dispatching of packets to the different carriers with the goal to maximize the utilization of the resources available in each CC, without penalizing the performance requirements of each slice.

We evaluate the effectiveness of the proposed solution with an open-source, realistic, end-to-end, full-stack network simulator for mmWaves [46] based on ns-3, which features the 3GPP channel model for mmWave frequencies and a 3GPP-like protocol stack with carrier aggregation. The results show that, compared to a mmWave network without slicing, the proposed solution reduces the latency of URLLC flows and increases the throughput of the eMBB streams, hence enhancing the QoS achieved by both slices at the same time.

The remainder of this chapter is organized as follows. In Section 5.2 we provide

a brief review of the state of the art regarding network slicing and CA solutions. Then, in Section 5.3, we introduce the slicing framework, focusing in particular on its novelty aspects compared to the currently available solutions in the literature. Section 5.4 provides a simulation-based performance analysis of the presented strategy, and finally we conclude this chapter and highlight possible future improvements in Section 5.5.

5.2 State of the art

This section will review relevant research efforts for the slicing of the RAN (Section 5.2.1) and carrier aggregation in sub-6 GHz and mmWave cellular networks (Section 5.2.2).

5.2.1 RAN Slicing

Although introducing network slicing at the RAN is still challenging, several 5G initiatives have been pushing for new frameworks to enable network slicing in mobile networks. [184] proposes a fully programmable network architecture based on a flexible RAN to enforce network slicing, also implementing a two-level MAC scheduler to share physical resources among slices, obtaining encouraging results in terms of throughput and resource allocation adaptability. Similarly, the authors of [185] envision fully virtualized LTE base stations that can be deployed on-the-fly to serve slices with different performance requirements Moreover, [186] analyzes the RAN slicing issue in a multi-cell network, presenting four different slicing approaches for splitting the radio resources among slices, and achieving high granularity and flexibility in the assignment of radio resources, as well as satisfactory levels of isolation. Paper [187] adapts a holistic approach to RAN slicing, proposing a framework that translates high-level service requests of the operators into a correct mapping of the physical layer resources. Finally, [188] proposes a novel latency-sensitive 5G RAN slicing solution for Industry 4.0 scenarios, where stringent latency requirements are common. This proposal, evaluated in industrial scenarios with mixed traffic types, is able to meet the latency requirements of delay-sensitive or time-critical applications, thus improving the QoS experienced by all traffic types through an efficient allocation of the resources to the slices. However, the schemes that have been proposed so far target traditional

sub-6 GHz deployments, while in this work we consider the application of network slicing to mmWaves cellular systems.

5.2.2 Carrier Aggregation

Carrier aggregation is a technique that the 3GPP has first introduced in the LTE specifications [189], and extended in NR [12], which enables different CCs to operate at different frequencies, and to use different MCS or retransmission processes, usually within the same base station. Moreover, CA allows the aggregation of licensed and unlicensed bands with LTE-U and Licensed-Assisted Access (LAA) [183]. The advantages that this approach can bring have been profoundly studied in the literature and eventually even implemented in actual deployments, but mostly within the realm of LTE-Advanced mobile networks: the employment of CA in such scenarios provides an increase of the available per user data-rate (since it can aggregate the radio resources across the spectrum) as well as the means for an agile interference management [190].

In 5G cellular systems, the capabilities of CA have been extended with the possibility of using up to 16 [191, 12] carriers with a bandwidth of up to 400 MHz. Moreover, as NR supports mmWave communications, it will be possible to combine carriers with different propagation properties (e.g., mmWave and sub-6-GHz) or in unlicensed and licensed bands (thanks to the extension of NR-U in the 60 GHz band) [15, 76], in order to increase the throughput and improve the reliability of transmissions [6]. In our previous work [42], we analyzed the performance of different CA schemes for mmWaves using an end-to-end network simulator [46], showing that CA improves the throughput of the network, due to the higher resilience to blockage given by macro-diversity and the higher efficiency of a per-carrier scheduling and MCS selection. However, even though the preliminary analysis carried out by means of simulation in [42] shows promising results, the application of this technique to mmWaves has not been exhaustively studied so far and presents some open challenges such as the introduction of joint-CC schedulers and MAC-PHY cross layers approaches.

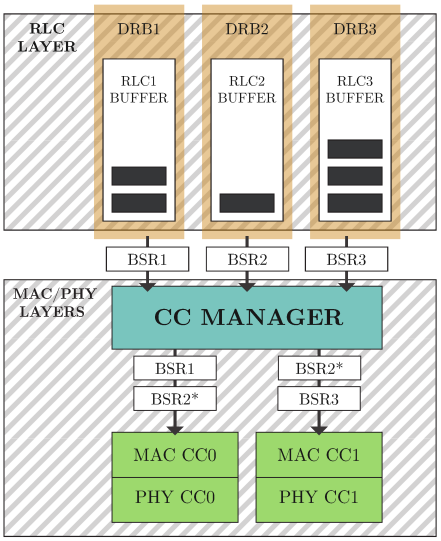


Fig. 5.1: Protocol stack of an NR device using CA, with a focus on the layers which play a role in the slicing framework. In our proposal the BSR messages coming from the above RLC layers are redistributed across the various CCs, possibly depicting different amounts of data compared to their original form (see BSR2 and BSR2*).

5.3 Efficient mmWave RAN slicing with CA

In this Section, we will describe the proposed RAN slicing framework for mmWave cellular networks, providing details on how CA can be used to perform slicing, and on the cross-carrier scheduling policy that manages to guarantee to each data stream the desired QoS.

The overall goal is to satisfy the requirements in terms of latency and reliability of URLLC flows, i.e., over-the-air delay below 1 ms and packet loss smaller than 10^{-6}, while maximizing the throughput of the eMBB flows that share the same radio interface. We designed the slicing framework to be robust with respect to the number of users per base station, the amount of eMBB traffic, and the configuration of the resource allocation in the access networks.

5.3.1 RAN Slicing Through CA

The CA technique involves the PHY and MAC layers, as well as the interaction between MAC and RLC. Figure 5.1 reports a simplified diagram of the protocol stack with the entities involved in the management of multiple carriers. During

the configuration phase, the base station notifies the availability of one or multiple carrier components, to which the UE could connect according to its capabilities. Once the connection setup is completed, the base station can manage the CCs by offloading users to different carriers, or by performing a cross-carrier scheduling for the users connected to multiple CCs. As described in the 3GPP specifications for NR [12], the inter-CC scheduling in CA happens in the highest portion of the MAC layer, which is interfaced with the different instances of RLC.* The RLC periodically sends to the MAC layer a BSR, a report with information about the occupancy of the different buffers (i.e., the size of the transmission and retransmission queues). The MAC layer then uses the BSRs to schedule the radio resources.

In this work, we propose to adapt the CA mechanism to achieve network slicing at the RAN. As previously highlighted, most of the solutions that have been introduced to perform slicing have been studied either to be deployed in the core network or, when implemented at the RAN, are based on ad hoc scheduling at the MAC layer with a single carrier. Conversely, the usage of CA, combined with mmWaves, has several benefits over a solution with a single carrier.

First of all, it allows the aggregation of multiple carriers, so that the telecom operators could use the available spectrum in a more flexible way. Additionally, CA enables the possibility to provide isolation among the different slices by serving each one with a different carrier. Finally, it makes it possible to exploit macro diversity, i.e., to allocate flows with different requirements in portions of the spectrum with distinct propagation characteristics. For example, a CC with a smaller carrier frequency exhibits a lower pathloss, but, at the same time, may be more constrained in terms of available bandwidth with respect to a CC at a higher frequency. This provides a natural fit to serve URLLC flows in the lower CC, as they could benefit from the improved propagation conditions but have limited needs in terms of bandwidth, and the eMBB traffic in the higher portion of the spectrum, trading reliability for a larger bandwidth. In our work, we follow this principle by always scheduling URLLC flows in the CC with the lowest carrier frequency.

In the proposed slicing framework, when a telecom operator needs to allocate

*In 3GPP networks, each end-to-end data flow is mapped to a Data Radio Bearer (DRB), which, in turn, corresponds to a specific pair of RLC and PDCP instances.

a new RAN slice for an end-to-end flow with a certain QoS level, it first checks if the base stations in the area where the slice should be served have CC available to host the slice. If this is the case, it specifies at the MAC layer of each base station the QoS requirements corresponding to the specific flow (e.g., whether it is an URLLC or eMBB flow). These requirements are expressed through a Quality Class Identifier (QCI), i.e., an indicator for the QoS of each end-to-end flow standardized by the 3GPP [192], associated to the BSRs generated at the RLC layer. Eventually, when the slice is operational, the MAC layer uses the QCI of the BSRs to map it to the proper CC. For example, in Figure 5.1, RLC3 serves an eMBB slice, and its BSRs are forwarded to CC1. Conversely, RLC1 is associated to an URLLC DRB, and will be scheduled on CC0. Notice that in this work we do not focus on the admission problem, but rather on optimization of the slice scheduling on the different CCs, as we will discuss in the next paragraphs.

5.3.2 Slice-aware Cross-Carrier Scheduling

As previously mentioned, CA enables, in principle, the orthogonal separation of the URLLC and eMBB slices in different CCs. However, as we will highlight in Section 5.4, this may lead to inefficiencies in the spectrum utilization, especially in the case where the slices have heterogeneous requirements in terms of bandwidth. In particular, even if the CC for the URLLC slices may be configured with a smaller bandwidth, the datarate difference between eMBB and URLLC flows, if not properly handled, can lead to the exhaustion of the available capacity in the eMBB slice, with idle resources in the CC for URLLC.

Therefore, as part of the proposed RAN slicing framework, we introduce MilliSlice, a cross-carrier scheduling component whose purpose is to improve the efficiency of the slicing process while avoiding detrimental effects on the QoS of the URLLC slices. Referring to Figure 5.1, this component is deployed in the CC manager, thus it does not require any modification in the per-carrier scheduling algorithm that the operator selects for each CC.

The slicing framework associates each slice to a *primary* carrier component, following the strategy described in the previous section, and, additionally, to one or more *secondary* CCs, in which the slice has a lower priority with respect to the slices that use these CCs as primary. The slices that have a low priority on a CC

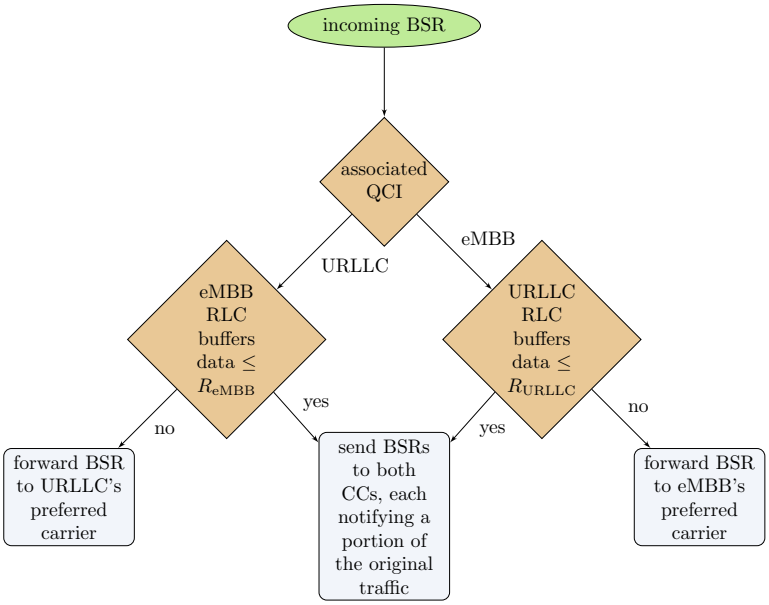


Fig. 5.2: Flow-chart of the BSR scheduling logic

will be served in that CC only if the flows that use it as primary will not occupy all the available resources. This makes the cross-carrier slicing mechanism adaptive to the load of each slice. Specifically, in the aforementioned case of URLLC and eMBB sources, the proposed method distributes the data across the CCs with the following criteria. The eMBB traffic shall be partially redirected towards its secondary CC if and only if the URLLC buffers (which consider this CC as primary) contain less data to transmit than a pre-determined threshold R_{eMBB}; a similar principle applies to the eMBB slices.

The process is based on an adaptive forwarding of the BSRs to the different CCs, as depicted in the flow chart of Figure 5.2. Notably, the component carrier manager, i.e., the entity in charge of splitting the traffic among the different carriers, tracks the buffer occupancy of the RLC layers with a sliding window mechanism. Then, once the RLC sends a BSR to the MAC, the CC manager checks the associated QCI and, if the buffer occupancy of the secondary carrier is above the predefined threshold, the BSR is forwarded to the primary CC only. Otherwise, the BSR is split across the primary and secondary CCs. The pseudocode in Alg. 5.1 extends this procedure for a generic number of secondary carrier components.

Algorithm 5.1 Cross-carrier scheduler implemented in the proposed RAN slicing framework.

Input: The incoming BSRs `BSR`, the `BufferOccupancyMap` at the CC manager, `QciCcMap`, associating QCIs to their primary carrier, and the set of thresholds R for each QCI

Output: `ChoosenCCs`, a map associating CCs and respective BSRs

Compute the aggregated RLC buffer occupancy (new packets + retransmissions), store it in `RlcLoads`
Consider `qci`, the QCI associated with the BSR `BSR` of a specific flow
if `Qci` $\in$ `QciCcMap`
    Add the primary CC to the list of available ones
    `ChoosenCCs[ QciCcMap[cci] ]` $\leftarrow$ `BSR`
    Check whether the RLC buffers of the various different slices contain less data than a given threshold, if so add them
    for all `entry` $\in$ `RlcLoads`
        `oQci` $\leftarrow$ QCI associated with `entry`
        if `qci` $\neq$ `oQci` **and** `RlcLoads[ oQci ]` $< R_{\texttt{oQci}}$
            `ChoosenCCs[ QciCcMap[oQci] ]` $\leftarrow$ `BSR`
    for all `cc` $\in$ `ChoosenCCs`
        `ChoosenCCs[cc]` $\rightarrow$ `BSR.TxQueueSize` $=$ `BSR.TxQueueSize` / **size**(`ChoosenCCs`)
return `ChoosenCCs`

Furthermore, we choose the carrier operating at lower frequency to be the primary for the URLLC flow, and to set the threshold $R_{\text{eMBB}} = 0$, so that the URLLC traffic is never redistributed across the CCs (i.e., it will be served only by its primary CC). This is due to the fact that URLLC packets would experience a lower average SINR on secondary carriers, as the primary is chosen to be the one with the lowest carrier frequency and, additionally, they would be handled with low priority in secondary CCs, thus impacting latency and reliability. Conversely, for the eMBB traffic, we set $R_{\text{URLLC}} = 1$ packet, so that these slices can be served by the secondary CC when the URLLC RLC buffers are empty.

5.4 Performance analysis

This section will provide insights on the performance that can be achieved using the proposed RAN slicing framework. The performance analysis has been carried out using simulations with the open-source network simulator ns-3, which al-

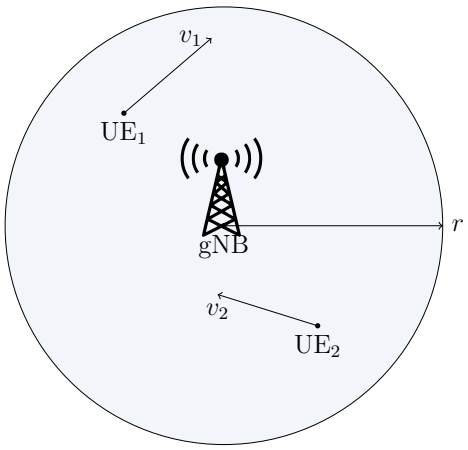


Fig. 5.3: Simulation scenario, with UEs randomly moving in a circle with radius r around a gNB.

lowed us to accurately analyze the end-to-end performance of the proposed slicing framework. Specifically, the simulations are based on the ns-3 mmWave module introduced in [46], which features the 3GPP channel model for mmWaves [123], to stochastically characterize propagation loss, fading, beamforming and interference in the wirless domain, a 3GPP-like protocol stack for gNBs and UEs, and, thanks to the integration with ns-3, the possibility of simulating different mobility patterns and the TCP/IP protocol stack.

To implement the slicing framework proposed in this work, we consider the implementation of CA for the ns-3 mmWave module described in [42]. The CC manager that behaves according to the policies described in Section 5.2.1 is an extension of the `MmWaveNoOpComponentCarrierManager` class, which adaptively forwards the BSRs from the RLC instances to the MACs of the various CCs. Additionally, we implemented a complete simulation script that can be used to instantiate slicing scenarios and compare different network configuration. The open-source code base associated to this work is publicly available,[†] so that researchers interested in the area of RAN slicing can use it to further extend this work.

[†] `github.com/signetlabdei/millislice`

5.4.1 Simulation Scenario and Parameters

We consider a scenario that models the coverage area of a cell in an urban, densely populated area. As represented in Figure 5.3, the simulation scenario consists of a single cell of radius 200 m, with one gNB at the center and N_U users that are uniformly dropped and move with random speed between 1 and 10 m/s. A remote host connected to the Internet holds eMBB and URLLC applications, modeled as UDP sources with different data rates, each generating downlink traffic for a specific user. The system operates at 28 GHz with a total bandwidth of 500 MHz. In case CA is used, an additional carrier component operating at 10 GHz is added and the overall bandwidth is divided among the two carriers according to the parameter cc_{ratio}, which defines the ratio between the bandwidth dedicated to CC_0 and that of CC_1, e.g., when $cc_{ratio} = 0.5$, each CC is configured with a bandwidth of 250 MHz. In our solution, CC_0 will act as preferred carrier for the eMBB slice, while CC_1 will be dedicated to URLLC flows. Finally, as previously mentioned, for this simulation campaign we set $R_{\text{eMBB}} = 0$ and $R_{\text{URLLC}} = 1$. With this configuration, URLLC data is never sent to the eMBB CC, while eMBB slices can be served by their secondary CC only if the RLC buffers corresponding to the URLLC slice are empty. For a more exhaustive list of simulation parameters, please refer to Table 5.1.

5.4.2 Network Configurations and Metrics

We consider two different baselines to benchmark the performance of the proposed slicing framework. The first ("no CA" in the plots) is a setup without CA and slicing, i.e., with a single carrier with the total system bandwidth B. The second ("CA, primary only" in the plots), instead, is a solution with slicing and CA, but without the adaptive cross-carrier scheduling, i.e., in which each slice has only a primary CC and cannot use the secondary CC.

We evaluated the performance of the proposed framework by analyzing the average end-to-end delay, aggregated throughput and packet loss ratio achieved at the application layer for both the eMBB and URLLC data flows. Moreover, to evaluate the per-carrier efficiency in terms of resource utilization, we defined the metric $\eta_{\text{CC}_\text{i}}$, which represents the portion of the consumed resources with respect

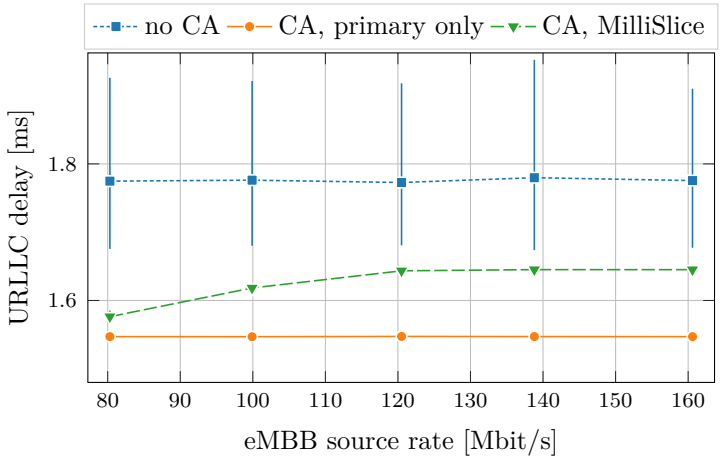
Table 5.1: Simulation parameters

Parameter	Value
Total System Bandwidth B	500 MHz
CC_0 center frequency f_0	28 GHz
CC_1 center frequency f_1	10 GHz
eMBB primary CC	CC_0
URLLC primary CC	CC_1
RLC Mode	Acknowledged
BSR timer	1 ms
cc_{ratio}	0.5
Number of URLLC UEs	10
Number of eMBB UEs	10
eMBB source rate	$[80, 100, 120, 140, 160]$ Mbit/s
URLLC source rate	$[1, 1.5, 2]$ Mbit/s
Radius r	200 m
UE speed	$\mathcal{U}\,[1, 10]$ m/s
R_{URLLC}	1 packet
R_{eMBB}	0

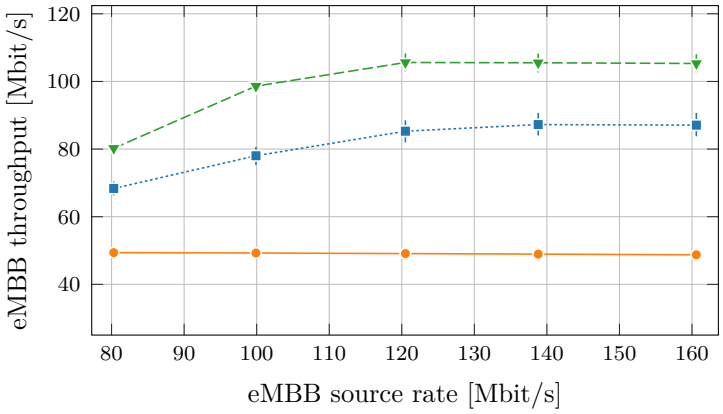
of the total available:

$$\eta_{\mathrm{CC_i}} = \frac{tx_{sym}[\mathrm{CC_i}]}{t_{sym} \cdot f_{frame} \cdot f_{subframe} \cdot f_{sym}} \times \frac{B_{CC_i}}{B} \tag{5.1}$$

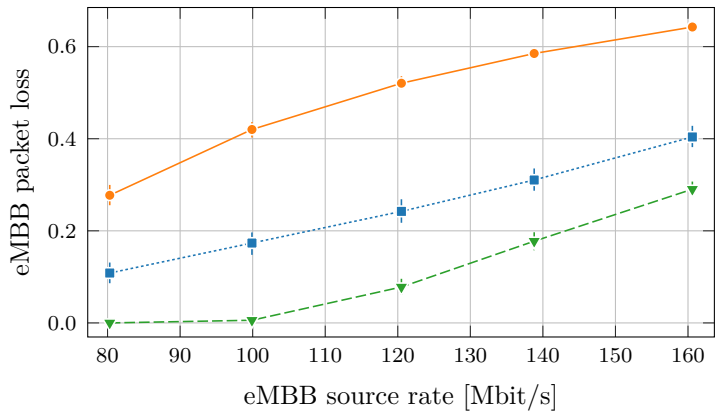
where $tx_{sym}[CC_i]$ is the total amount of OFDM symbols transmitted through $\mathrm{CC_i}$, t_{sym} is the simulation time in seconds, f_{frame} is the number of frames in a second, $f_{subframe}$ is the number of subframes within a frame, and f_{sym} is the number of the OFDM symbols which can be transmitted in a subframe. Moreover, the weight B_{CC_i}/B represents the portion of system bandwidth dedicated to $\mathrm{CC_i}$, and is applied to achieve a normalized result.



(a) Average URLLC delay.



(b) Average eMBB throughput.



(c) Average eMBB packet loss.

Fig. 5.4: Per-user performance metrics achieved for different values of the eMBB source rate; the URLLC data-rate is fixed at 1.0 Mbit/s.

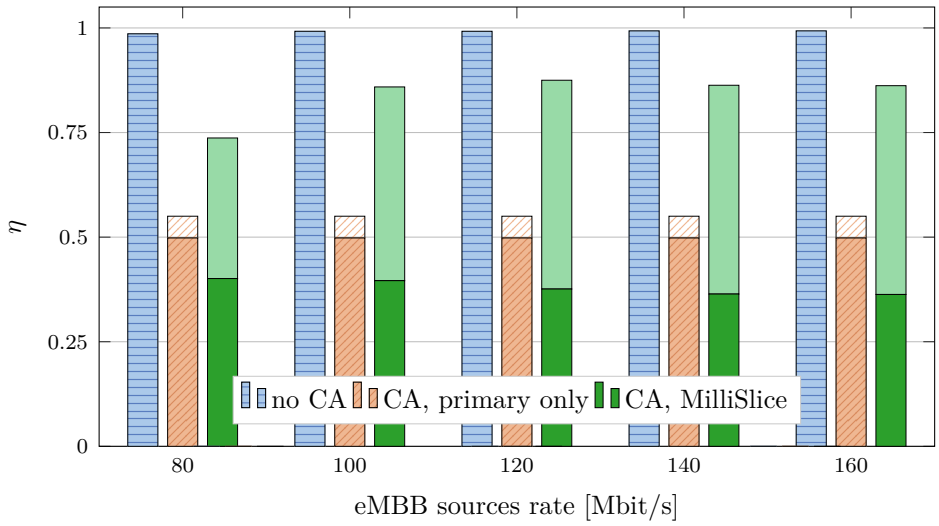


Fig. 5.5: Evaluation of the resource utilization versus different values of eMBB source rate and with URLLC data-rate fixed at 1.0 Mbit/s. The darker portions of the bars represent η_{CC_0}, the lighter represent η_{CC_1}.

5.4.3 Results

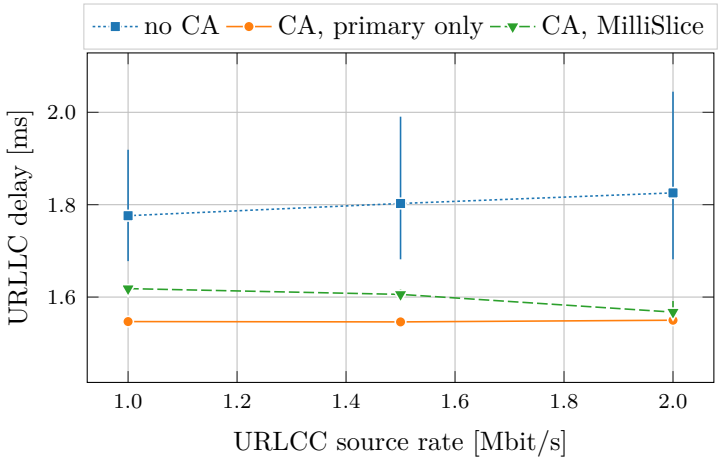
In Figure 5.4, we compared the performance achieved by the three strategies over different values of the eMBB source rate. Although all the solutions are able to guarantee a reliable delivery of the URLLC traffic, it can be noticed that the introduction of RAN slicing by means of carrier aggregation is beneficial for the delay: indeed, both the primary only and MilliSlice solutions show lower URLLC delay compared with the standard approach. In particular, the lowest delay is reasonably achieved when the two flows are completely isolated, because the usage of dedicated carriers allows URLLC transmissions to be independently scheduled, without incurring additional delay due to the presence of other eMBB packets in the queue. Moreover, the possibility to employ a carrier operating at a lower frequency ensures a more reliable data delivery, making it possible to achieve the correct reception of each packet with a smaller number of MAC and RLC layer retransmissions, thus reducing the delay. However, the advantage that the complete isolation provides for URLLC traffic comes at the price of sacrificing the QoS experienced by the eMBB slice, which exhibits lower throughput (Figure 5.4b) and higher packet loss compared with the other solutions (Figure 5.4c). In this case, the carrier component dedicated to the eMBB flow does not provide

enough resources to satisfy the offered traffic and becomes saturated. Instead, MilliSlice is able to achieve the best performance for the eMBB services while minimizing the URLLC delay with respect to standard systems, and thus represents a viable solution to achieve network slicing at the RAN level. Thanks to an elastic scheduling algorithm, our solution is able to efficiently exploit the available resources by allowing the congested eMBB slice to use the carrier dedicated to the URLLC flow when idle. This behavior is confirmed by Figure 5.5, which represents the resource utilization achieved by the three different approaches, possibly showing the portion used by either CC_0 (darker) or CC_1 (lighter) when CA is employed. It can be seen that with MilliSlice more than 80% of the system resources are exploited and the load is equally distributed among the two carriers. In contrast, with the primary only approach the carrier dedicated to URLLC is poorly utilized and about 45% of the available resources are wasted. Moreover, the more agile link adaptation provided by CA [42] enables MilliSlice to achieve a higher performance gain with respect to the single carrier approach, even using a smaller amount of resources.

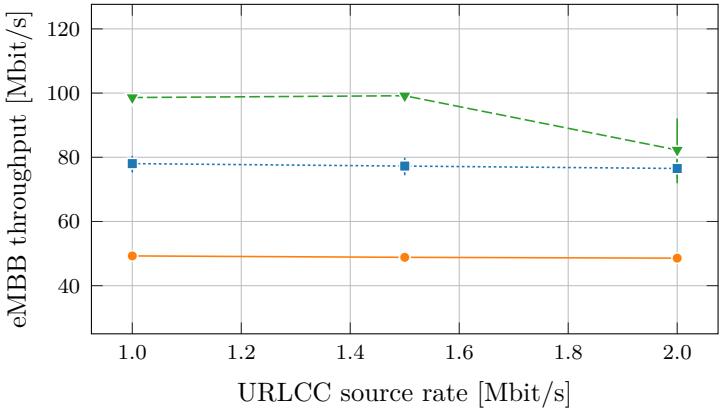
If, on the other hand, we analyze the effectiveness of MilliSlice across different URLLC source data-rates, we can recognize a similar general trend of the various metrics: in Figure 5.6 our solution exhibits a higher throughput and lower packet loss for the eMBB flow compared with the other solutions, coupled with a reduction of the URLLC delay with respect to the single carrier approach.

However, by observing Figures 5.6b and 5.6c, it can be noticed that the gain introduced by MilliSlice decreases when increasing the rate of the URLLC sources. This phenomenon can be interpreted as follows: as the amount of URLLC traffic increases, the BSR arrival occurrences indicating that the respective RLC buffers as empty significantly decrease; in turn, given our threshold choices, this results in a reduction of the scheduling instances implementing a redistribution of the traffic across the different CCs, hence the inability to sustain the eMBB demands. Nevertheless, we deem possible to significantly enhance the effectiveness of our CC usage policy by coupling such strategy with an ad hoc, slicing-oriented, MAC layer scheduling, as such choice would enable different and specifically more aggressive BSR redistribution strategies by the component carrier manager.

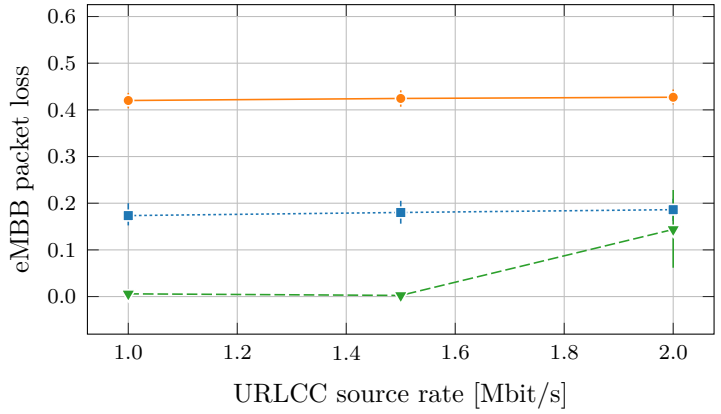
To evaluate the robustness of the proposed scheduling algorithm to possible scenario variations, we analyzed the system behavior by varying the number of



(a) Average URLLC delay.



(b) Average eMBB throughput.



(c) Avergage eMBB packet loss ratio.

Fig. 5.6: Per-user performance metrics achieved for different values of URLLC source rate; the eMBB data-rate is fixed at 100 Mbit/s.

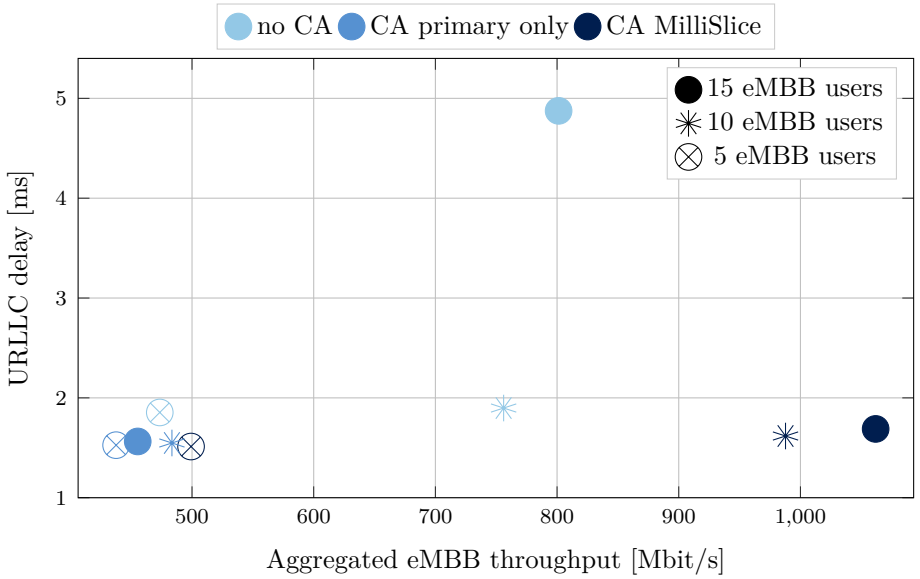
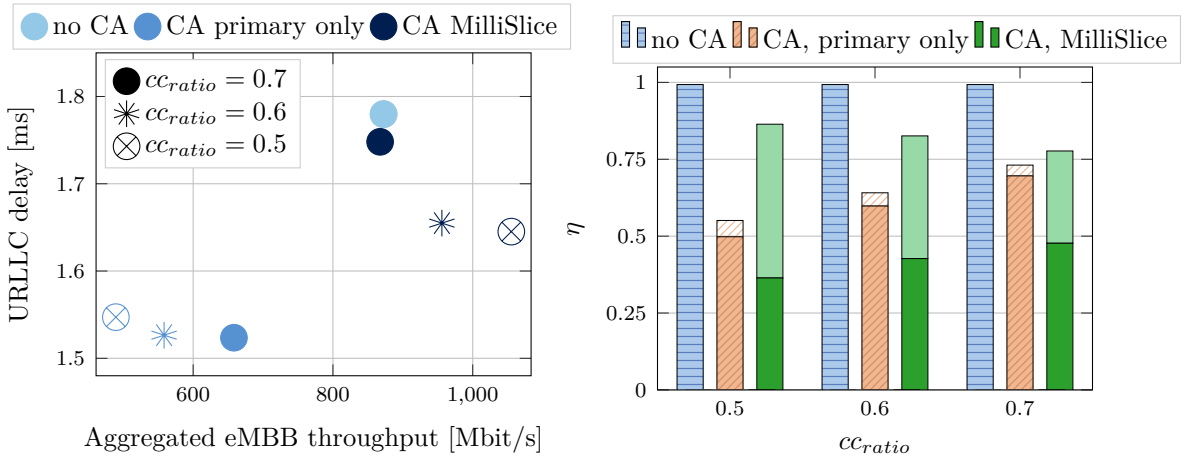


Fig. 5.7: Average URLLC delay versus aggregated eMBB throughput.

users. The results are shown in Figure 5.7, in which each point represents the achieved performance in terms of average URLLC delay and aggregated eMBB throughput when considering a certain amount of users. One one hand, in the single carrier case, the lack of any slicing strategy makes the URLLC performance susceptible to the increase of the number of eMBB sources. On the other, the static carrier assignment isolates the two traffic types onto their favored carrier and lacks any degree of adaptability to the offered eMBB traffic. Intead, MilliSlice manages to scale well and sustain different amounts of eMBB sources while keeping the URLLC delay under 2 ms.

Finally, in Figure 5.8a we can observe how our proposed solution shows poor adaptation capabilities with respect to a variation of the cc_{ratio}: as one of the carriers starts to gain possession of most of the bandwidth, the simplicity of our traffic redistribution strategy, coupled with the lack of ad hoc MAC layer scheduling solutions, starts to show some limitations, even though it still outperforms the other solutions. In particular, such loss in the effectiveness of our policy is driven by a sub-optimal exploitation of the system bandwidth: as depicted by Figure 5.8b, the CC whose dedicated resources are lower tends to be backlogged, while the other one does not absorb as much traffic as it would be capable.



(a) Average URLLC delay versus aggregated eMBB throughput.

(b) Evaluation of the resource utilization over different values of cc_{ratio}. The darker portions of the bars represent η_{CC_0}, the lighter represent η_{CC_1}.

Fig. 5.8: Evaluation of the system behavior when changing distribution of the system bandwidth among the carrier components by means of the parameter cc_{ratio}. The URLLC sources rate is fixed to 1.0 Mbit/s, while the eMBB sources rate is fixed to 140 Mbit/s.

5.5 Conclusions and Future Work

The variety of services that 5G networks will have to support requires both the exploitation of previously unexplored portions of the spectrum (i.e., the mmWave frequencies) and of additional flexibility in the RAN configuration. In this chapter, we proposed to combine two enablers of 5G networks, i.e., network slicing and carrier aggregation, to support in the same radio interface simultaneous transmission of URLLC and eMBB traffic flows. Specifically, we proposed a simple but effective policy for the distribution of the various traffic flows among different slices, mapped across multiple carrier components, also exploiting the diversity of the different frequency bands available at mmWaves. We implemented such solution in the ns-3 mmWave module, and carried out an extensive simulation campaign, benchmarking our solution with a number of metrics against two different baseline policies. The promising results and the effectiveness of the proposed solution showed that network slicing through carrier aggregation, especially when coupled with an adaptive cross-carrier scheduling, can sustain heterogeneous 5G requirements.

Future work will focus on more refined solutions, aimed at improving the operations of the schedulers that operate at the carrier component level, to make them aware of the kind of traffic flow they need to support, and to integrate more advanced policies in the proposed slicing framework.

6

Towards Millimeter Wave Vehicular Networks

6.1 Introduction

The rapid evolution towards 5G wireless networks will accelerate the adoption of solutions for Connected Intelligent Transportation Systems (C-ITSs) to deliver improved traffic safety and efficiency through autonomous driving [193]. These systems, whose market estimates are in the order of 7 trillions USD, promise to make the number of road accidents drop by as much as 90%, while carbon emissions will reduce by more than 60%. The hands-free driving environment of C-ITSs can also reduce drivers' stress and tedium, as well as increase their productivity. C-ITSs could save over 2.7 billion unproductive hours annually in the US in work commutes, according to some estimates [194].

When fully commercialized, C-ITSs will support several use cases whose requirements will likely exceed the capacity of current communication technologies for vehicular networks [195, 196]. For example, for cooperative perception services, where vehicles exchange processed sensor data to improve the coverage and accuracy of environmental perceptions, the data rate requirements can reach up to approximately 1 Gbps for high-quality uncompressed images. For advanced safety applications, instead, latency must be very small (i.e., less than 100 ms for high

degree of automation) to ensure prompt reactions to unpredictable events [197].

The potential of Connected Autonomous Vehicles (CAVs) will be fully unleashed through V2X wireless communications, providing connectivity to and from cellular base stations (V2I) and among vehicles (V2V). Today, the two key access technologies that enable V2X communications are IEEE 802.11p and 3GPP Cellular-V2X (C-V2X) that, however, fall short of fulfilling the foreseen extreme traffic demands (e.g., in terms of very high throughput, ultra low latency and ultra high reliability) of future vehicular services.

In this regard, different standardization activities are currently being promoted by the IEEE and the 3GPP, with the 802.11bd [198] and NR V2X [199] specifications, respectively, to overcome the limitations of current technology. Both standards aim at boosting the wireless capacity by encompassing the possibility of using, besides traditional sub-6 GHz frequencies (that may support only basic safety services), the lower part of the mmWave spectrum, which features the availability of large chunks of untapped bandwidth. This would enable data rates in the order of hundreds of megabits per second [70] to support more advanced use cases (from semi- or fully-automated driving to cooperative perception), and improve over 3GPP C-V2X and IEEE 802.11p, which can reach – at most – a few tens of megabits per second [200]. Additionally, the unique characteristics of the mmWave signal, including the channel sparsity and the high temporal and angular resolution, may be used for very accurate positioning of vehicles, a critical requirement for most future vehicular services [201]. However, communication at mmWaves introduces serious challenges for the whole protocol stack and requires the maintenance of directional transmissions [200], due to severe path and penetration losses: even though IEEE and 3GPP research activities are in their initial stages, adequate discussion on whether (and how) standardization proposals will be able to overcome such limitations is still missing.

In this chapter, we discuss how mmWave operations can be efficiently integrated in IEEE 802.11bd and 3GPP NR V2X systems. Specifically, we focus on the V2V component of these specifications, and, unlike existing literature reviewing vehicular standard developments, e.g., [202], we shed light on potential shortcomings that future releases need to overcome to fully enable V2V operations at mmWaves. We focus on PHY, MAC, and higher-layer design challenges, including the issues related to channel estimation, synchronization, mobility management, resource

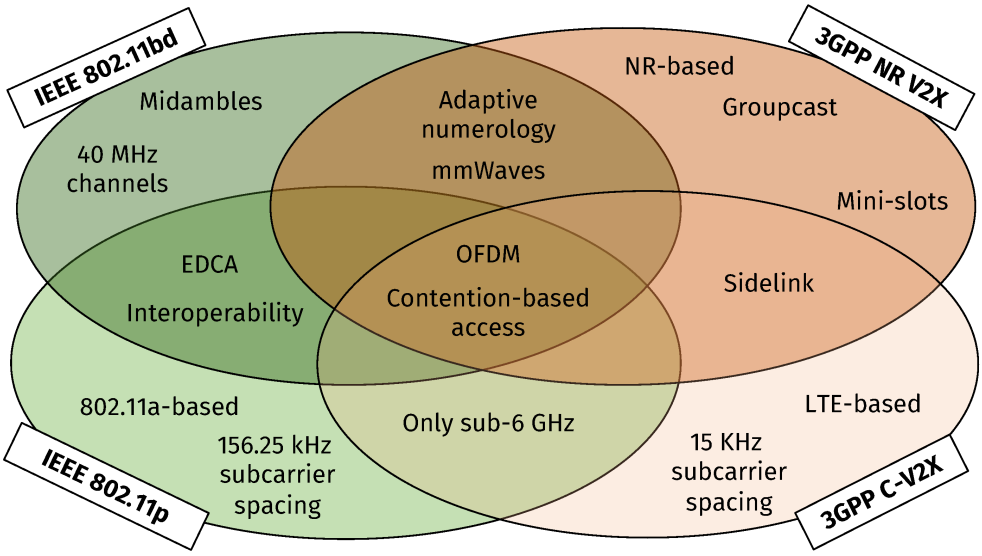


Fig. 6.1: Common characteristics and differences of the different V2V specifications.

allocation and congestion and flow control. Besides stimulating further research towards mmWave-compliant IEEE/3GPP specifications, we validate mmWave solutions in view of the strict requirements of future vehicular systems, a research challenge that is still largely unexplored. To this aim, we introduce MilliCar, a novel full-stack simulator for NR-V2X-compliant V2V networks operating at mmWave frequencies. Moreover, we present the results of a performance evaluation carried out using MilliCar, which shows that mmWave communications can be efficiently exploited in vehicular scenarios but within a limited range.

The rest of the chapter is organized as follows. In Section 6.2, we describe the ongoing standardization efforts for next-generation vehicular networks which foresee the support of mmWave communications. In Section 6.3, we discuss about the open challenges related to high-frequency operations considering the whole protocol stack. We introduce MilliCar in Section 6.4, while present our end-to-end performance evaluation and discuss the results in Section 6.5. Finally, we conclude the chapter in Section 6.6.

6.2 V2V Standardization Activities

The Institute of Electrical and Electronics Engineers (IEEE) and the 3GPP are standardizing next-generation networks for vehicular applications with IEEE

802.11bd and 3GPP NR V2X. They are designed to safely coexist with previous generations of the same technology (i.e., 802.11p and C-V2X, respectively) in the same deployment [202], are based on OFDM, with an adaptive physical layer design, and can use both sub-6 GHz and mmWave bands with contention-based schemes, as highlighted in Figure 6.1. Nonetheless, they also present some distinct characteristics, which are inherited from the different original designs of 3GPP and IEEE networks. NR V2X uses a sidelink for V2V operations, which could also be scheduled, while 802.11bd is based on the 802.11 Enhanced Distribution Channel Access (EDCA). Other differences, related to the specific physical layer and signaling configurations, will be discussed in the following paragraphs.

6.2.1 IEEE 802.11bd

In March 2018, the IEEE formed the 802.11 Next Generation V2X (NGV) Study Group, to improve the 802.11 MAC and PHY layers for V2X communications. The current V2X IEEE specifications, i.e., Wireless Access in Vehicular Environments (WAVE), with 802.11p for the PHY and MAC layers, is derived from 802.11a - 2009, and is no longer able to guarantee the present and future needs of vehicular applications.

The new amendment (commonly known as 802.11bd) targets communications at 5.9 GHz and, optionally, in the spectrum from 57 GHz to 71 GHz. NGV receivers must be able to interpret also 802.11p messages, while transmitters have to guarantee coexistence and backward compatibility between 802.11p and 802.11bd. 802.11bd reduces the End-to-End (E2E) latency, increases the throughput and the communication range (up to twice those yielded by 802.11p), and doubles the relative speed between vehicles (i.e., up to 500 km/h). To meet these requirements, 802.11bd foresees:

- the usage of Low-density Parity Check Code (LDPC) codes with midambles, i.e., specific portions of a frame in between OFDM data symbols that improve the channel estimation in fast varying channels [203];
- flexible sub-carrier spacing, with up to 40 MHz channel bandwidth.

No specifications for mmWaves have been released yet by the IEEE, except for a proposal to upgrade part of the PHY and lower MAC layers to those designed

Table 6.1: Millimeter wave challenges in IEEE 802.11bd and NR V2X standards.

	Open Challenges	Explanation
PHY Layer	Numerology design	Longer slots lead to channel variations
	Multiple antenna arrays	Synchronization with distributed antennas
	Joint radar and communication	Based on IEEE 802.11ad (static and indoor) scenarios
	Broad/multi/groupcast communication	Directionality precludes broadcast operations
	Channel estimation	Challenging in time-varying channel
	Synchronization	Synchronization signals need to be directional
MAC Layer	Mobility management	Directionality complicates vehicle discovery and retransmissions
	Resource allocation	CSMA strategies suffer from increased deafness
	Interference management	Unscheduled and autonomous sidelink transmission prevents interference coordination
Higher Layers	Multi-hop and routing	Routing is complicated by highly volatile links
	Multi-RAT support	RATs coexistence in the same frequency band, vehicle, and/or deployment
	Congestion and flow control	Suboptimal interaction between channel variability and transport layer rate estimation
Modeling	Channel design	Characterization of mmWave channels in dynamic environments

for 11ad/11ay high data rate scenarios [204], although these standards have been designed to target indoor communications. Therefore, there is an ongoing discussion on how to address the specific challenges of this frequency range, with preliminary studies based on 802.11ad/ay.

6.2.2 NR V2X

The 3GPP has specified in Study Items for Releases 15 and 16 that C-V2X (defined specifically for LTE in Release 14, but with a forward compatible design) will be extended into NR V2X, enabling next generation use cases such as vehicle platooning and advanced and remote driving, and high-data-rate sensor data transmissions.

The novelties investigated by the 3GPP are:

- direct measurement of the Sidelink (SL) channel, or decoding of Physical Sidelink Control Channel (PSCCH) transmissions, to identify occupied SL resources;
- multiplexing of different logical channels [199], along with the definition of the resource allocation modes 1, where the base station schedules the resources, and 2, which lets the UE autonomously select the sidelink transmission resources. Mode 2 is the more likely candidate for an initial deployment of NR V2X, given that mode 1 would require cellular network operators to upgrade their base stations to the NR V2X specifications, with increased deployment and management costs;
- support of mini-slot scheduling, i.e., the possibility to immediately schedule a transmission in just a portion of the 14 OFDM symbols specified for an NR slot, for latency-critical services;
- improvement of the localization accuracy of vehicles, leveraging the additional spatial and angular degrees of freedom provided by operations at mmWaves and the utilization of large antenna arrays;

With respect to the PHY layer numerology, no specifications have been released yet; the assumption has been to use a flexible numerology as described in 3GPP Release 15, with sub-carrier spacings of 60 and 120 KHz in Frequency Range

2 (FR2), i.e., between 24.25 and 52.6 GHz. Many other features are derived from NR. Moreover, no further specifications have been provided about resource allocation and channel sensing at mmWaves. Channel access schemes have not yet been specified for Release 16 and, due to lack of time until the end of the current release, NR V2X SL enhancements will be discussed from Release 17 on. As of December 2019, 3GPP Release 17 NR V2X activities include (i) SL evaluation methodology updates; (ii) low-power low-latency resource allocation enhancement, especially for mode 2; (iii) SL discontinuous reception options for broadcast, groupcast, and unicast; and (iv) support of new SL frequency bands for single-carrier operations, including FR2-specific enhancements [205]. Finally, a channel model for V2X communication in the sub-6 GHz band (FR1) and at mmWaves (FR2) is described in [66].

6.3 V2V Operations at MmWaves: Open Challenges

As introduced in Section 6.1, V2V standardization is moving full pace ahead, and will need to address the challenges introduced in the whole protocol stack by mmWave frequencies. The following subsections and Table 6.1 discuss open issues and research directions.

6.3.1 PHY Layer Challenges

Numerology design Both 802.11bd and NR V2X RATs support a flexible PHY frame structure, to address different QoS requirements. A longer symbol duration (i.e., a smaller subcarrier spacing) reduces the impact of noise, but may also suffer from sudden channel variations [24], making mmWave V2V communications more challenging. As a consequence, the NR V2X frame structure can be configured in a *self-contained* fashion, i.e., different sub-frames can be associated to a different numerology. In this way, it is possible to arrange a shorter symbol duration to support high-data-rate low-latency applications (e.g., for cooperative perception and/or remote driving services), while a lower subcarrier spacing can be reserved for narrowband communications to exchange basic safety information.

Multiple antenna arrays mmWave networks must establish directional transmissions to sustain an acceptable communication quality with beamforming. This is achieved using high-dimensional phased arrays, possibly placed in

distributed locations [24]. Distributed antennas improve the spectral efficiency by exploiting spatial diversity, thereby resulting in less correlated channels, but raise synchronization issues and require the design of efficient transmit power allocation and resource management mechanisms [206]. In these regards, zero-forcing and intra-block diagonalization schemes offer a good trade-off between capacity and system complexity considering power constraints, even though more advanced studies are needed before distributed antenna solutions can be applied to vehicular networks.

Joint radar and communication The use of mmWaves in a vehicular context is not new, with automotive radars operating in the 77 GHz spectrum. Dual-functional stacks integrating radar and V2V communications have already been investigated in the literature [207], but not combined yet in V2V specifications. Spectrum isolation or interference mitigation schemes typically enable their coexistence, but a better performance would be achieved by multiplexing both sensing and data on the same waveform, thereby improving resource utilization while reducing hardware cost and size.

Broad/multi/groupcast communication Directionality may preclude broadcast communications at mmWaves, if different directions cannot be used simultaneously (as in analog beamforming). On the other hand, transceivers with hybrid and digital technologies can beamform towards as many directions as the number of radio-frequency chains in the phased array, thereby achieving broad/multi/groupcast communications. Such architectures, however, are currently limited by hardware design and suffer from high energy consumption and computational complexity, which is critical considering the limited on-board resources of budget car models. To be energy efficient, digital/hybrid beamformers will need to use appropriate precoding techniques as well as converters with one or few bits of resolution. Discontinuous reception (DRX) modes, which enable receiving vehicles to temporarily disable their radio-frequency front end, can offer significant power savings when the traffic is intermittent, as in the case of vehicular scenarios [208].

Channel estimation Tracking the channel quality in multiple spatial directions will increase the channel estimation overhead at mmWaves. This is particularly challenging in V2V applications, where the channel varies quickly over time, and the initial estimate may rapidly become obsolete. Even though IEEE 802.11bd foresees the use of midambles [203] to handle channel variations, beamformed mmWave transmissions require specifically tailored channel estimation and precoding techniques. Furthermore, the exchange of channel state information (e.g., through the new Physical Sidelink Feedback Channel (PSFCH) in NR V2X) needs to be timely, to avoid the feedback of stale information in scenarios with a highly variable channel (e.g., because of the increased Doppler effect at mmWaves) [200].

Synchronization IEEE 802.11bd and 3GPP NR V2X mode 2 (a) specifications support autonomous sidelink operations without base stations [198, 199]. In this case, vehicles should maintain or acquire time and frequency synchronization with other users. To this end, synchronization signals can be exchanged in pre-defined resource pools, even though the directional nature of the communication at mmWaves may slow down the rate at which such information is acquired, thereby compromising robust synchronization.

6.3.2 MAC Layer Challenges

The issues that mmWaves introduce at the MAC layer in V2V scenarios stem from the lack of omnidirectional sensing and signaling, due to beamformed communications. Beamforming, indeed, introduces deafness to vehicles which are not beam-aligned, and complicates the design of channel access and neighbor discovery schemes. Moreover, these challenges add to those typical of the MAC layer in vehicular ad hoc scenarios.

Vehicle discovery and mobility management Directionality complicates an efficient and quick discovery of neighboring vehicles [200]. In the Vehicle-to-Network (V2N) context, the base stations have fixed locations. In V2V scenarios, instead, both endpoints move and could be within reach for just a few seconds. Therefore, 802.11bd and NR V2X signaling schemes should allow the vehicles to

discover each other quickly, even when considering mmWave directional transmissions, and rapidly adapt the communication endpoint in highly mobile environments. Moreover, the volatility of the connection caused by the mmWave channel and by the mobility of vehicles makes retransmissions more complex. MmWave systems can hence leverage automotive sensors, including Light Detection and Rangings (LiDARs) and videocameras, that gather information about the environment and classify surrounding objects: acquisitions from these sensors can then be used to reduce the overhead associated with link configuration and beam management, since, for example, the transmitter can detect the position of the receiver and estimate the optimal direction of communication.

Channel access and resource allocation As mentioned in Section 6.2, the 3GPP will likely introduce contention-based channel access in NR V2X (i.e., with the aforementioned mode 2), as in IEEE 802.11bd. When both specifications will be extended to mmWaves, they will need to cope with the interaction between directionality and the channel sensing schemes. The classic Carrier Sense Multiple Access (CSMA) strategies, prone to the hidden node problem even in sub-6 GHz bands, suffer from increased deafness at mmWaves. Moreover, contention avoidance messages, which broadcast the intent to occupy the channel, may not be received by every vehicle. While this issue can be partially alleviated by scheduling transmissions based on the congestion level, more accurate solutions will be discussed in the standards groups [202]. Finally, the high mobility of the nodes may introduce unforeseen collisions (e.g., when a transmitting vehicle changes path) but also free up channel resources (e.g., when a vehicle moves outside the communication area). Therefore, the design of efficient uncoordinated channel access procedures in dynamic vehicular scenarios at mmWaves is even more challenging than in WLAN systems. Notice that the highly-volatile nature of the mmWave channel in the vehicular scenario may create a larger response time for the Adaptive Modulation and Coding (AMC) scheme loop at the MAC layer, hence requiring a margin to compensate for the possible outdated CQI: this may lead to a suboptimal use of the transmission capacity.

Interference management Directional communications at mmWaves can isolate the users, reducing the interference and leading the network towards a

noise-limited regime [24]. Nevertheless, the degree of isolation depends on the density of vehicles and the level of spatial multipath, and interference may not be negligible in some deployments. In these scenarios, interference management schemes may help improve the network capacity by scheduling transmissions to minimize the interference. For example, the infrastructure-based and ad hoc deployments can be mixed to allow the network and the vehicles to coordinate and decide which resources should be blanked to avoid interference with V2V communications. For the out-of-coverage case (supported by both 802.11bd and NR V2X), instead, vehicles autonomously determine sidelink transmission resources, thus further complicating interference management.

6.3.3 Higher Layer Challenges

Multi-RAT support In next-generation V2V networks, different technologies will coexist in the same vehicle and deployment, using the same or different frequency bands. For example, multi-connectivity techniques, that combine sub-6 GHz and mmWave bands, could provide additional robustness to V2V operations. The different RATs from the 3GPP and IEEE should therefore be aware of each other, possibly with a user-plane integration at some layer. This integration can be exploited to efficiently disseminate the information over the different RATs, to combine the benefits of complementary technologies, and make up for the limitations of a mmWave standalone system.

Multi-hop communications and routing Multi-hop relaying schemes can extend the limited mmWave range for V2V. In particular, far-away vehicles may be interested in communicating through other vehicles that act as relays. V2V network operations will have to cope with efficient routing and successful delivery of packets in networks with highly volatile links, exacerbating the issues that traditionally affect vehicular ad hoc networks [209]. While routing is generally performed at the network layer, for such challenging scenarios a cooperation with the 3GPP and IEEE stacks could enable faster routing updates, based on continuous and prompt refresh of the links available as next hops.

Congestion and flow control Communication in V2V scenarios will be mostly bursty and among two peer vehicles, exploiting a massive amount of band-

width in the mmWave bands. For such short flows, TCP may not be needed, and could actually worsen the performance. The congestion window growth could indeed throttle the rate available at the application layer. With multi-hop communications and longer flows, instead, congestion control is needed. In this case, however, the available congestion control algorithms may suffer from the sub-optimal interaction between the channel variability and the rate estimation at the transport layer, with consequent high latency and low resource utilization. Cross-layer solutions, in which the transport protocol is aware of the actual performance of the wireless RAT, would allow higher layers to quickly adapt and use the optimal operating mode for single- and multi-hop scenarios. Finally, another challenge is how to provide reliability at the transport layer, e.g., through retransmissions, network coding or other Forward Error Correction (FEC) schemes.

6.3.4 Modeling Challenges

Accurate channel and protocol stack modeling at mmWaves is an essential step towards proper vehicular protocol design and performance characterization. The 3GPP has specified how to characterize mmWave propagation for NR V2X in [66], without, however, investigating second order statistics (e.g., spatio-temporal correlation). This prevents the applicability of existing models to dynamic environments. Additionally, the effect of the correlation among signals in a multipath environment, e.g., the role played by reflections from adjacent vehicles, is currently underestimated. The impact of Doppler and fading, which is critical at high frequencies, has also not yet been numerically characterized. In this sense, new measurements and the usage of ray tracing techniques could provide further insights on the performance of V2V communications at mmWaves, together with full-stack simulations [70] and real-world experiments.

The research community has started proposing solutions to improve the performance of V2V communications at mmWaves [210, 211]. As of today, the lack of testbeds for mmWave V2V scenarios makes simulation the preferred means for the performance evaluation of novel networking designs. However, to the best of our knowledge, an open-source, publicly available network simulator that integrates mmWaves and V2V scenarios is not currently available: simulation tools for mmWaves, indeed, only support fixed infrastructure scenarios [46, 47], while simulators for ad-hoc communications (e.g., Device-to-Device (D2D)) only model sub-6 GHz frequencies [212].

In this section, we introduce MilliCar, an open-source ns-3 module for V2V mmWave networks*. The module introduces a characterization of sidelink PHY and MAC layers that follow the 3GPP numerologies for NR V2X [199], and enables the study and development of beamforming, link adaptation and medium access techniques for mmWave V2V in end-to-end, full-stack simulations. Additionally, the module features the 3GPP channel model introduced in [66], which has been designed for vehicular simulations in urban and highway scenarios. MilliCar integrates the LTE Service Access Point (SAP) to connect the MAC layer to RLC and PDCP, and implements a new ns-3 `NetDevice` (i.e., `MmWaveVehicularNetDevice`) to take care of the integration with the TCP/IP stack of ns-3. Finally, the module incorporates (i) a helper, that can be used to easily set up simulations; (ii) unit tests, to guarantee that the module behaves as expected even when adding new features; and (iii) several examples, to simulate scenarios with a varying number of vehicles and different deployments.

MilliCar was developed as a standalone module (e.g., with respect to the other ns-3 modules that support mmWaves [46, 47]), to separate the sidelink implementation from that of scheduled cellular protocol stacks. One of the design goals of this module, indeed, is a lean implementation, and the possibility to extend it with new features without having to deal with the complexity of protocol stacks that have not been designed from the start to support a sidelink. Nonetheless, the MilliCar module can still rely on the higher layers from the LTE module through SAPs, to run end-to-end simulations with an NR-V2X-like protocol stack. We

*Available at https://github.com/signetlabdei/millicar

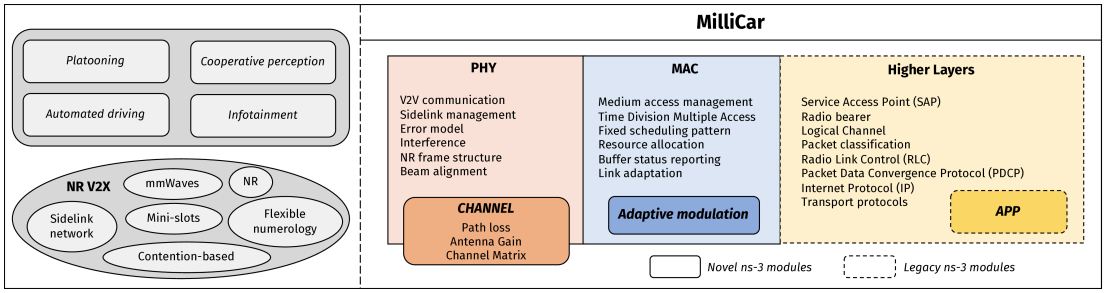


Fig. 6.2: General Overview of the MilliCar Module, with the Features Implemented at each Layer of the Stack

believe that this constitutes a good tradeoff between integration with ns-3 and flexibility to develop new components.

Our simulation module aims at providing researchers with the proper tool to evaluate the effectiveness of novel solutions and overcome the challenges related to V2V communications at mmWaves, such as those presented in Section 6.3. The possible use cases include the perfomance evaluation of channel access and beamforming schemes, resource scheduling algorithms, and solutions for multi-hop forwarding and routing.

6.4.1 An ns-3 Module for NR V2X

In the following sections, we describe the main characteristics of the MilliCar module for V2V networking. The general features of each component of the module are depicted in Figure 6.2, while Figure 6.3 provides a simplified UML diagram.

Notice that MilliCar is integrated with ns-3, and reuses a number of data structures and classes (for example - to hold the configuration parameters of the frame structure) from the NYU/UNIPD `mmwave` module for cellular communications [46].

Vehicular Channel Model Implementation

The accurate characterization of the channel behavior is paramount to obtain reliable simulation results. Therefore, MilliCar implements the propagation and fading models that the 3GPP suggests for V2V communications at mmWaves [213,

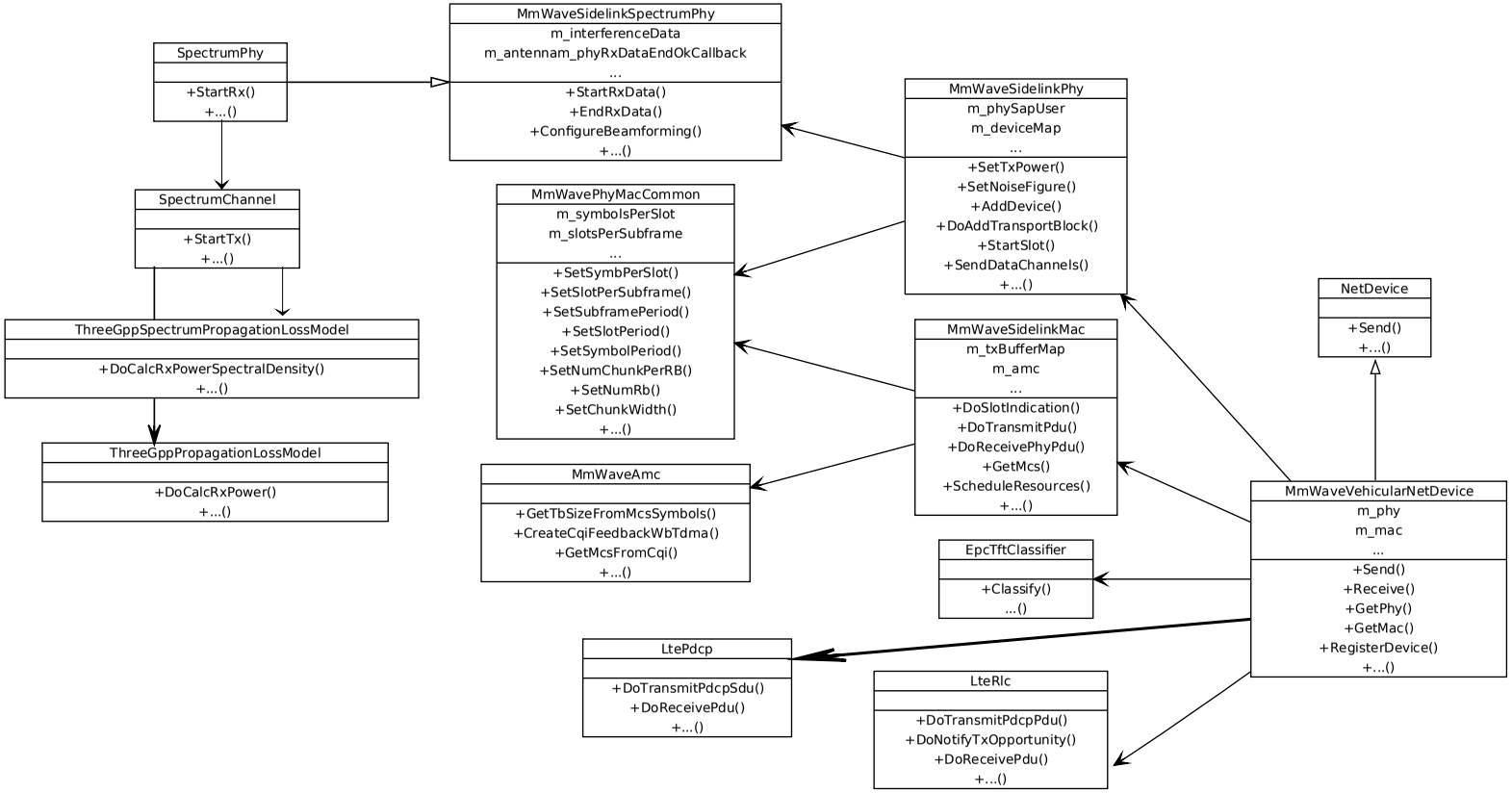


Fig. 6.3: Simplified UML Diagram of the MilliCar Classes

66], by means of the APIs provided by the ns-3 `spectrum` and `propagation` module and described in Section 2.2.2.

The model supports two different scenarios *V2V-Urban*, which emulates signal propagation among vehicles in urban environments, and *V2V-Highway*, which models highway scenarios. In particular, devices communicating through the same wireless channel are attached to a single instance of `SpectrumChannel` that accounts for the modeling of the propagation phenomena using the interfaces `PropagationLossModel` and `SpectrumPropagationLossModel`. The channel condition model is implemented by the classes `ThreeGppV2vUrbanChannelConditionModel` and `ThreeGppV2vHighwayChannelConditionModel`, and supports three different channel states, i.e., LOS, NLOS, and NLOSv. Moreover, the pathloss model is implemented by the classes `ThreeGppV2vUrbanPropagationLossModel` and `ThreeGppV2vHighwayPropagationLossModel`, which extend the `PropagationLossModel` and interact with the channel condition model to retrieve the channel state. Finally, the fast fading model is implemented by the class `ThreeGppChannelModel`, while the computation of the overall channel gain is handled by `ThreeGppSpectrumPropagationLossModel` which extends the `SpectrumPropagationLossModel` interface. Readers interested in a detailed description of the channel model may refer to Chapter 2.

Physical Layer

The MilliCar physical layer is composed of two classes, namely, `MmWaveSidelinkSpectrumPhy` and `MmWaveSidelinkPhy`. `MmWaveSidelinkSpectrumPhy` extends the abstract class `SpectrumPhy` and acts as an interface between `MmWaveVehicularNetDevice` and `SpectrumChannel`. In fact, it handles the transmission and reception operations through the methods `StartTxDataFrames` and `StartRx`. The method `StartTxDataFrames` generates the signal to be transmitted over the channel, represented by the structure `MmWaveSidelinkSpectrumSignalParameters`. Then, it forwards it to the `SpectrumChannel` instance by calling the method `SpectrumChannel::StartTx`. Conversely, when a signal is received from `SpectrumChannel`, the method `StartRx` checks whether or not it can be decoded by applying an error model and, if so, forwards it to the upper layer. The error model that is currently supported by our module is based on the one described in [46], which

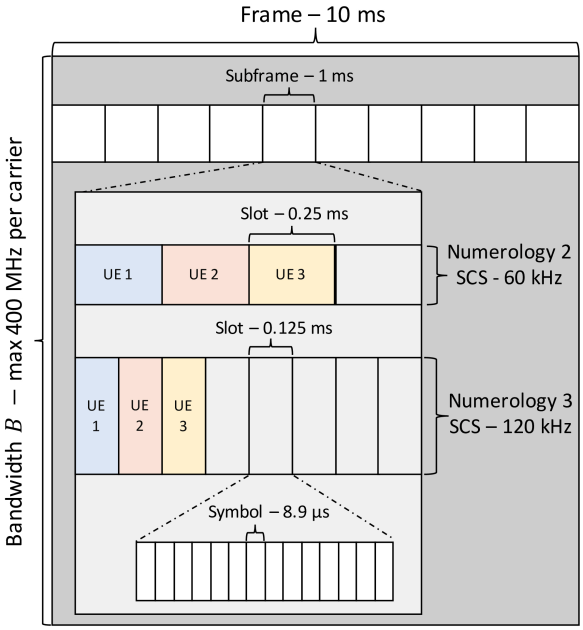


Fig. 6.4: Frame Structures Supported by MilliCar

derives the error probability taking as input the received SINR and the MCS used to encode the signal. To compute the SINR, `MmWaveSidelinkSpectrumPhy` relies on the classes `mmWaveInterference` and `mmWaveChunkProcessor`, of the `mmwave` module [46]. Moreover, `MmWaveSidelinkSpectrumPhy` takes care of the periodic generation of the CSI reported to the upper layers.

The class `MmWaveSidelinkPhy` is in charge of maintaining the system synchronization, and of managing the physical channel used for the transmission and reception of the transport blocks (our module currently supports the modeling of Physical Sidelink Shared Channel (PSSCH) only). The frame structure used by MilliCar is compliant with NR specifications, i.e., a frame of 10 ms is divided in 10 subframes, each containing a variable number of slots. Each slot is composed of 14 OFDM symbols, whose duration depends on the selected numerology configuration [16]. Following the proposal in [199], our module currently supports NR numerologies 2 and 3, i.e., with 4 and 8 slots per subframe, respectively, corresponding to a Subcarrier Spacing (SCS) of 60 kHz or 120 kHz. Figure 6.4 shows the implemented frame structure, where different colors represent different possi-

ble allocation patterns. A transmission buffer is used to store the transport blocks to be sent during the first available slot, together with information regarding the MCS to use and the allocated OFDM symbols.

The method `StartSlot` marks the beginning of each slot and takes care of transmitting the transport blocks stored in the buffer, by scheduling multiple calls to the method `StartTxDataFrames` of `MmWaveSidelinkSpectrumPhy`. Moreover, `MmWaveSidelinkPhy` takes care of forwarding the received transport blocks to the upper layer and managing the beamforming operations to properly point the beam towards the other end device (at this stage, perfect beam alignment is assumed, and further refinements are left for future work).

MAC Layer

The MAC layer functionalities are implemented in the `MmWaveSidelinkMac` class, which includes: (i) the management of the medium access, (ii) the scheduling of the available resources, (iii) the support of transmission and reception over multiple logical channels, and (iv) the link adaptation.

The 3GPP study item for NR V2X [199] considers both in-coverage (mode 1) and out-of-coverage (mode 2) options for resource allocation. MilliCar natively supports mode 2, which is the more likely to be implemented in an early deployment of NR V2X, given that mode 1 would require an update of base stations following standard specifications [214, 199]. Also, mode 2 is of particular interest for researchers since it poses several challenges that remain to be addressed. In particular, `MmWaveSidelinkMac` implements a TDMA-based access scheme, where different vehicles transmit in different slots, as generally assumed for directional mmWave operations [46]. Similarly to mode 2c defined in [199], the MAC layer is pre-configured with a fixed scheduling pattern, which determines how the slots are assigned to the vehicles on a per-subframe basis. By default, each vehicle can use a single slot per subframe, but this pattern can be customized using the `SetSfAllocationInfo` method.

At the beginning of each slot, the MAC layer retrieves the scheduling pattern and executes `DoSlotIndication` to decide whether to perform the transmit or receive operation. In case the slot is intended for transmission, the method `ScheduleResources` divides the available resources among the active logical chan-

nels[†] and notifies the scheduling decision to the upper layers. Then, it builds the transport block using the Service Data Units (SDUs) received from the higher layers, which is then forwarded to the PHY layer by calling the method `AddTransportBlock`. To avoid the allocation of unnecessary resources, the buffers at the upper layers are monitored through a periodic buffer status reporting procedure. Such reports are used to decide the amount of resources to be reserved for each logical channel. Conversely, if the slot is dedicated to another device, the PHY layer is informed about a possible incoming reception by the MAC, which then performs de-multiplexing operations to map the received packets onto the proper logical channels.

Moreover, automatic link adaptation functionalities are provided based on CSI reports received from the PHY. This mechanism is handled by the `MmWaveAmc` class, which uses the last received CSI report to determine the optimal modulation and coding scheme to be used for the transmission.

Integration with the Higher Layers

MilliCar also provides full integration with the higher layers of the protocol stack ensuring, by means of SAP, high flexibility for future improvements of the stack design. We attach to each `MmWaveSidelinkMac` object multiple instances of `LteRlc`, each connected to an `LtePdcp` object. Closing the gap, a specific class implementing the SAP is used to connect the PDCP object to our ad hoc `MmWaveVehicularNetDevice`, envisioning in this way a full bottom-up and top-down integration. The instances of these layers for each end-to-end connection are managed inside the `MmWaveVehicularNetDevice` class, which extends `NetDevice` and implements all the virtual classes commonly used to set up the communication to and from the TCP/IP stack.

In order for two nodes to communicate, a radio bearer must be set up. Once a `MmWaveVehicularNetDevice` is associated to each node, the method `MmWaveVehicularNetDevice::ActivateBearer` is executed on both communication endpoints. This function accepts as input an integer number representing the `bearerID`, the RNTI of the destination (an integer number that differentiates distinct nodes in the network) and the IP address of the pairing node. In particular, each

[†]A logical channel represents an end-to-end connection at the MAC and physical layers.

`bearerID` must unequivocally identify a radio bearer, and cannot be shared among different pairs of devices. The consistency of this assignment (along with that of the RNTI) among different nodes is guaranteed in the helper's configuration method, which will be described in Section 6.4.1. At the MAC layer, a bearer is mapped to a logical channel identifier, as defined by the 3GPP standard. However, at this stage of development, logical channels and radio bearers have a one-to-one correspondence.

The operations carried out by the `ActivateBearer` method are:

- the creation of a rule to classify packets generated from different sources, using `EpcTft::PacketFilter`;
- the instantiation of an `LteRlc` object, which can be identified by the RNTI of the destination node and the logical channel identifier. The RLC object is then linked to the MAC layer instance associated to the node;
- the creation of an `LtePdcp` object, which has to be connected to the `MmWaveVehicularNetDevice` and the RLC object created in the previous step.

After these steps, the RLC and PDCP objects are stored in a dedicated structure, i.e., `SidelinkRadioBearerInfo`, which is then identified with the univocal `bearerID` and saved in the `m_bearerToInfoMap` variable. Currently, the version of the RLC supported by this module is `LteRlcUm`, which provides segmentation and concatenation but no retransmissions.

Once `MmWaveVehicularNetDevice::Send` receives a packet from the IP layer, it accesses the `m_tftClassifier` variable to retrieve the `bearerID` that associates the RNTI and the logical channel identifier to the packet, and stores them in the `TransmitPdcpSduParameters` struct of `LtePdcpSapProvider`. This is then forwarded to the PDCP. Conversely, in the reception phase, a packet is simply sent from PDCP to the `NetDevice`, and from the `NetDevice` to the upper layers.

Helpers and Test Framework

The mmWave vehicular module is also equipped with *helpers* to allow the users to easily set up the simulation, and *unit tests* to check basic functionalities of the module and facilitate future class developments.

HELPERS The main helper is `MmWaveVehicularHelper` which (i) creates and configures the objects for the channel computation; (ii) computes the parameters for the frame structure, according to the selected 3GPP numerology; (iii) installs the networking stack on the vehicles; and (iv) connects groups of vehicles that will communicate together. The first operation is performed during initialization, and relies on three `StringValue` attributes that configure the propagation loss model (`PropagationLossModel`), the fading model (`SpectrumPropagationLossModel`), and the propagation delay model (`PropagationDelayModel`). A typical configuration would include the propagation and fading classes described in Section 6.4.1, without a delay model, as this is included in the spectrum model. However, the user can change and select different options (e.g., a simple Friis propagation loss) and combine also a delay model. The `Numerology` attribute, which is linked to the `SetNumerology` method, accepts 2 or 3 as integer value, to select among the two different numerologies currently foreseen for NR V2X.

The method `InstallMmWaveVehicularNetDevices` accepts a container of `Node` objects, and returns the `NetDeviceContainer` with the `MmWaveVehicularNetDevice` objects. Additionally, for each vehicle, this method sets up the instances of the PHY and MAC layers, configures the antenna at the vehicle and connects it to the channel.

The `MmWaveVehicularHelper` also configures and connects to each device another helper, called `MmWaveVehicularTracesHelper`, which, at runtime, generates a trace of the SINR and MCS for each transmitted packet.

Finally, `PairDevices` configures a bearer and connects, pair by pair, all the devices in a container of `NetDevices` passed as input argument. This makes it possible to create multiple groups of vehicles, with independent scheduling patterns, that generate interference with concurrent transmissions. The vehicles in the same group are all logically connected, thus packets could (in principle) be exchanged among any pair of nodes.

UNIT TESTS The unit test suite contains four tests. `MmWaveVehicularSidelinkSpectrumPhyTestSuite` has a test case that checks if the SNR computed by `MmWaveSidelinkSpectrumPhy` for a single transmission is in line with the expected SNR (considering isotropic antennas and an ideal channel). Similarly, `MmWaveVehicularInterferenceTestSuite` tests if the interference among two

groups of vehicles is correctly computed. `MmWaveVehicularRateTestCase`, instead, features vehicles equipped with a full protocol stack (with UDP at the transport layer), and tests full buffer transmissions for different values of the MCS (i.e., from 0 to 28), checking if there is a one-to-one correspondence between transmitted and received packets.

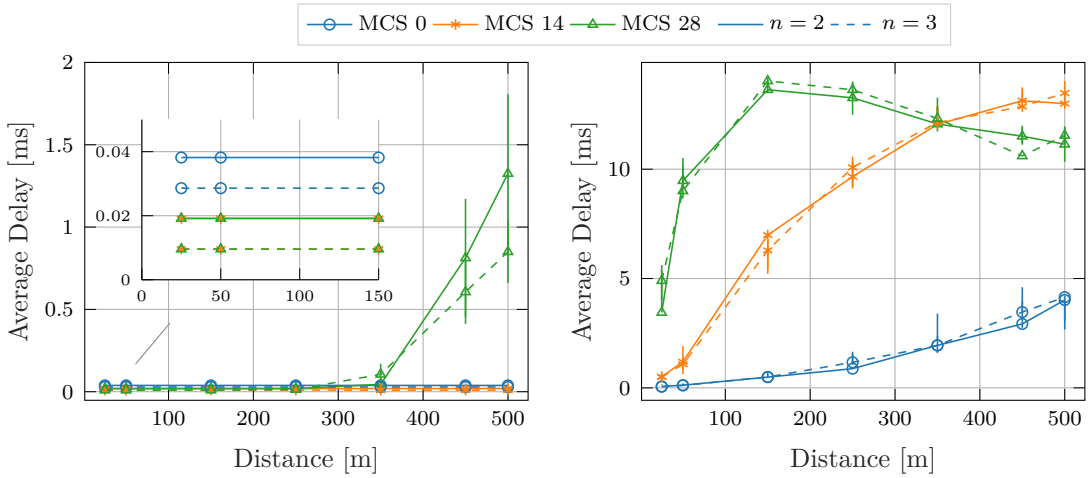
6.5 Performance Evaluation of mmWave V2V Communications

In this section, we target the following two objectives. First, we validate the main functionalities of the MilliCar module through an extensive simulation campaign. Second, we investigate the impact of several system parameters on the end-to-end network performance. More specifically, we examine the effect of the inter-vehicle distance, the propagation scenario, the selected numerology, the MCS, and the RLC reordering timer on two metrics, i.e., the average communication delay and the Packet Reception Ratio (PRR), which are indicators of the robustness of the connection. We believe that our performance analysis will help stimulate more research on the design and evaluation of mmWave vehicular networks, as well as guide standardization decisions towards the most promising architectural configuration(s) for V2V deployments.

We developed two simulation scenarios, i.e., Scenario A, described in Section 6.5.1, and Scenario B, described in Section 6.5.2. In particular, the former has been designed to analyze the impact of different system parameters, namely the MCS, the inter-vehicle distance, the RLC configuration, and the selected numerology, on the end-to-end communication performance, and to compare the system behavior in different propagation scenarios. The latter, instead, considers the presence of multiple groups of vehicles travelling on the same road and sharing

Table 6.2: Simulation parameters.

	Scenario A	**Scenario B**
Distance	$\{25, \ldots, 500\}$ m	40 m
Speed	20 m/s	20 m/s
Propagation scenario	[Urban , Highway]	Highway
Antenna size	4×4	$\{1 \times 1, 2 \times 2, 4 \times 4\}$
Bandwidth	100 MHz	100 MHz
Carrier frequency	28 GHz	28 GHz
Numerology	$\{2, 3\}$	3
MCS	$\{0, 14, 28\}$	$\{0, 28\}$
RLC mode	Unacknowledged	Unacknowledged
RLC reordering timer	$\{1, \ldots, 100\}$ ms	10 ms
RLC buffer size	512 kBytes	512 kBytes
UDP source rate	800 kbps	$\{10, 50, 100\}$ Mbps



(a) LOS propagation.

(b) NLOSv propagation.

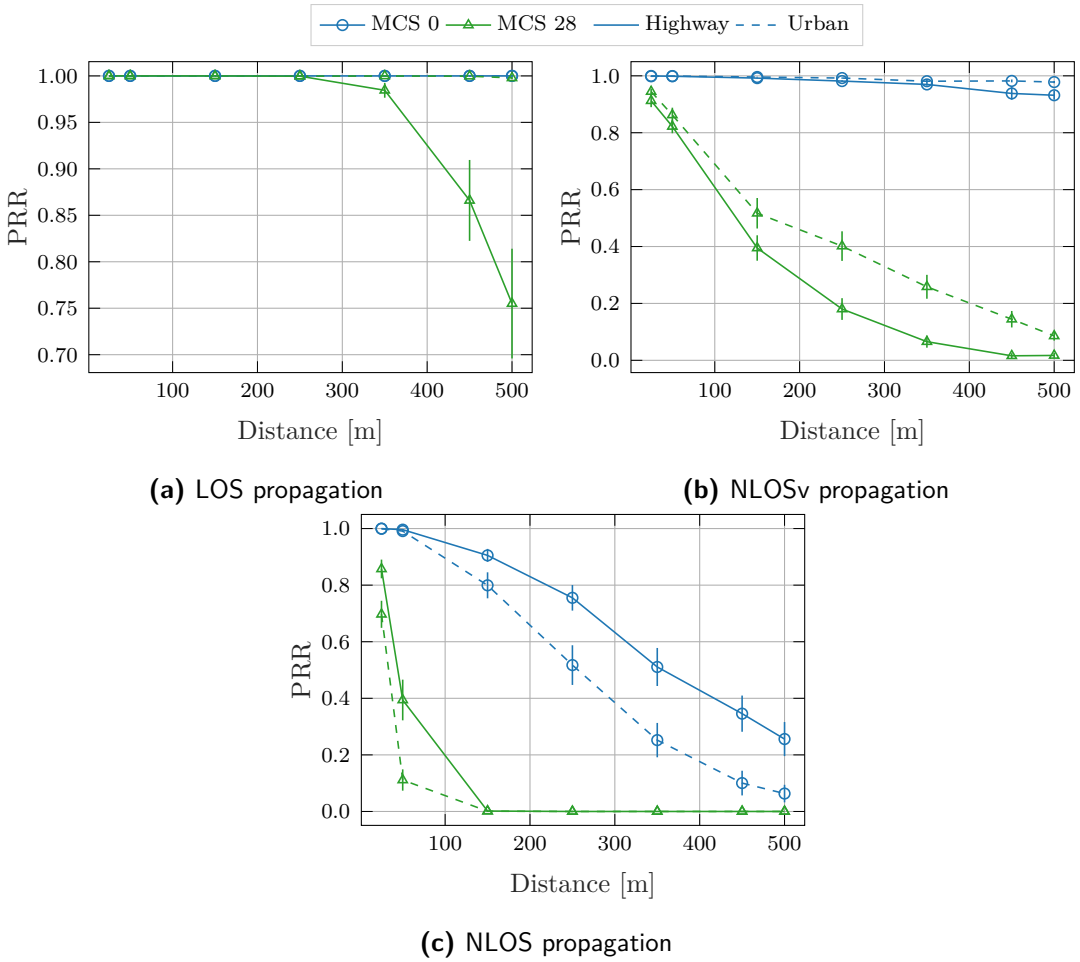
Fig. 6.5: Performance comparison of different numerologies (n) and MCSs, for a highway scenario.

the wireless channel, and evaluates the performance achieved with different modulation and coding schemes and antenna settings. For each scenario, we carried out a simulation campaign and computed some metrics of interest by averaging the results of multiple independent runs. Table 6.2 summarizes the parameters used in our simulations.

6.5.1 Impact of Numerology, MCS, and RLC Parameters

In Scenario A, two vehicles proceed one in front of the other at a constant speed of 20 m/s, keeping the same distance during the whole simulation. One vehicle, i.e., the server, generates packets of 100 Bytes at a fixed rate of 800 kbps, and sends them to the other vehicle, i.e., the client, using a UDP application.

First, we studied the performance of the end-to-end delay and PRR at increasing values of the inter-vehicle distance, focusing on the 3GPP Highway scenario. We assessed these metrics for LOS and NLOSv channel conditions and, for a more detailed insight, we compared the results obtained using different numerologies (i.e., $n = 2$ and $n = 3$) and MCSs. In Figure 6.5a we show that, in the LOS regime, at lower distances numerology 3 guarantees the lowest average delay. This is motivated by the fact that, as described in Section 6.4.1, for this numerology the subframe in divided into 8 slots (compared to 4 slots for numerology 2), resulting in shorter OFDM symbols, to fit the same subframe duration. On the other hand,



(a) LOS propagation

(b) NLOSv propagation

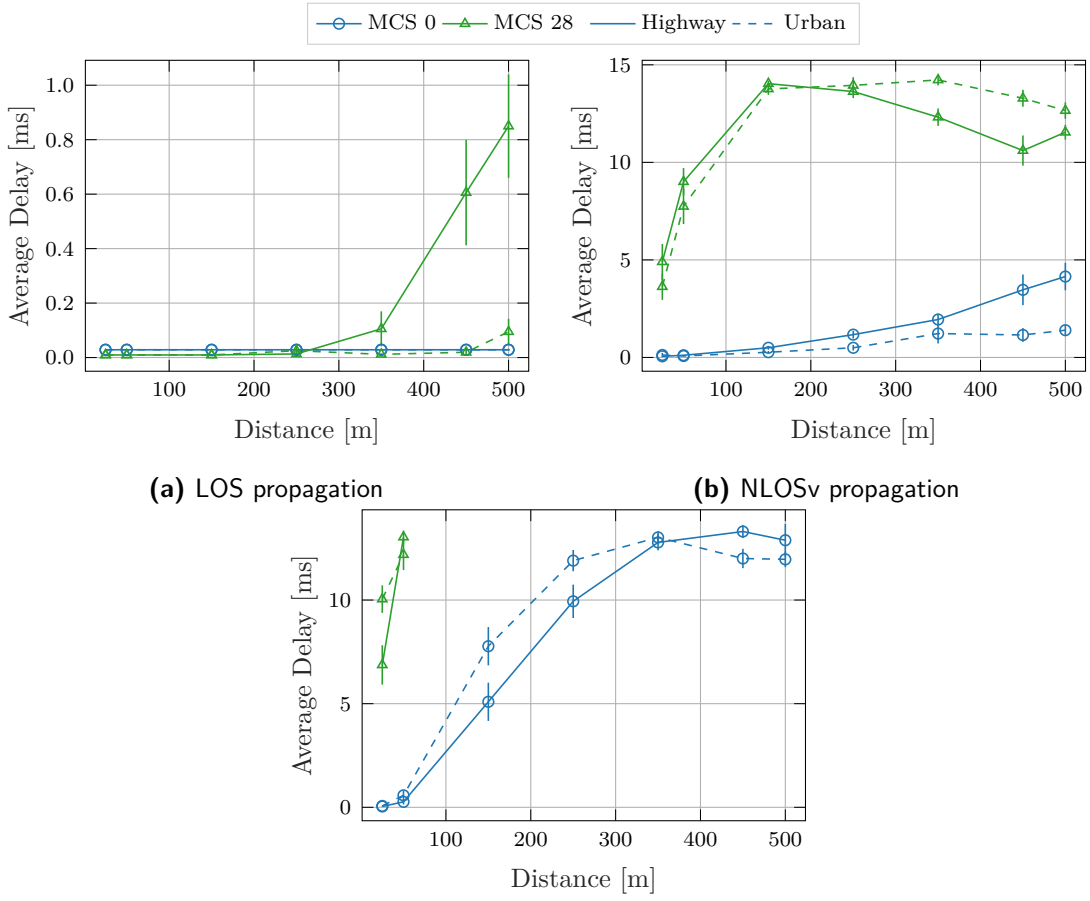
(c) NLOS propagation

Fig. 6.6: PRR for different channel conditions and propagation scenarios, numerology $n = 3$, and packet size 100 Bytes.

as illustrated in Figure 6.5b, it is more difficult to observe a difference between the different numerologies if we consider the NLOSv channel condition, which generally results in a significantly higher end-to-end delay compared to the LOS case. In NLOSv, in fact, some packets may be lost due to a bad channel state. In addition, as the receiving RLC entity implements a reordering procedure for all the received Packet Data Units (PDUs), it has to wait for missing packets in the receiving window until the reordering timer expires: this may increase the packet delay regardless of the numerology that is selected. It should also be mentioned that, when using MCS 28 in LOS, the average delay grows remarkably if we increase the inter-vehicle distance above 250 m as a result of degraded channel conditions, as shown in Figure 6.5a. However, at such long distances, in a real scenario the path would likely be obstructed by other vehicles and, in this case, we expect that the delay will evolve as shown in Figure 6.5b for the NLOSv regime. Moreover, it can be noticed that for MCS 28 in NLOSv, the average delay shows a decreasing behavior when considering distances above 150 m. This is a consequence of the high packet loss rate experienced at such distances, which results in less congested buffers for the remaining packets. As a side note, our results also confirm that better resilience is offered by MCS 0, which guarantees a delay as low as around 1.3 ms, even at 250 m in NLOSv.

In Figures 6.6 and 6.7 we plot the PRR and average delay, respectively, as a function of the inter-vehicle distance and the channel conditions, i.e., urban or highway, for a fixed numerology $n = 3$. In particular, Figures 6.6a and 6.6b exemplify that better end-to-end performance can be obtained using the urban path loss configuration with LOS and NLOSv conditions, hence resulting in a lower latency, as represented by Figures 6.7a and 6.7b. This is motivated by the fact that, in an urban environment, the communication benefits from reflections from walls and/or environmental blockages, which are more likely in street canyons. In this scenario, however, static objects are also more likely to completely block the signal, thus resulting in communication outage. This is demonstrated by Figures 6.6c and 6.7c, where the trend of the curves is switched and urban propagation results in reduced PRR (up to -80%) and increased latency (up to $+50\%$) compared to highway propagation.

Finally, in Figure 6.8 we study how the average end-to-end delay and the PRR are affected by different values of the RLC reordering timer. In particular, we can

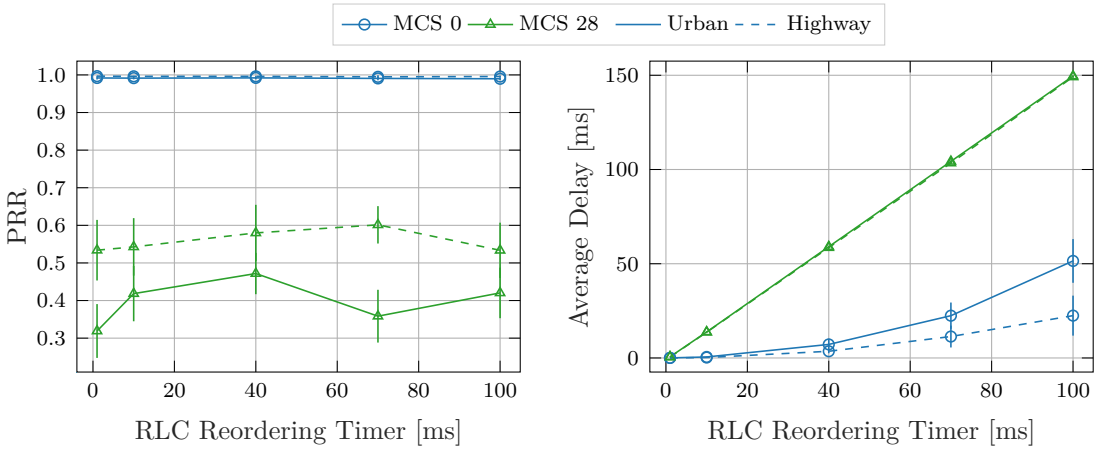


(a) LOS propagation

(b) NLOSv propagation

(c) NLOS propagation

Fig. 6.7: Average delay for different channel conditions and propagation scenarios, numerology $n = 3$, and packet size 100 Bytes.



(a) PRR **(b)** Average delay

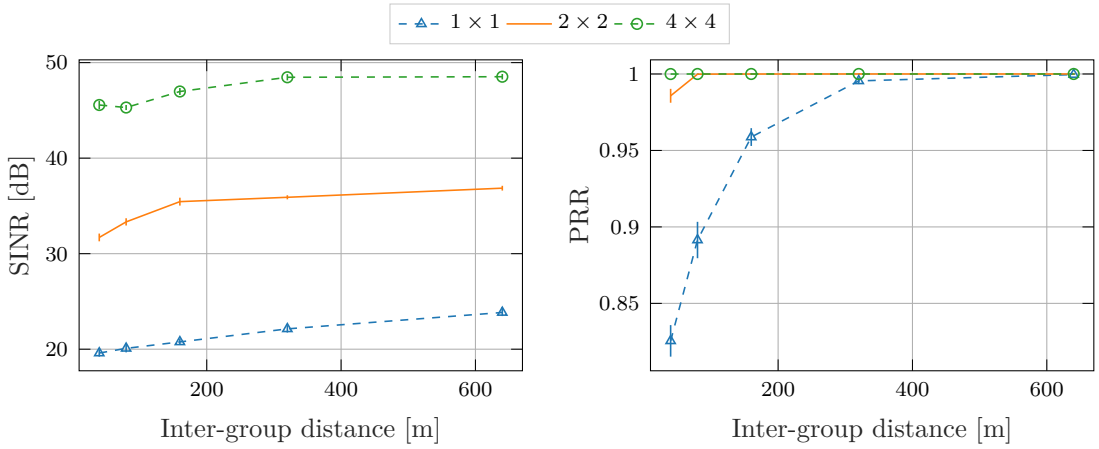
Fig. 6.8: Performance comparison as a function of the RLC reordering timer, for numerology $n = 3$, NLOSv channel condition, and inter-vehicle distance equal to 150 m.

see from Figure 6.8a that, for all the modulation and coding schemes considered, higher values of the reordering timer do not significantly affect the reception ratio. Since we are using RLC unacknowledged mode and we are not implementing any HARQ techniques at the MAC layer, lost packets are not retransmitted[‡], therefore, if there are some missing packets in the receiving window, the reordering timer associated to each packet has to expire before they can be forwarded to the upper layers, which results in an increased experienced delay, as shown in Figure 6.8b.

6.5.2 Impact of Interference and Resource Allocation

In Scenario B, we considered two groups of vehicles traveling in the same direction on different lanes. Each group is composed of two vehicles, one behind the other, moving at a constant speed of 20 m/s and keeping a safety distance of 40 m. Within a group, the rear vehicle acts as a server and generates data packets which are sent to the front vehicle. We considered an ON-OFF traffic model, in which a UDP source keeps switching between the ON and the OFF states. During the ON state, the source generates packets at a constant rate for 100 ms, while in the OFF state it stays idle for a random amount of time, which follows an

[‡]The PRR is expected to improve with higher values of the reordering timer when retransmissions are used, but this study is left for future work.



(a) Average SINR

(b) Average PRR

Fig. 6.9: Performance comparison as a function of the inter-group distance for different antenna configurations and MCS 14. The UDP source rate is 10 Mbps.

exponential distribution with mean 100 ms. All vehicles operate at 28 GHz with a bandwidth of 100 MHz, possibly interfering in case of concurrent transmissions, and are equipped with a UPA of $N \times M$ antenna elements to establish directional communications.

In this context, we evaluate the impact of interference on the communication performance by considering different system configurations. In Figure 6.9a, we plot the average SINR experienced as a function of the inter-group distance, i.e., the distance between the two groups of vehicles, and of the antenna size. It can be seen that the SINR increases with the inter-group distance as a consequence of the weaker effect of the interference. The trend is similar for all the antenna configurations, but the SINR curve has a different offset depending on the number of antenna elements that are used. Indeed, Figure 6.9b demonstrates that the average PRR at the application layer increases for larger antenna arrays, which are able to focus the transmitted power on narrower beams, hence achieving a higher directivity that can possibly reduce the interference. Specifically, only the 4×4 configuration is able to provide a reliable data delivery (i.e., PRR $\cong 1$), regardless of the inter-group distance. For all other antenna architectures, perfect reception is guaranteed when the inter-group distance is higher than 80 m and 600 m for the 2×2 and 1×1 configurations, respectively.

The presence of a centralized scheduling mechanism could prevent the occur-

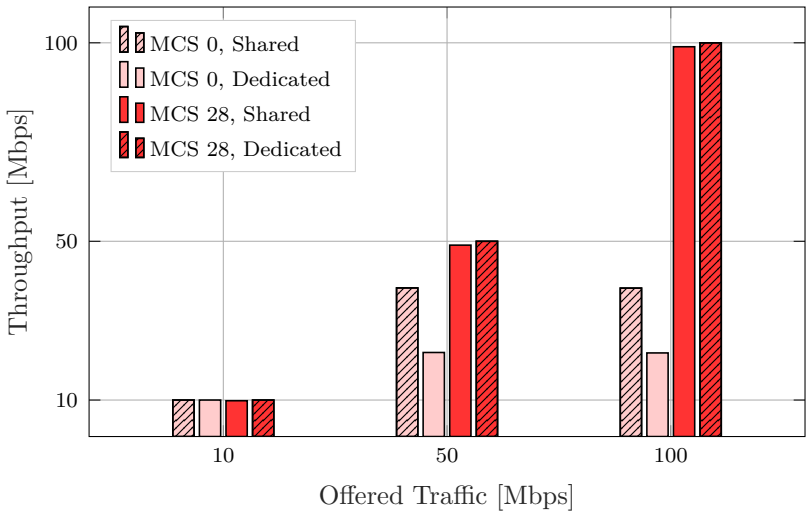


Fig. 6.10: Average throughput as a function of the UDP source rate and the modulation and coding scheme, with or without the orthogonal scheduling of the resources. The antenna configuration is 2×2.

rence of packet collisions by splitting the available resources among the groups in an orthogonal manner. However, reducing the amount of resources accessible to each terminal may limit the achievable throughput. Moreover, the usage of a robust modulation and coding scheme, coupled with a high antenna gain, could provide protection against the interference and enable proper communication performance even without a scheduling mechanism, but at the cost of a lower capacity. We evaluated this trade off by analyzing the average throughput achieved for different modulation and coding schemes, either with or without the orthogonal split of the radio resources among the groups, as reported in Figure 6.10. With MCS 0, i.e., the most robust modulation and coding scheme, neither strategy is able to always satisfy the offered traffic. However, in case of shared resources, the system provides a higher throughput thanks to the larger amount of available resources. Instead, for MCS 28 both strategies offer enough resources to accommodate the offered traffic. We notice that the use of directional antennas mitigates the effect of interference and guarantees high performance even without a coordinated scheduling mechanism.

6.6 Conclusions and Future Work

The wireless networking standardization bodies have started focusing on new market verticals to find new use cases and applications for 5G and beyond. This chapter provided an overview of the ongoing standardization activities for vehicular communications at mmWaves, showing similarities and differences between the IEEE 802.11bd and 3GPP NR V2X specifications. Moreover, we detailed the main challenges related to high-frequency operations considering the whole protocol stack, and introduced MilliCar, the first implementation of an open-source ns-3 module for the simulation of NR-V2X networks at mmWaves. The module enables end-to-end, full-stack simulations of vehicular networks with a 3GPP channel model for V2V propagation and fading at mmWaves, physical and MAC layers redesigned for NR V2X, and integration with the higher layers of the protocol stack from ns-3.

We used MilliCar to evaluate the end-to-end performance of V2V networks operating at mmWaves considering two simulation scenarios in which vehicles operate through directional communications to exchange data packets using a UDP application, and we investigated the impact of several system-level parameters, including the numerology, the MCS, the antenna array size, the RLC reordering timer, the propagation scenario, and the communication distance. Our results demonstrated that mmWave communications can be efficiently exploited in vehicular scenarios but within a limited range. Proper beamforming design could mitigate the effect of interference among groups of communicating vehicles and improve the efficiency by increasing the reuse of the available resources, thereby ensuring higher communication performance, though at the cost of additional complexity (e.g., to align the transmit/receive beams). Also, we proved that, while the RLC reordering timer does not impact much on the PRR, it makes the average end-to-end delay increase significantly, especially for increasing MCSs. Finally, we analyzed the effect of different scheduling options and we concluded that for MSC 0, i.e., the modulation and coding scheme that yields the lowest datarate at the physical layer, the system provides a higher throughput when sharing the available resources, thanks to the robustness of this MCS against errors caused by the interference together with the intrinsic directionality offered by mmWave communications.

As part of our future work, we will integrate new features to the MilliCar module based on the latest proposals in the 3GPP NR V2X standardization process, including a more realistic beam management mechanism and a dedicated medium access control scheme.

7 Conclusions

mmWave communications represent one of the main technological advancements of 5G and beyond cellular systems, having the potential to provide gigabit-per-second data rates. Besides, the large amount of radio resources available in these bands can be leveraged to improve the communication diversity (e.g., by means of multi-connectivity techniques, such as CA) and enable the seamless coexistence of multiple services within the same network (e.g., by means of network slicing). However, communicating at these frequencies involves several intricacies which arise from the hostile propagation conditions experienced by mmWave signals. To fully exploit the advantages that mmWave communications can bring, cellular systems need to revised, including adjustments in the radio protocol stack and adopting new architectural solutions.

In this thesis, we provided novel solutions to overcome the limitations posed by mmWaves and exploit their potential in the context of 5G and beyond cellular networks. After introducing the use cases and requirements for 5G systems, we outlined the main technological enablers and described the main features of the 3GPP 5G NR standard. Then, we focused on mmWave communications and outlined the challenges related to the peculiar propagation conditions experienced at these frequencies.

In Chapter 2, we presented novel simulation models for the accurate evaluation of next-generation wireless systems. In particular, we introduced the first SCM for

ns-3, which supports a wide frequency range, including mmWave spectrum bands, and the modeling of different propagation environments, including cellular and vehicular scenarios. Moreover, we presented a flexible framework for the modeling of antenna arrays and beamforming operations, supporting multiple radiation patterns and array shapes, and different beamforming schemes.

In Chapter 3, we provided an overview of IAB, a novel technology developed by the 3GPP to deploy wireless-backhauled base stations, which represents a valuable solution towards dense network deployments. To identify the advantages and drawbacks of IAB, we carried out a simulation campaign and evaluated the full stack performance of an IAB deployment, considering different settings and traffic patterns. Our results showed that IAB can efficiently relay cell-edge traffic, although the benefits decrease for more congested networks. Moreover, we proposed a novel, semi-centralized, resource management scheme for IAB, which efficiently splits the available resources between the access and backhaul interfaces. Through system level simulations, we demonstrated that our scheme is able to improve the overall network performance with respect to a distributed scheduling approach.

In Chapter 4, we analyzed the integration of MU-MIMO HBF in 5G mmWave networks and discussed the interplay between beamforming operations and the higher layers of the protocol stack. We considered different beam design and scheduling strategies, and evaluated the end-to-end network performance by means of system-level simulations. In particular, we demonstrated that the uncoordinated interaction between the beamformer and the scheduler leads to sub-optimal performance. A joint design of these two entities is therefore fundamental to exploit the capabilities of MU-MIMO HBF mmWave systems.

In Chapter 5, we presented a new RAN slicing framework for 5G and beyond cellular systems operating at mmWave frequencies. This framework is based on CA and exploits the presence of different CCs to support multiple slices simultaneously. We focused on how to serve URLLC and eMBB slices that share the same radio access resources, without compromising the quality of service of the users in either of the two. We evaluated the performance of our solution by means of system level simulations, and compared it against two different baseline policies. The results showed that our framework is able to improve the throughput of the eMBB slice and reduce the delay of the URLLC slice, while preserving isolation

among the traffic flows.

In Chapter 6, we discussed the potential and challenges of mmWave communications to deliver next-generation vehicular services. After reviewing the main standardization activities for vehicular communications, we presented MilliCar, a novel simulation tool for the evaluation of mmWave V2V networks based on the 3GPP NR V2X standard. Using this tool, we carried out an extensive simulation campaign to evaluate the end-to-end performance of mmWave vehicular communications, considering different scenarios and parameters. Our results demonstrated that mmWaves are able to provide high communication performance also in vehicular contexts, thus representing a promising enabler towards connected, cooperative, and intelligent transportation systems.

List of Publications

Publications in International Journals

[J1] M. Pagin, T. Zugno, M. Polese, M. Zorzi, "Resource Management for 5G NR Integrated Access and Backhaul: a Semi-centralized Approach," *IEEE Transaction on Wireless Communications*, Early Access.

[J2] F. Gomez Cuba, T. Zugno, J. Kim, M. Polese, S. Bahk, M. Zorzi, "Hybrid Beamforming in 5G mmWave Networks: a Full-stack Perspective," *IEEE Transaction on Wireless Communications*, Early Access.

[J3] T. Zugno, M. Drago, M. Giordani, M. Polese, M. Zorzi, "Towards Standardization of Millimeter Wave Vehicle-to-Vehicle Networks: Open Challenges and Performance Evaluation," *IEEE Communication Magazine*, vol. 58, no. 9, pp. 79-85, September 2020.

[J4] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, M. Zorzi, "Integrated Access and Backhaul in 5G mmWave Networks: Potentials and Challenges," *IEEE Communication Magazine*, vol. 58, no. 3, pp. 62-68, March 2020.

Publications in Conference and Workshop Proceedings

[C1] T. Zugno, M. Drago, S. Lagen, Z. Ali, M. Zorzi, "Extending the ns-3 Spatial Channel Model for Vehicular Scenarios," in *Proceedings of the Workshop on ns-3 (WNS3 '21)*, Virtual Event, USA, 2021.

[C2] M. Lecci, T. Zugno, S. Zampato, M. Zorzi, "A Full-Stack Open-Source Framework for Antenna and Beamforming Evaluation in mmWave 5G NR," *Proceedings of IEEE International Conference on Communications (ICC)*, Montreal, Canada, 2021.

[C3] F. Gomez Cuba, T. Zugno, J. Kim, M. Polese, S. Bahk, M. Zorzi, "Full-stack Hybrid Beamforming in mmWave 5G Networks," *Proceedings of the 19th IEEE Mediterranean Communication and Computer Networking Conference (MedComNet 2021)*, Virtual Event, 2021.

[C4] T. Zugno, M. Drago, M. Giordani, M. Polese, M. Zorzi, "NR V2X Communications at Millimeter Waves: An End-to-End Performance Evaluation," Proceedings of the IEEE Globecom 2020, Taipei, Taiwan, 2020.

[C5] T. Zugno, F. Campagnaro, M. Zorzi, "Controlling in real-time an ASV-carried ROV for quay wall and ship hull inspection through wireless links in harbor environments," *Proceedings of the IEEE Global OCEANS 2020*, Singapore-U.S. Gulf Coast, 2020.

[C6] M. Pagin, F. Agostini, T. Zugno, M. Polese, M. Zorzi, "Enabling RAN Slicing Through Carrier Aggregation in mmWave Cellular Networks," *Proceedings of the 18th IEEE Mediterranean Communication and Computer Networking Conference (MedComNet 2020)*, Arona, Italy, 2020.

[C7] T. Zugno, M. Polese, N. Patriciello, B. Bojović, S. Lagen and M. Zorzi, "Implementation of A Spatial Channel Model for ns-3," *Proceedings of the Workshop on ns-3 (WNS3 '20)*, Gaithersburg, MD, USA, 2020.

[C8] M. Drago, T. Zugno, M. Polese, M. Giordani, M. Zorzi, "Millicar - An ns-3 Module for MmWave NR V2X Networks," *Proceedings of the Workshop on ns-3 (WNS3 '20)*, Gaithersburg, MD, USA, 2020.

[C9] M. Polese, T. Zugno, M. Zorzi, "Implementation of Reference Public Safety Scenarios in ns-3," *Proceedings of the Workshop on ns-3 (WNS3 '19)*, Firenze, Italy, 2019.

[C10] T. Zugno, M. Polese, M. Lecci, and M. Zorzi. "Simulation of Next-generation Cellular Networks with ns-3: Open Challenges and New Directions," *Proceedings of the 2019 Workshop on Next-Generation Wireless with ns-3 (WNGW 2019)*, Firenze, Italy, 2019.

[C11] T. Zugno, M. Polese, M. Zorzi. "Integration of Carrier Aggregation and Dual Connectivity for the ns-3 mmWave Module," *Proceedings of the 10th Workshop on ns-3 (WNS3 '18)*, Mangalore, India, 2018..

References

[1] Ericsson, "Ericsson Mobility Report," Tech. Rep., June 2021.

[2] ITU-R, "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," Recommendation ITU-R M.2083, Sep. 2015.

[3] A. Aijaz, M. Simsek, M. Dohler, and G. Fettweis, *Shaping 5G for the Tactile Internet*. Springer International Publishing, 2017.

[4] S. K. Rao and R. Prasad, "Impact of 5G Technologies on Industry 4.0," *Wireless Personal Communications*, vol. 100, no. 1, pp. 145–159, Mar. 2018.

[5] International Telecommunication Union (ITU), "Setting the Scene for 5G: Opportunities and Challenges," ITU Tematic Report, 2018.

[6] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[7] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[8] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: the Dominant Theme for Wireless Evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[9] T. L. Marzetta, "Massive MIMO: An Introduction," *Bell Labs Technical Journal*, vol. 20, pp. 11–22, Mar. 2015.

[10] M. Condoluci and T. Mahmoodi, "Softwarization and virtualization in 5G mobile networks: Benefits, trends and challenges," *Computer Networks*, vol. 146, pp. 65–84, Dec. 2018.

[11] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega *et al.*, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Communications magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[12] 3GPP, "NR and NG-RAN Overall Description," TS 38.300 (Rel. 15), 2018.

[13] Parkvall, Stefan and Dahlman, Erik and Furuskar, Anders and Frenne, Mattias, "NR: The New 5G Radio Access Technology," *IEEE Communications Standards Magazine*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[14] 3GPP, "NR; Base Station (BS) radio transmission and reception," Technical Specification (TS) 38.104, Jan. 2021, v17.0.0.

[15] Z. Khan, H. Ahmadi, E. Hossain, M. Coupechoux, L. A. Dasilva, and J. J. Lehtomäki, "Carrier Aggregation/Channel Bonding in Next Generation Cellular Networks: Methods and Challenges," *IEEE Network*, vol. 28, no. 6, pp. 34–40, Nov 2014.

[16] 3GPP, "NR; Physical channels and modulation," TS 38.211 (Rel. 16), 2020.

[17] ——, "Study on New Radio (NR) access technology," TS 38.912 (Rel. 16), 2020.

[18] H. Kim, J. Kim, and D. Hong, "Dynamic TDD Systems for 5G and Beyond: A Survey of Cross-Link Interference Mitigation," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 4, pp. 2315–2348, 2020.

[19] G. Naik, B. Choudhury, and J. Park, "IEEE 802.11bd 5G NR V2X: Evolution of Radio Access Technologies for V2X Communications," *IEEE Access*, vol. 7, pp. 70 169–70 184, 2019.

[20] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity," TS 37.340 (Rel. 15), 2018.

[21] ——, "NR; Integrated Access and Backhaul (IAB) radio transmission and reception," Technical Specification (TS) 38.174, Jun. 2020, v0.1.0.

[22] Balasubramanian, Bharath and Daniels, E. Scott and Hiltunen, Matti and Jana, Rittwik and Joshi, Kaustubh and Sivaraj, Rajarajan and Tran, Tuyen X. and Wang, Chengwei, "RIC: A RAN Intelligent Controller Platform for AI-Enabled Cellular Networks," *IEEE Internet Computing*, vol. 25, no. 2, pp. 7–17, Mar. 2021.

[23] H. T. Friis, "A note on a simple transmission formula," *Proceedings of the IRE*, vol. 34, no. 5, pp. 254–256, 1946.

[24] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, First Quarter 2019.

[25] Y. Banday, G. Mohammad Rather, and G. R. Begh, "Effect of atmospheric absorption on millimetre wave frequencies for 5g cellular networks," *IET Communications*, vol. 13, no. 3, pp. 265–270, Feb. 2019.

[26] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Communications Surveys Tutorials*, vol. 20, no. 2, Dec. 2018.

[27] H. M. Rahim, C. Y. Leow, T. A. Rahman, A. Arsad, and M. A. Malek, "Foliage attenuation measurement at millimeter wave frequencies in tropical vegetation," in *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, 2017, pp. 241–246.

[28] K. Du, O. Ozdemir, F. Erden, and I. Guvenc, "Sub-terahertz and mmwave penetration loss measurements for indoor environments," *arXiv preprint arXiv:2103.02745*, 2021.

[29] A. Schumacher, R. Merz, and A. Burg, "A mmwave bridge concept to solve the cellular outdoor-to-indoor challenge," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–6.

[30] S. Collonge, G. Zaharia, and G. Zein, "Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 2396–2406, Nov. 2004.

[31] G. R. MacCartney, S. Deng, S. Sun, and T. S. Rappaport, "Millimeter-Wave Human Blockage at 73 GHz with a Simple Double Knife-Edge Diffraction Model and Extension for Directional Antennas," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Montreal, QC, Canada, 2016.

[32] U. T. Virk and K. Haneda, "Modeling Human Blockage at 5G Millimeter-Wave Frequencies," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 3, pp. 2256–2266, Mar. 2020.

[33] I. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions and link-budget," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2018.

[34] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter Wave Cellular Networks: A MAC Layer Perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, Jul. 2015.

[35] T. Zugno, M. Polese, M. Lecci, and M. Zorzi, "Simulation of Next-Generation Cellular Networks with ns-3: Open Challenges and New Directions," in *Proceedings of the 2019 Workshop on Next-Generation Wireless with ns-3*, ser. WNGW 2019, Florence, Italy, 2019, pp. 38–41.

[36] T. R. Henderson, M. Lacage, G. F. Riley, C. Dowell, and J. Kopena, "Network Simulations with the ns-3 Simulator," *SIGCOMM demonstration*, vol. 14, no. 14, p. 527, 2008.

[37] Senza Fili Consulting, "The Economics of Small Cells and WiFi Offload," 2012.

[38] SIGNALS Research Group, "Street Light Small Cells - A Revolution in Mobile Operator Network Economics," October 2014.

[39] 3GPP, "NR; Study on integrated access and backhaul; Release 15," TR 38.874, 2018.

[40] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[41] S. Mondal and J. Paramesh, "A reconfigurable 28-/37-GHz MMSE-adaptive hybrid-beamforming receiver for carrier aggregation and multi-standard MIMO communication," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 5, pp. 1391–1406, May 2019.

[42] T. Zugno, M. Polese, and M. Zorzi, "Integration of Carrier Aggregation and Dual Connectivity for the ns-3 mmWave Module," in *Proceedings of the 10th Workshop on ns-3*, ser. WNS3 '18, Surathkal, India, 2018, pp. 45–52.

[43] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-Wave Vehicular Communication to Support Massive Automotive Sensing," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 160–167, December 2016.

[44] M. Polese, F. Restuccia, A. Gosain, J. Jornet, S. Bhardwaj, V. Ariyarathna, S. Mandal, K. Zheng, A. Dhananjay, M. Mezzavilla, J. Buckwalter, M. Rodwell, X. Wang, M. Zorzi, A. Madanayake, and T. Melodia, "MillimeTera: Toward A Large-Scale Open-Source MmWave and Terahertz Experimental Testbed," in *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, ser. mmNets '19. Los Cabos, Mexico: ACM, 2019, pp. 27–32.

[45] S. K. Saha, Y. Ghasempour, M. K. Haider, T. Siddiqui, P. D. Melo, N. Somanchi, L. Zakrajsek, A. Singh, R. Shyamsunder, O. Torres *et al.*, "X60: A programmable testbed for wideband 60 GHz WLANs with phased arrays," *Computer Communications*, vol. 133, pp. 77–88, Jan. 2019.

[46] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-End Simulation of 5G mmWave Networks," *IEEE Com-*

munications Surveys and Tutorials, vol. 20, no. 3, pp. 2237–2263, Third quarter 2018.

[47] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simulation Modelling Practice and Theory*, vol. 96, p. 101933, 2019.

[48] H. Assasa and J. Widmer, "Extending the IEEE 802.11 ad Model: Scheduled Access, Spatial Reuse, Clustering, and Relaying," in *Proceedings of the Workshop on ns-3*. ACM, 2017, pp. 39–46.

[49] H. Assasa, J. Widmer, T. Ropitault, and N. Golmie, "Enhancing the ns-3 IEEE 802.11ad Model Fidelity: Beam Codebooks, Multi-Antenna Beamforming Training, and Quasi-Deterministic MmWave Channel," in *Proceedings of the 2019 Workshop on Ns-3*, ser. WNS3 2019. New York, NY, USA: Association for Computing Machinery, 2019, pp. 33–40.

[50] L. Lanante, S. Roy, S. E. Carpenter, and S. Deronne, "Improved Abstraction for Clear Channel Assessment in ns-3 802.11 WLAN Model," in *Proceedings of the 2019 Workshop on ns-3*, ser. WNS3 2019. Florence, Italy: Association for Computing Machinery, 2019, pp. 49–56.

[51] H. Assasa, J. Widmer, T. Ropitault, A. Bodi, and N. Golmie, "High Fidelity Simulation of IEEE 802.11ad in ns-3 Using a Quasi-Deterministic Channel Model," in *Proceedings of the 2019 Workshop on Next-Generation Wireless with ns-3*, ser. WNGW 2019. New York, NY, USA: Association for Computing Machinery, 2019, pp. 22–25.

[52] M. Rebato, M. Polese, and M. Zorzi, "Multi-Sector and Multi-Panel Performance in 5G mmWave Cellular Networks," in *IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, December 2018, pp. 1–6.

[53] M. Zhang, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "ns-3 Implementation of the 3GPP MIMO Channel Model for Frequency Spectrum above 6 GHz," in *Proceedings of the Workshop on ns-3*, Porto, Portugal, 2017.

[54] N. Baldo and M. Miozzo, "Spectrum-aware Channel and PHY layer modeling for ns3," in *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[55] P. Ferrand, M. Amara, S. Valentin, and M. Guillaud, "Trends and challenges in wireless channel modeling for evolving radio access," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 93–99, July 2016.

[56] M. Polese and M. Zorzi, "Impact of Channel Models on the End-to-End Performance of mmWave Cellular Networks," in *IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018.

[57] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 60–67, Oct 2004.

[58] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, February 1987.

[59] 3GPP, "Study on Channel Model for Frequencies from 0.5 to 100 GHz," TR 38.901 (Rel. 15), 2018.

[60] O. H. Koymen, A. Partyka, S. Subramanian, and J. Li, "Indoor mm-Wave Channel Measurements: Comparative Study of 2.9 GHz and 29 GHz," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[61] K. A. Remley, J. A. Gordon, D. Novotny, A. E. Curtin, C. L. Holloway, M. T. Simons, R. D. Horansky, M. S. Allman, D. Senic, M. Becker, J. A. Jargon, P. D. Hale, D. F. Williams, A. Feldman, J. Cheron, R. Chamberlin, C. Gentile, J. Senic, R. Sun, P. B. Papazian, J. Quimby, M. Mujumdar, and N. Golmie, "Measurement Challenges for 5G and Beyond: An Update from the National Institute of Standards and Technology," *IEEE Microwave Magazine*, vol. 18, no. 5, pp. 41–56, July 2017.

[62] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, Jan 2017.

[63] A. Maltsev, A. Pudeyev, A. Lomayev, and I. Bolotin, "Channel modeling in the next generation mmWave Wi-Fi: IEEE 802.11ay standard," in *22th European Wireless Conference*, May 2016.

[64] M. Lecci, P. Testolina, M. Giordani, M. Polese, T. Ropitault, C. Gentile, N. Varshney, A. Bodi, and M. Zorzi, "Simplified Ray Tracing for the Millimeter Wave Channel: A Performance Evaluation," in *Information Theory and Applications Workshop (ITA)*, 2020.

[65] 3GPP, "Study on Channel Model for Frequency Spectrum Above 6 GHz," TR 38.900 (Rel. 14), 2017.

[66] ——, "Study on Evaluation Methodology of New Vehicle-to-Everything (V2X) Use Cases for LTE and NR," TR 37.885, V15.3.0, 2019.

[67] M. Boban, X. Gong, and W. Xu, "Modeling the Evolution of Line-of-Sight Blockage for V2V Channels," in *IEEE 84th Vehicular Technology Conference (VTC-Fall)*, Montreal, QC, Canada, 2016.

[68] M. Lecci, M. Polese, C. Lai, J. Wang, C. Gentile, N. Golmie, and M. Zorzi, "Quasi-Deterministic Channel Model for mmWaves: Mathematical Formalization and Validation," in *IEEE GLOBECOM*, Taipei, Taiwan, Dec. 2020.

[69] M. Lecci, P. Testolina, M. Polese, M. Giordani, and M. Zorzi, "Accuracy vs. Complexity for mmWave Ray-Tracing: A Full Stack Perspective," https://arxiv.org/abs/2007.07125.

[70] M. Drago, T. Zugno, M. Polese, M. Giordani, and M. Zorzi, "MilliCar: An ns-3 Module for MmWave NR V2X Networks," in *Proceedings of the 2020 Workshop on ns-3 (WNS3 2020)*, Gaithersburg, MD, USA, Jun. 2020.

[71] S. Lagen, K. Wanuga, H. Elkotby, S. Goyal, N. Patriciello, and L. Giupponi, "New Radio Physical Layer Abstraction for System-Level Simulations of 5G

Networks," in *Proceedings of IEEE International Conference on Communications (to be published)*, ser. IEEE ICC, Dublin (Ireland), June 2020.

[72] B. Bojovic, L. Giupponi, Z. Ali, and M. Miozzo, "Evaluating unlicensed LTE technologies: LAA vs LTE-U," *IEEE Access*, vol. 7, pp. 89 714–89 751, 2019.

[73] Q. Chen, X. Xu, and H. Jiang, "Spatial multiplexing based nr-u and wifi coexistence in unlicensed spectrum," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA, Sept. 2019, pp. 1–5.

[74] R. Maldonado, C. Rosa, and K. I. Pedersen, "Latency and reliability analysis of cellular networks in unlicensed spectrum," *IEEE Access*, vol. 8, pp. 49 412–49 423, 2020.

[75] N. Patriciello, S. Goyal, S. Lagen, L. Giupponi, B. Bojovic, A. Demir, M. Beluri, "NR-U and WiGig Coexistence in 60 GHz Bands," arXiv preprint arXiv:2001.04779, 2019.

[76] S. Lagen, L. Giupponi, S. Goyal, N. Patriciello, B. Bojović, A. Demir, and M. Beluri, "New Radio Beam-Based Access to Unlicensed Spectrum: Design Challenges and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 8–37, First quarter 2020.

[77] K. Z. Ghafoor, M. Guizani, L. Kong, H. S. Maghdid, and K. F. Jasim, "Enabling efficient coexistence of dsrc and c-v2x in vehicular networks," *IEEE Wireless Communications (Early Access)*, pp. 2–8, 2019.

[78] M. Rebato, F. Boccardi, M. Mezzavilla, S. Rangan, and M. Zorzi, "Hybrid spectrum sharing in mmwave cellular networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 2, pp. 155–168, June 2017.

[79] F. Boccardi, H. Shokri-Ghadikolaei, G. Fodor, E. Erkip, C. Fischione, M. Kountouris, P. Popovski, and M. Zorzi, "Spectrum pooling in mmwave networks: Opportunities, challenges, and enablers," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 33–39, November 2016.

[80] P. Nuggehalli, "LTE-WLAN Aggregation [Industry Perspectives]," *IEEE Wireless Communications*, vol. 23, no. 4, pp. 4–6, Aug. 2016.

[81] WBA and NGMN Alliance, "RAN Convergence Paper," version 1.0, Aug. 2019.

[82] C. A. Balanis, *Antenna Theory: Analysis and Design*. Wiley-Interscience, 2005.

[83] M. S. Rabbani and H. Ghafouri-Shiraz, "Size Improvement of Rectangular Microstrip Patch Antenna at Mmwave and Terahertz Frequencies," *Microwave and Optical Technology Letters*, vol. 57, no. 11, pp. 2585–2589, Aug. 2015.

[84] Y. M. Abdelkader, M. M. Hamada, and A. N. Mohieldin, "System Level Co-Simulation Approach for Ultra-Wideband Massive MIMO Beam Forming Phased Array Transmitters," in *31st International Conference on Microelectronics (ICM)*, Cairo, Egypt, Mar. 2019.

[85] Y. J. Cho, G. Suk, B. Kim, D. K. Kim, and C. Chae, "RF Lens-Embedded Antenna Array for mmWave MIMO: Design and Performance," *IEEE Communications Magazine*, vol. 56, no. 7, pp. 42–48, Jul. 2018.

[86] C. Menudier, J. Lintignat, S. Mons, P. Médrel, N. Delhote, E. Ngoya, S. Bila, M. Thévenot, B. Jarry, P. Gamand, J. Sombrin, and D. Baillargeat, "Design and optimization of multielement antennas and RF circuits for beamforming with a reduced number of RF Front-ends," in *IEEE MTT-S International Microwave Workshop Series on 5G Hardware and System Technologies (IMWS-5G)*, Dublin, Ireland, Aug. 2018.

[87] P. Testolina, M. Lecci, M. Polese, M. Giordani, and M. Zorzi, "Scalable and Accurate Modeling of the Millimeter Wave Channel," in *International Conference on Computing, Networking and Communications (ICNC)*, Feb. 2020.

[88] H. S. Dhillon and G. Caire, "Wireless Backhaul Networks: Capacity Bound, Scalability Analysis and Design Guidelines," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6043–6056, Nov 2015.

[89] F. Khan and Z. Pi, "mmWave mobile broadband (MMB): Unleashing the 3–300 GHz spectrum," in *34th IEEE Sarnoff Symposium*, Princeton, NJ, USA, 2011.

[90] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1508–1522, Mar. 2017.

[91] M. N. Islam, S. Subramanian, and A. Sampath, "Integrated Access Backhaul in Millimeter Wave Networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, USA, 2017, pp. 1–6.

[92] M. Polese, M. Giordani, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "End-to-End Simulation of Integrated Access and Backhaul at mmWaves," in *IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Sep. 2018.

[93] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth Partitioning and Downlink Analysis in Millimeter Wave Integrated Access and Backhaul for 5G," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8195–8210, Dec 2018.

[94] M. Polese, M. Giordani, A. Roy, D. Castor, and M. Zorzi, "Distributed Path Selection Strategies for Integrated Access and Backhaul at mmWaves," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2018.

[95] A. Ometov, D. Moltchanov, M. Komarov, S. V. Volvenko, and Y. Koucheryavy, "Packet Level Performance Assessment of mmWave Backhauling Technology for 3GPP NR Systems," *IEEE Access*, vol. 7, pp. 9860–9871, 2019.

[96] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP: Standards and Design Principles," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems (MMSys)*, 2011.

[97] V. Gambiroza, B. Sadeghi, and E. W. Knightly, "End-to-end performance and fairness in multihop wireless backhaul networks," in *Proceedings of the*

10th Annual International Conference on Mobile Computing and Networking (MOBICOM), Philadelphia, PA, USA, 2004, pp. 287–301.

[98] M. N. Islam, N. Abedini, G. Hampel, S. Subramanian, and J. Li, "Investigation of performance in integrated access and backhaul networks," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Honolulu, HI, USA, 2018.

[99] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Max-min rates in self-backhauled millimeter wave cellular networks," *arXiv preprint arXiv:1805.01040*, 2018.

[100] W. Lei, Y. Ye, and M. Xiao, "Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 970–979, Sep. 2020.

[101] M. E. Rasekh, D. Guo, and U. Madhow, "Interference-aware routing and spectrum allocation for millimeter wave backhaul in urban picocells," in *53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, 2015.

[102] M. Bilal, M. Kang, S. C. Shah, and S.-G. Kang, "Time-Slotted Scheduling Schemes for Multi-hop Concurrent Transmission in WPANs with Directional Antenna," *ETRI Journal*, vol. 36, no. 3, pp. 374–384, Jun. 2014.

[103] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," in *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, San Francisco, CA, USA, 2003.

[104] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth partitioning and downlink analysis in millimeter wave integrated access and backhaul for 5G," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8195–8210, Dec. 2018.

[105] C. Saha and H. S. Dhillon, "Millimeter Wave Integrated Access and Backhaul in 5G: Performance Analysis and Design Insights," *IEEE Journal on*

Selected Areas in Communications, vol. 37, no. 12, pp. 2669–2684, Dec. 2019.

[106] R. Singh and P. Kumar, "Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 127–142, Jan. 2019.

[107] B. Ji, C. Joo, and N. Shroff, "Throughput-optimal scheduling in multihop wireless networks without per-flow information," *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 634–647, Apr. 2013.

[108] T. K. Vu, C.-F. Liu, M. Bennis, M. Debbah, and M. Latva-Aho, "Path selection and rate allocation in self-backhauled mmWave networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Spain, 2018.

[109] J. García-Rois, F. Gómez-Cuba, M. R. Akdeniz, F. J. González-Castaño, J. C. Burguillo, S. Rangan, and B. Lorenzo, "On the analysis of scheduling in dynamic duplex multihop mmWave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6028–6042, Nov. 2015.

[110] F. Gomez-Cuba and M. Zorzi, "Optimal link scheduling in millimeter wave multi-hop networks with space division multiple access," in *2016 Information Theory and Applications Workshop (ITA)*, La Jolla, CA, USA, 2016.

[111] F. Gómez-Cuba and M. Zorzi, "Optimal Link Scheduling in Millimeter Wave Multi-hop Networks with MU-MIMO radios." *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1839–1854, Mar. 2020.

[112] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," in *29th IEEE Conference on Decision and Control*, Honolulu, HI, USA, 1990.

[113] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[114] F. P. Kelly, A. K. Maulloo, and D. K. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, Apr. 1998.

[115] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated Access and Backhaul in 5G mmWave Networks: Potential and Challenges," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 62–68, Mar. 2020.

[116] B. Korte and J. Vygen, *Combinatorial Optimization*. Springer Berlin Heidelberg, 2002.

[117] H. N. Gabow, "Data structures for weighted matching and nearest common ancestors with linking," in *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, California, USA, 1990.

[118] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead," *Computer Networks (COMNET)*, vol. 182, Aug. 2020.

[119] 3GPP, "NR; Medium Access Control (MAC) protocol specification," Technical Specification (TS) 38.321, Jul. 2020, v16.1.0.

[120] ——, "CSI feedback for Type I codebook," Huawei, HiSilicon, Technical Document (TDoc) R1-1713763, 08 2017.

[121] ——, "NR; Backhaul Adaptation Protocol (BAP) specification," Technical Specification (TS) 38.340, Mar. 2021, v16.4.0.

[122] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An Open Source Product-Oriented LTE Network Simulator Based on Ns-3," in *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '11, Miami, Florida, USA, 2011.

[123] T. Zugno, M. Polese, N. Patriciello, B. Bojović, S. Lagen, and M. Zorzi, "Implementation of a spatial channel model for ns-3," in *Proceedings of the 2020 Workshop on ns-3*, ser. WNS3 2020, Gaithersburg, MD, USA, 2020.

[124] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Proc. Information Theory and Applications Workshop (ITA)*, San Diego, CA, USA, Feb. 2013.

[125] S. Sun, T. Rappaport, R. Heath, A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Communications Magazine*, vol. 52, no. 12, pp. 110–121, Dec. 2014.

[126] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, Sep. 2019.

[127] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Fundamental capacity of MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 684–702, Nov. 2002.

[128] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Aspects of favorable propagation in massive MIMO," in *22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 76–80.

[129] 3GPP, "NR; Multiplexing and channel coding, v16.1.0," *TS 38.212*, Apr. 2020.

[130] ——, "NR; Physical layer procedures for control, v16.1.0," *TS 38.213*, Apr. 2020.

[131] B. Mondal, V. Sergeev, A. Sengupta, G. Ermolaev, A. Davydov, E. Kwon, S. Han, and A. Papathanassiou, "MU-MIMO and CSI feedback performance of NR/LTE," in *53rd Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, USA, 2019.

[132] S. Han, C. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[133] S. Dutta, C. N. Barati, D. Ramirez, A. Dhananjay, J. F. Buckwalter, and S. Rangan, "A case for digital beamforming at mmWave," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 756–770, Feb. 2020.

[134] D. J. Love and R. W. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Transactions on Information Theory*, vol. 51, no. 8, pp. 2967–2976, July 2005.

[135] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, "Hybrid MMSE precoding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19 167–19 181, 2017.

[136] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, January 2014.

[137] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1432–1443, July 2017.

[138] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6481–6494, July 2015.

[139] C. Li, R. Cai, and D. Liu, "A suboptimal STDMA scheduling for concurrent transmissions in mmWave wireless networks," in *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2014.

[140] Y. Niu, Y. Li, D. Jin, L. Su, and D. Wu, "Blockage robust and efficient scheduling for directional mmwave WPANs," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 2, pp. 728–742, Feb. 2015.

[141] J. Jiang and D. Kong, "Joint user scheduling and MU-MIMO hybrid beamforming algorithm for mmWave FDMA massive MIMO system," *International Journal of Antennas and Propagation*, Nov. 2016.

[142] T. E. Bogale, L. B. Le, and A. Haghighat, "User scheduling for massive MIMO OFDMA systems with hybrid analog-digital beamforming," in *IEEE International Conference on Communications (ICC)*, Jun. 2015, pp. 1757–1762.

[143] V. N. Ha, D. H. N. Nguyen, and J. Frigon, "Joint subchannel allocation and hybrid precoding design for mmWave multi-user OFDMA systems," in *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017, pp. 1–5.

[144] J. Jing, K. Deting, L. Guangyue, W. Junxuan, and Y. Wujun, "User scheduling algorithm for mmWave FDMA Massive MU-MIMO system with hybrid beamforming," in *Proceedings of the 8th International Conference on Signal Processing Systems*, ser. ICSPS 2016. Auckland, New Zealand: Association for Computing Machinery, Nov. 2016.

[145] G. Lee and Y. Sung, "A new approach to user scheduling in massive multi-user MIMO broadcast channels," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1481–1495, Apr. 2018.

[146] E. Kim, J. Kwak, and S. Chong, "Virtual beamforming and user scheduling for sub-array architecture in mmWave networks," *IEEE Communications Letters*, vol. 23, pp. 168–171, Jan. 2019.

[147] J. P. González-Coma and L. Castedo, "Power efficient scheduling and hybrid precoding for time modulated arrays," *IEEE Access*, vol. 8, pp. 21 063–21 076, 2020.

[148] J. Liu and E. S. Bentley, "Hybrid-beamforming-based millimeter-wave cellular network optimization," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2799–2813, Dec. 2019.

[149] M. Cheng, J. B. Wang, J. Cheng, J. Y. Wang, and M. Lin, "Joint scheduling and precoding for mmwave and sub-6GHz dual-mode networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 098–13 111, Nov. 2020.

[150] D. Moltchanov, A. Ometov, P. Kustarev, O. Evsutin, J. Hosek, and Y. Koucheryavy, "Analytical TCP model for millimeter-wave 5G NR systems in dynamic human body blockage environment," *Sensors*, vol. 20, no. 14, p. 3880, Jul. 2020.

[151] S. Choi, J. Song, J. Kim, S. Lim, S. Choi, T. T. Kwon, and S. Bahk, "5G K-SimNet: End-to-end performance evaluation of 5G cellular systems," in *16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Las Vegas, NV, USA, USA, Jan. 2019.

[152] S. Sur, I. Pefkianakis, X. Zhang, and K.-H. Kim, "WiFi-assisted 60 GHz wireless networks," in *Proc. of ACM MobiCom*, Snowbird, Utah, USA, 2017, pp. 28–41.

[153] M. Zhang, M. Polese, M. Mezzavilla, J. Zhu, S. Rangan, S. Panwar, and M. Zorzi, "Will TCP work in mmWave 5G cellular networks?" *IEEE Communications Magazine*, vol. 57, no. 1, pp. 65–71, Jan. 2019.

[154] M. K. Müller, F. Ademaj, T. Dittrich, A. Fastenbauer, B. R. Elbal, A. Nabavi, L. Nagel, S. Schwarz, and M. Rupp, "Flexible multi-node simulation of cellular mobile communications: the vienna 5g system level simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 1–17, 2018.

[155] S. Martiradonna, A. Grassi, G. Piro, and G. Boggia, "Understanding the 5G-air-simulator: A tutorial on design criteria, technical components, and reference use cases," *Computer Networks*, vol. 177, p. 107314, August 2020.

[156] Y. Wang, J. Xu, and L. Jiang, "Challenges of system-level simulations and performance evaluation for 5G wireless networks," *IEEE Access*, vol. 2, pp. 1553–1561, 2014.

[157] S. Cho, S. Chae, M. Rim, and C. G. Kang, "System level simulation for 5G cellular communication systems," in *Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2017, pp. 296–299.

[158] M. Liu, P. Ren, Q. Du, W. Ou, X. Xiong, and G. Li, "Design of system-level simulation platform for 5G networks," in *IEEE/CIC International Conference on Communications in China (ICCC)*, July 2016, pp. 1–6.

[159] M. Han, J. W. Lee, C. G. Kang, and M. J. Rim, "5G K-SimSys: Open/modular/flexible system level simulator for 5G system," in *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Oct 2018, pp. 1–2.

[160] M. Kim, S. Ko, and S. Kim, "Enhancing TCP end-to-end performance in millimeter-wave communications," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–5.

[161] M. Lecci, T. Zugno, S. Zampato, and M. Zorzi, "A full-stack open-source framework for antenna and beamforming evaluation in mmWave 5G NR," in *IEEE International Conference on Communications (ICC)*, 2021.

[162] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved Handover Through Dual Connectivity in 5G mmWave Mobile Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, Sept 2017.

[163] M. R. Akdeniz, Y. Liu, S. Rangan, and E. Erkip, "Millimeter wave picocellular system evaluation for urban deployments," in *IEEE Globecom Workshops*, 2013, pp. 105–110.

[164] M. Akdeniz, Y. Liu, M. Samimi, S. Sun, S. Rangan, T. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[165] S. Akoum, O. El Ayach, and R. W. Heath, "Coverage and capacity in mmWave cellular systems," in *Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2012, pp. 688–692.

[166] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "A comparison of MIMO techniques in downlink millimeter wave cellular networks with hybrid beamforming," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 1952–1967, May 2016.

[167] 3GPP, "NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone," TS 38.102-2 (Rel. 16), 2020.

[168] F. Gómez-Cuba and A. J. Goldsmith, "Compressed sensing channel estimation for OFDM with non-Gaussian multipath gains," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 47–61, 2020.

[169] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive mimo with hybrid precoding over frequency-selective fading channels," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1259–1262, 2016.

[170] M. Mezzavilla, M. Miozzo, M. Rossi, N. Baldo, and M. Zorzi, "A lightweight and accurate link abstraction model for the simulation of LTE networks in ns-3," in *Proceedings of the 15th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '12, 2012, pp. 55–60.

[171] M. Casoni and N. Patriciello, "Next-generation TCP for ns-3 simulator," *Simulation Modelling Practice and Theory*, vol. 66, pp. 81 – 93, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1569190X15300939

[172] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of computation*, vol. 44, no. 170, pp. 463–471, Apr. 1985.

[173] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE 77th Vehicular Technology Conference (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[174] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[175] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44 596–44 612, 2019.

[176] "ns-3 | a discrete-event network simulator for internet systems," https://www.nsnam.org/, accessed: 2020-06-22.

[177] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 7, pp. 2207–2225, Jul. 2016.

[178] F. Gómez-Cuba and F. J. González-Castaño, "Improving third-party relaying for LTE-A: A realistic simulation approach," in *IEEE International Conference on Communications (ICC)*, Sydney, NSW, Australia, Jun. 2014.

[179] A. Frotzscher, U. Wetzker, M. Bauer, M. Rentschler, M. Beyer, S. Elspass, and H. Klessig, "Requirements and current solutions of wireless communication in industrial automation," in *IEEE International Conference on Communications Workshops (ICC)*. IEEE, 2014, pp. 67–72.

[180] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, Fourth quarter 2018.

[181] NGMN Alliance, "Description of network slicing concept," NGMN 5G P1 Requirements & Architecture Work Stream End-to-End Architecture Delivarable, 2016.

[182] K. I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. G. U. Garcia, and Y. Wang, "Carrier Aggregation for LTE-advanced: Functionality and Performance Aspects," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 89–95, June 2011.

[183] R. Zhang, M. Wang, L. X. Cai, Z. Zheng, X. Shen, and L. L. Xie, "LTE-Unlicensed: the Future of Spectrum Aggregation for Cellular Networks," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 150–159, June 2015.

[184] A. Ksentini and N. Nikaein, "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, June 2017.

[185] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '17, Snowbird, Utah, USA, 2017.

[186] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On radio access network slicing from a radio resource management perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, October 2017.

[187] S. D'Oro, F. Restuccia, and T. Melodia, "Toward Operator-to-Waveform 5G Radio Access Network Slicing," *IEEE Communications Magazine*, vol. 58, no. 4, pp. 18–23, April 2020.

[188] J. García-Morales, M. C. Lucas-Estañ, and J. Gozalvez, "Latency-Sensitive 5G RAN Slicing for Industry 4.0," *IEEE Access*, vol. 7, pp. 143 139–143 159, September 2019.

[189] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description," TS 36.300 (Rel. 15), 2018.

[190] K. I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. G. U. Garcia, and Y. Wang, "Carrier aggregation for LTE-advanced: functionality and performance aspects," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 89–95, June 2011.

[191] 3GPP, "Study on New Radio Access Technology - Physical Layer Aspects," TR 38.802 (Rel. 14), 2017.

[192] 3GPP, "Policy and charging control architecture," TS 23.203 (Rel. 16), 2019.

[193] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014.

[194] L. M. Clements and K. M. Kockelman, "Economic effects of automated vehicles," *Transportation Research Record*, vol. 2606, pp. 106–114, Jan. 2017.

[195] M. Giordani, A. Zanella, T. Higuchi, O. Altintas, and M. Zorzi, "Performance Study of LTE and mmWave in Vehicle-to-Network Communications," *IEEE 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, 2018.

[196] M. Giordani, A. Zanella, T. Higuchi, O. Altintas, and M. Zorzi, "On the Feasibility of Integrating mmWave and IEEE 802.11p for V2V Communications," in *IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Aug 2018.

[197] 3GPP, "Service requirements for enhanced V2X scenarios - Rel. 15," TS 22.186, 2018.

[198] IEEE, "802.11 NGV proposed PAR," Study Group on 802.11bd (TGbd) 802.11-18/0861r8, May 2019.

[199] 3GPP, "Study on NR Vehicle-to-Everything (V2X)," TR 38.885, March 2019.

[200] M. Giordani, A. Zanella, and M. Zorzi, "Millimeter wave communication in vehicular networks: Challenges and opportunities," in *6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, May 2017.

[201] H. Wymeersch, G. Seco-Granados, G. Destino, D. Dardari, and F. Tufvesson, "5G mmWave positioning for vehicular networks," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 80–86, Dec 2017.

[202] G. Naik, B. Choudhury, and J. Park, "IEEE 802.11bd 5G NR V2X: Evolution of Radio Access Technologies for V2X Communications," *IEEE Access*, vol. 7, pp. 70 169–70 184, May 2019.

[203] IEEE, "PHY Numerology Discussions," Study Group on 802.11bd (TGbd) 802.11-19/0686, May 2019.

[204] ——, "OCB for 60 GHz V2X," Study Group on 802.11bd (TGbd) 802.11-19/1162, July 2019.

[205] 3GPP, "NR Sidelink enhancement," RP-193231, 2019.

[206] X. You, D. Wang, B. Sheng, X. Gao, X. Zhao, and M. Chen, "Cooperative distributed antenna systems for mobile communications [Coordinated and Distributed MIMO]," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 35–43, June 2010.

[207] P. Kumari, J. Choi, N. González-Prelcic, and R. W. Heath, "IEEE 802.11ad-Based Radar: An Approach to Joint Vehicular Communication-Radar System," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3012–3027, April 2018.

[208] S. H. Ali Shah, S. Aditya, S. Dutta, C. Slezak, and S. Rangan, "Power Efficient Discontinuous Reception in THz and mmWave Wireless Systems," in *IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019.

[209] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, and W. Xu, "Connected roads of the future: Use cases, requirements, and design considerations for vehicle-to-everything communications," *IEEE Vehicular Technology Magazine*, vol. 13, no. 3, pp. 110–123, July 2018.

[210] C. Perfecto, J. Del Ser, and M. Bennis, "Millimeter-Wave V2V Communications: Distributed Association and Beam Alignment," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2148–2162, Sep. 2017.

[211] I. Mavromatis, A. Tassi, R. J. Piechocki, and A. Nix, "Efficient V2V Communication Scheme for 5G mmWave Hyper-Connected CAVs," in *IEEE In-*

ternational Conference on Communications Workshops (ICC Workshops), May 2018, pp. 1–6.

[212] R. Rouil, F. J. Cintrón, A. Ben Mosbah, and S. Gamboa, "Implementation and Validation of an LTE D2D Model for ns-3," in *Proceedings of the Workshop on ns-3*, ser. WNS3 '17. Porto, Portugal: Association for Computing Machinery, 2017, pp. 55–62.

[213] M. Giordani, T. Shimizu, A. Zanella, T. Higuchi, O. Altintas, and M. Zorzi, "Path Loss Models for V2V mmWave Communication: Performance Evaluation and Open Challenges," in *IEEE 2nd Connected and Automated Vehicles Symposium*, September 2019.

[214] T. Zugno, M. Drago, M. Giordani, M. Polese, and M. Zorzi, "Towards Standardization of Millimeter Wave Vehicle-to-Vehicle Networks: Open Challenges and Performance Evaluation," *IEEE Communications Magazine*, vol. 58, no. 9, pp. 79–85, 2020.