5-2023

Enhanced Mobile Networking using Multi-connectivity and Packet Duplication in Next-Generation Cellular Networks

Prabodh Mishra
Clemson University, pmishra@clemson.edu

Enhanced Mobile Networking using Multi-connectivity and Packet Duplication in Next-Generation Cellular Networks

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Computer Engineering

---

by
Prabodh Kumar Mishra
May 2023

Accepted by:
Dr. Kuang-Ching Wang, Committee Chair
Dr. Harlan Russell
Dr. Linke Guo
Dr. Long Cheng
Dr. Jim Martin

Abstract

Modern cellular communication systems need to handle an enormous number of users and large amounts of data, including both users as well as system-oriented data. 5G is the fifth-generation mobile network and a new global wireless standard that follows 4G/LTE networks. The uptake of 5G is expected to be faster than any previous cellular generation, with high expectations of its future impact on the global economy. The next-generation 5G networks are designed to be flexible enough to adapt to modern use cases and be highly modular such that operators would have the flexibility to provide selective features based on user demand that could be implemented without investment in additional infrastructure. Thus, the underlying cellular network that is capable of delivering these expectations must be able to handle high data rates with low latency and ultra-reliability to fulfill these growing needs.

Communication in the sub-6 GHz range cannot provide high throughputs due to the scarcity of spectrum in these bands. Using frequencies in FR2 or millimeter wave (mmWave) range for communication can provide large data rates and cover densely populated areas, but only over short distances as they are susceptible to blockages. This is why dense deployments of mmWave base stations are being considered to achieve very high data rates. But, such architectures lack the reliability needed to support many V2X applications, especially under mobility scenarios. As we have discussed earlier, 5G and beyond 5G networks must also account for UE's mobility as they are expected to maintain their level of performance under different mobility scenarios and perform better than traditional networks. Although 5G technology has developed significantly in recent years, there still exists a critical gap in understanding how all these technologies would perform under mobility. There is a need to analyze and identify issues that arise with mobility and come up with solutions to overcome these hurdles without compromising the performance of these networks. Multi-connectivity (MC) refers to simultaneous connectivity with multiple radio access technologies

or bands and potentially represents an important solution for the ongoing 5G deployments towards improving their performance. To address the network issues that come with mobility and fill that gap, this dissertation investigates the impact of multi-connectivity on next-generation networks from three distinct perspectives, 1) mobility enhancement using multi-connectivity in 5G networks, 2) improving reliability in mobility scenarios using multi-Connectivity with packet duplication, and 3) single grant multiple uplink scheme for performance improvement in mobility scenarios.

The traditional macro-cell architecture of cellular networks that cover large geographical areas will struggle to deliver the dense coverage, low latency, and high bandwidth required by some 5G applications. Thus, 5G networks must utilize ultra-dense deployment of access points operating at higher mmWave frequency bands. But, for such dense networks, user mobility could be particularly challenging as it would reduce network efficiency and user-perceived service quality due to frequent handoffs. Multi-connectivity is seen as a key enabler in improving the performance of these next-generation networks. It enhances the system performance by providing multiple simultaneous links between the user equipment (UE) and the base stations (BS) for data transfer. Also, it eliminates the time needed to deal with frequent handoffs, link establishment, etc. Balancing the trade-offs among handoff rate, service delay, and achievable coverage/data rate in heterogeneous, dense, and diverse 5G cellular networks is, therefore, an open challenge. Hence, in this dissertation, we analyze how mobility impacts the performance of current Ultra-dense mmWave network (UDN) architecture in a city environment and discuss improvements for reducing the impact of mobility to meet 5G specifications using multi-connectivity.

Current handover protocols, by design, suffer from interruption even if they are successful and, at the same time, carry the risk of failures during execution. The next-generation wireless networks, like 5G New Radio, introduce even stricter requirements that cannot be fulfilled with the traditional hard handover concept. Another expectation from these services is extreme reliability that will not tolerate any mobility-related failures. Thus, in this dissertation, we explore a novel technique using packet duplication and evaluate its performance under various mobility scenarios. We study how packet duplication can be used to meet the stringent reliability and latency requirements of modern cellular networks as data packets are duplicated and transmitted concurrently over two independent links. The idea is to generate multiple instances (duplicates) of a packet and transmit them simultaneously over different uncorrelated channels with the aim of reducing the packet failure probability. We also propose enhancements to the packet duplication feature to improve radio resource utilization.

The wide variety of use cases in the 5G greatly differs from the use cases considered during the design of third-generation (3G) and fourth-generation (4G) long-term evolution (LTE) networks. Applications like autonomous driving, IoT applications, live video, etc., are much more uplink intensive as compared to traditional applications. However, the uplink performance is often, by design, lower than the downlink; hence, 5G must improve uplink performance. Hence, to meet the expected performance levels, there is a need to explore flexible network architectures for 5G networks. In this work, we propose a novel uplink scheme where the UE performs only a single transmission on a common channel, and every base station that can receive this signal would accept and process it. In our proposed architecture, a UE is connected to multiple mmWave capable distributed units (DUs), which are connected to a single gNB-central unit. In an ultra-dense deployment with multiple mmWave base stations around the UE, this removes the need to perform frequent handovers and allows high mobility with reduced latency. We develop and evaluate the performance of such a system for high throughput and reliable low latency communication under various mobility scenarios.

To study the impact of mobility on next-generation networks, this work develops and systematically analyzes the performance of the 5G networks under mobility. We also look into the effect of increasing the number of users being served on the network. As a result, these studies are intended to understand better the network requirements for handling mobility and network load with multi-connectivity. This dissertation aims to achieve clarity and also proposes solutions for resolving these real-world network mobility issues.

Dedication

I dedicate this dissertation to my wife, colleague, and best friend, Snigdhaswin Kar, who has been my biggest source of inspiration and encouragement throughout this journey.

I also dedicate this to my parents and loved ones, who have always been supportive of my endeavors and taught me to pursue my dreams and goals. Thank you all for your prayers and blessings.

Acknowledgments

I would like to thank my adviser, Dr. Kuang-Ching Wang, for his constant support, guidance, and encouragement throughout my research work. I am deeply grateful to Dr. Wang for giving me an opportunity and providing invaluable input that helped guide this work.

I would like to thank my committee members Dr. Harlan Russell, Dr. Linke Guo, Dr. Long Cheng, and Dr. Jim Martin, for their advice and support during my research. I am very grateful for their time and help during my research.

Finally, I would like to thank my peers and members of my research group for their support and help during my graduate studies.

This dissertation is based upon work supported by the National Science Foundation under Grants 1419199, 1743363, 2027208, and 1935966.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cellular communication is the most popular way to connect people together for real-time communication and data transmission these days. Modern cellular communication systems handle an enormous number of users and large amounts of data, including both users as well as system-oriented data. Since their initial arrival in the late 1970s, cellular networks, and technology has evolved considerably, with successive generations (2G through 4G) representing significant milestones in the development of mobile connectivity. 5G is the fifth-generation mobile network and a new global wireless standard that follows 4G/LTE networks. While 5G technology is built on components from previous generations of wireless technologies, it expands the focus from typical consumer use cases like phone calls, text messaging, and web browsing to larger-scale industrial and government uses. Thus, the uptake of 5G is expected to be faster than any previous cellular generation, with high expectations of its future impact on the global economy. Analysts forecast there will be 4.4 billion 5G connections, and 5G will account for nearly half of all mobile subscriptions by 2027 [23].

5G networks and beyond have the potential to support and enable new applications in areas like artificial intelligence (AI) and the Internet of Things (IoT) to benefit sectors as wide-ranging as defense, manufacturing, medicine, and education. Thus, the 3rd Generation Partnership Project (3GPP), an independent global standardization committee responsible for defining the technical specifications for wireless standards since the introduction of 3G, specifies three different classes of 5G services: Ultra-Reliable Low Latency Communication (URLLC), enhanced Mobile Broadband (eMBB), and massive Machine Type Communication (mMTC). URLLC applications like remote control of critical infrastructure, vehicles, and medical procedures require high reliability and very low

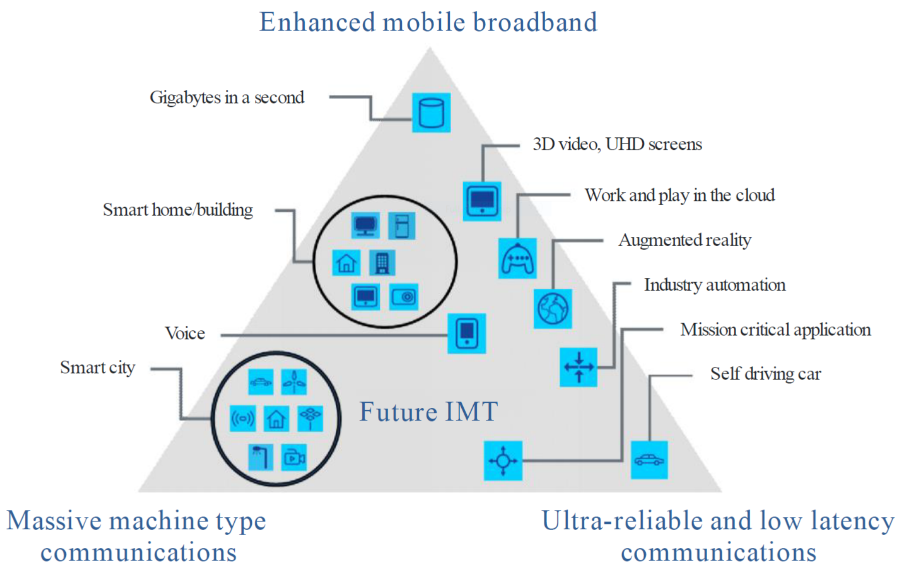


Figure 1.1: Usage scenarios of 5G networks and beyond *Source: Rec. ITU-R M.2083* [35]

latency for error-free faster access to data. eMBB applications require high throughput and capacity to enable new immersive experiences such as VR and AR with faster, more uniform data rates at lower cost-per-bit. And finally, mMTC needs high scalability and is meant to seamlessly connect a massive number of embedded sensors in virtually everything through the ability to scale down in data rates, power, and mobility by providing extremely lean and low-cost connectivity solutions. The next-generation 5G networks are designed to be flexible enough to adapt to new use cases and be highly modular such that operators would have the flexibility to provide selective features based on demand that could be implemented without investment in additional infrastructure. Figure 1.1 illustrates some examples of envisioned usage scenarios for 5G and beyond networks and gives an idea of how the requirements for different use cases may vary.

3GPP is still actively working on new releases, which would comprise more standardized features and specifications developers can use and implement for 5G technology. However, the initial requirements and specifications for 5G were defined in 2015 by The International Telecommunications Union (ITU) [35] and have since been integrated by 3GPP into Releases 15 [3] and 16 [4]. These ambitious specifications represented a step up from 4G performance and aimed to address the requirements of the emerging applications discussed above. 5G specifications defined expected throughputs of up to 20 Gbps, i.e., 20 times faster than 4G networks. Latencies of 1 millisecond as compared to 10 milliseconds for 4G and connection densities of 1 million devices per square kilometer (10 times more than 4G) to support the growing numbers of IoT devices and sensors.

Along with all the above demanding specifications, ITU also proposed that 5G networks should enable high mobility up to 500 km/h with acceptable QoS (Quality of Service), envisioned in particular for high-speed trains. More recently, developments in other areas of technology have resulted in applications being developed that expect seamless connectivity under various types of mobility scenarios. Use cases like automated driving incorporates functions requiring a high degree of real-time and precise coordination between a vehicle and its surrounding environment. 3GPP specifications in TS 22.186 list the system performance requirements to support various enhanced vehicle-to-everything or V2X scenarios [8]. The requirements for latency, reliability and data rate vary widely among the different V2X scenarios and cannot be classified into a single class of service. Thus, the underlying cellular network must be capable of delivering high data rates with low latency and high reliability to fulfill these growing needs.

Such enhanced wireless connectivity will therefore require the allocation of a new spectrum to support the 5G network, as well as more efficient use of the existing spectrum through hardware and software innovations. The 5G New Radio (NR) [14] is a new air interface developed by 3GPP. It has a flexible transmission time interval (TTI) structure with reduced processing times to meet the low-latency constraints and achieve reliability requirements. The NR air interface can operate over two frequency ranges, FR1 (sub-6 GHz) and FR2 (above 24 GHz), each range with specific propagation characteristics. Communication in the FR1 or sub-6 GHz range can be coupled with technologies such as massive multiple-input/multiple-output (MIMO) antennas to deliver reliable, cost-effective, and highly scalable mobile broadband access. Using frequencies in FR2 or millimeter wave (mmWave) range for communication can provide large data rates and cover densely populated areas, but only over short distances as they are susceptible to blockages. This is why dense deployments of mmWave base stations are being considered to achieve very high data rates. But, such architectures lack the reliability needed to support many V2X applications, especially under mobility scenarios.

Considering that the lower frequency bands allow for broad coverage while the higher frequency bands deliver faster speeds and better quality, practical deployment of a 5G network will require the usage of three types of radio frequencies: below 1 GHz, 1-6 GHz, and above 24 GHz (mmWaves). The decision of what bands to be used for different purposes would be influenced not just by the country's spectrum distribution policy but also influenced by international convening bodies like the ITU. Furthermore, 5G network deployments need to combine low, mid, and high-band spectrums to reach both remote areas and ensure good speed of the connection. Thus, an ideal deployment would enable the use of mmWave links as much as possible for high throughput and low latency but also provide low or mid-band links for support to ensure the network is reliable. This is primarily the reason why we need to redefine the network architecture design for 5G as compared to traditional wireless networks.

As we have discussed earlier, 5G and beyond 5G networks must also account for UE's mobility as they are expected to maintain their level of performance under different mobility scenarios and perform better than traditional networks. Although 5G technology has developed significantly in recent years, there still exists a critical gap in understanding how all these technologies would perform under mobility. There is a need to analyze and identify issues that arise with mobility and come up with solutions to overcome these hurdles without compromising the performance of these networks. Multi-connectivity (MC) refers to simultaneous connectivity with multiple radio access technologies or bands and potentially represents an important solution for the ongoing 5G deployments towards improving their performance. To address the network issues that come with mobility and fill that gap, this dissertation investigates next-generation networks from three distinct perspectives as follows:

1. Mobility enhancement using Multi-Connectivity in 5G networks.
2. Improving reliability in mobility scenarios using Multi-Connectivity with Packet Duplication.
3. Single grant multiple uplink scheme for performance improvement in mobility scenarios.

First, this dissertation analyses the issue of mobility in 5G networks from the perspective of network architecture and proposes changes like multi-connectivity needed in the current architecture. Second, we propose enhancement using packet duplication to improve the performance of the multi-connectivity architecture to support enhanced mobility. Finally, the third topic looks into a novel architecture focusing on a single grant-based multiple uplink transmission and its performance under mobility. Further, the motivation behind this work is to better understand network behavior for the specialized use case where the requirement is to achieve low latency and high reliability, similar to URLLC use cases, along with achieving high throughputs. This type of use case is becoming more

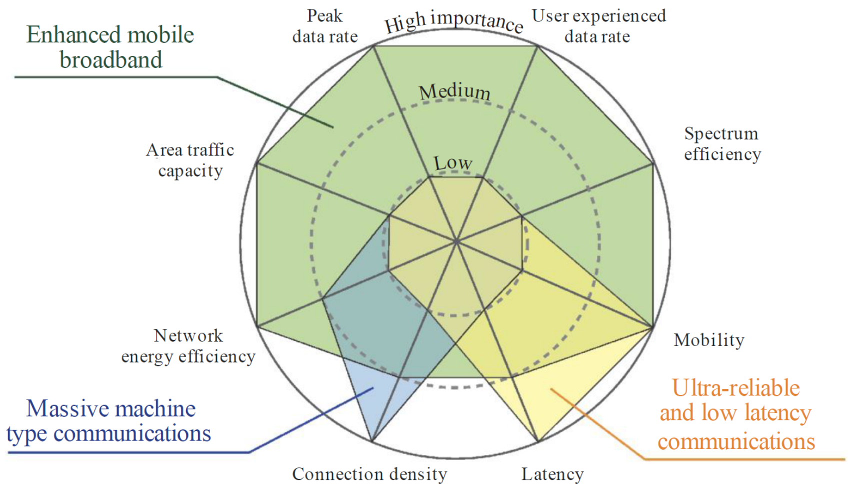


Figure 1.2: Importance of key 5G capabilities in different usage scenarios *Source: Recommendation ITU-R M.2083* [35]

and more common with modern applications. As a result, this dissertation focuses on evaluating network performance in mobility scenarios from two aspects. The first aspect looks at the UE performance in a city-like environment, and the second looks at how the network scales when the load is increased by analyzing the effects of multiple users on these deployments' performance. Thus, the experiments in each of the above three topics are designed accordingly. To better understand the effects of mobility, this dissertation also looks into the number of handovers and their effects on each deployment.

1.1 Mobility Enhancement using Multi-Connectivity in 5G Networks

5G networks must support a large user base and serve a diverse set of applications with very different use cases. Figure 1.2 uses a three-step scale of "high," "medium," and "low" to show the importance of each key network capability for the three 5G services classes, i.e., eMBB, URLLC, and mMTC as defined by 3GPP. However, the previous generation's mobile network architecture was designed to meet requirements for voice and conventional mobile broadband services only. This network architecture has proven to be insufficiently flexible to support diversified 5G services as described in multiple 3GPP version upgrades due to various network elements and complex interfaces.

The traditional macro-cell architecture of cellular networks that cover large geographical areas will struggle to deliver the dense coverage, low latency, and high bandwidth required by some 5G applications. Thus, 5G networks must utilize ultra-dense deployment of access points operating at higher mmWave frequency bands. But, for such dense networks, user mobility could be particularly challenging as it would reduce network efficiency and user-perceived service quality due to frequent handoffs. Multi-connectivity (MC) is seen as a key enabler in improving the performance of these next-generation networks [53]. Simultaneous connectivity to multiple radio access technologies or multiple bands represents an important solution for ongoing 5G deployments. It enhances the system performance by providing multiple simultaneous links between the user equipment (UE) and the base stations (BS) for data transfer. Also, it eliminates the time needed to deal with frequent handoffs, link establishment, etc.

User mobility directly impacts the design and performance of cellular wireless networks in terms of resource management aspects like channel allocation, traffic load estimation, etc. radio propagation aspects like signal strength variation, interference level, etc., and location management aspects such as location area planning. The mobility protocols ensure the delivery of useful data and control signals via a radio interface link to the desired user when commuting between coverage areas. However, providing seamless mobility to all devices in a diverse, dense, multi-tier network like 5G with traditional protocols would expectedly have repercussions on network performance. The 5G and next-generation cellular networks would require flexibility in the network topology to ensure the level of performance that is expected of such services. Balancing the trade-offs among handoff rate, service delay, and achievable coverage/data rate in heterogeneous, dense, and diverse 5G cellular networks is, therefore, an open challenge. Hence, this dissertation in this section analyzes how mobility impacts the performance of current Ultra-dense mmWave network (UDN) architecture in a city environment and discusses improvements for reducing the impact of mobility to meet 5G specifications using multi-connectivity.

1.2 Improving Reliability in Mobility Scenarios using Multi-Connectivity with Packet Duplication

As discussed, millimeter wave technology is a key enabler for 5G use cases; however, mmWave links have a lower transmission range and are subject to blockages. Deploying small cells along with overlapping macro base stations is a potential solution for improving latency and bandwidth. But, such an architecture leads to low cell resource utilization and adds to infrastructure deployment costs due to imbalanced network load [52]. As suggested in the previous section, multi-connectivity could potentially enhance the system performance by providing multiple simultaneous links between the user equipment (UE) and the base stations. In such systems, packet duplication can be used to meet the stringent reliability and latency requirements of modern cellular networks as data packets are duplicated and transmitted concurrently over two independent links.

This dissertation hence explores a novel idea to improve the performance of multi-connected networks even further. We consider a new radio dual-connectivity (NR-DC) system with packet duplication. The idea is to generate multiple instances (duplicates) of a packet and transmit them simultaneously over different uncorrelated channels with the aim of reducing the packet failure probability. We also propose an enhancement to the packet duplication feature based on the power level reported by the UE for the attached gNB to improve radio resource utilization. The creation and removal of duplicate, redundant packets are performed at the Packet Data Convergence Protocol (PDCP) layer of the Radio Access Network.

Current handover protocols, by design, suffer from interruption even if they are successful and, at the same time, carry the risk of failures during execution. The next-generation wireless networks, like 5G New Radio, introduce even stricter requirements that cannot be fulfilled with the traditional hard handover concept. Particularly, 5G URLLC use cases, such as in transport and manufacturing, demand cell handover latency very close to zero milliseconds, which is a significant reduction from the 30-60 millisecond handover allowed in 4G LTE systems. Another expectation from these services is extreme reliability that will not tolerate any mobility-related failures. Consequently, softer handover concepts where the UE is multi-connected to a source gNB and one or more target base stations are needed. Thus, the next-generation cellular networks further require ultra-reliability in the handover mechanism to ensure the reliability of services is not affected. Thus, in this section of the dissertation, we analyze the system and evaluate its performance with packet duplication under various mobility scenarios for reliable low-latency communication. We also look into how this new approach impacts handovers in a multi-connectivity-based deployment and the overall performance of a 5G network.

1.3 Single Grant Multiple Uplink Scheme for Performance Improvement in Mobility Scenarios

The 5G and beyond 5G networks are considered to be an extremely heterogeneous ecosystem of various technologies like ultra-dense networks (UDN), mmWave approach, IoT-based smart cities, multi-tier network architecture, multiple-input, and multiple-output (MIMO) platforms, etc. The wide variety of use cases in the 5G greatly differs from the use cases considered during the design of third-generation (3G) and fourth-generation (4G) long-term evolution (LTE) networks. Applications like autonomous driving, IoT applications, live video, etc., are much more uplink intensive as compared to traditional applications. However, the uplink performance is often, by design, lower than the downlink; hence, 5G must improve uplink performance.

Simultaneous connectivity through multiple links, like in the case of multi-connectivity, involves independent transmissions to the connected base stations. In the uplink direction, this means that the UE must perform multiple transmissions, which leads to higher power consumption. It further depends on the implementation of multiple stacks that must operate in parallel at the physical (PHY), medium access control (MAC), and radio link control (RLC) layers, leading to increased UE complexity. Hence, to meet the expected performance levels, there is a need to explore flexible network architectures for 5G networks.

In this work, we propose a novel uplink scheme where the UE performs only a single transmission on a common channel, and every base station that can receive this signal would accept and process it. In our proposed architecture, a UE is connected to multiple mmWave capable distributed units (DUs), which are connected to a single gNB-CU. A gNB consists of a gNB-control unit (CU) and one or more gNB-distributed units. CU controls the operation of DUs over the fronthaul interface. The Radio Resource Control (RRC) layer in the gNB-CU configures the uplink grants for a given UE and will provide this uplink grant information to all the gNB-DUs such that all the DUs in the range of the UE's uplink communication can receive and accept the transmission. Each DU that receives the packet processes it and sends it to a common PDCP sub-layer in the gNB-CU that is responsible for the removal of duplicate packets. Only the packet that arrives first is processed and passed on to the upper layers, whereas the copies of the protocol data unit (PDU) from other later DUs are discarded. Thus, in a dense deployment of gNB-DUs, this technique can increase the probability of successful transmission and hence, increase the network's reliability. In an ultra-dense deployment with multiple mmWave base stations around the UE, this removes the need to perform frequent handovers and allows high mobility with reduced latency. In this work, we develop and evaluate the performance of such a system for high throughput and reliable low latency communication under various mobility scenarios.

1.4 Problems and Objectives

With the estimated explosive growth in mobile internet traffic for 5G networks and given the limitations in mobility support, the current network architecture and protocols raise a major challenge for performance enhancement and handover management. In many cases, these designs result in traffic flows where the UE must pass through the same process irrespective of QoS requirement or UE location, etc., leading to single points of failure, suboptimal routing, unnecessary access to mobility resources, and scalability problems. The architecture and protocols of previous generations often favor simplicity at the cost of performance and overall latency, which is unsuitable for 5G and beyond networks. As a result, 5G networks demand new and unconventional protocols and management mechanisms to ensure reliability, low latency, and ease of network operability. Therefore, this dissertation investigates network challenges related to mobility and explores the multi-connectivity techniques as a solution to three basic questions (1) What is the impact of mobility on current architecture in a city environment, and how can multi-connectivity improve this? (2) How can we use multi-connectivity to improve upon the latency and reliability of the network to meet URLCC use case requirements under mobility? And (3) Given the novel 5G uses cases and larger expected uplink traffic, is there a technique that could address these challenges by reinventing the current standards, and how does it perform under mobility?

To study the impact of mobility on next-generation networks, this work develops and systematically analyzes the overall mobility problem in three chapters that refer to the above three questions but with very distinct requirements. Question 1 takes a deep understanding of the current architecture, deployment strategy, and behavior pattern under mobility. Question 2, requires a clear and holistic understanding of the workings and interactions of each protocol layer. And question 3, needs an unconventional, outside-the-box approach to the current standards and protocols. As a result, these studies are intended to understand better the network requirements for handling mobility with multi-connectivity. This dissertation aims to achieve clarity in the three areas of discussion and also propose solutions for resolving these real-world network mobility issues.

The rest of the dissertation is organized as follows: Chapter 2 provides the relevant background knowledge, related works, and motivations for the dissertation work. Chapter 3 goes into the details of how we analyze the problems associated with mobility in next-generation network architectures. Chapter 4 presents details of the implementation and evaluation methodology for the novel techniques to support enhanced mobility. Chapter 5 discusses the issues and experimental design details for studying the impact of mobility on the suggested novel single grant multiple uplink schemes. Chapter 6 summarizes and concludes the dissertation.

Chapter 2

Background

The previous generations of mobile networks were mainly designed with the goal of offering fast and reliable mobile data services to network users. With 5G, that scope has been extended to include a broad range of wireless services that would be delivered to the end-user across multiple access platforms and multi-layer networks. In this chapter, we discuss some of the background details of our work. In Section 2.1, we provide details of the 5G network architecture evolution, the 5G Radio Access Network (RAN) protocol stack, 5G deployment options, etc. As this dissertation's primary emphasis is on understanding network mobility-related challenges and solutions, in Section 2.2, we discuss handovers in more detail. Before we move into the background work related to the topics covered in this dissertation, section 2.3 presents details about the simulation environment used in our work. Section 2.4 presents the work and deployment strategies for multi-connectivity in 5G networks. In Section 2.5 we provide details of the work that was conducted related to the packet duplication in multi-connectivity architecture for improving reliability. And Section 2.6 discusses the background for the design changes to support the proposed single grant uplink scheme.

2.1 5G Evolution

5G network architecture can effectively be viewed as a dynamic, coherent, and flexible framework of multiple advanced technologies working together to support a variety of applications. 5G utilizes a more intelligent architecture, with Radio Access Networks no longer constrained by base station proximity or complex infrastructure. 5G also leads the way towards disaggregated and virtual RAN, with new interfaces creating additional data access points. The Third Generation Partnership Project (3GPP) is the standards development organization that has developed the new 5G system architecture (5GS). It includes a new 5G core network (5GC) and the 5G New Radio (NR) access network. The new core is critical for 5G as it is designed with new service types in mind and also to reap the benefits of modern-day cloud infrastructures [21]. In the following subsections, we first discuss the 5G core network architecture in Section 2.1.1. We then discuss the 5G RAN architecture in Section 2.1.2. Next, we discuss how 5G deployment can roll out along with the current deployment in Section 2.1.3. Section 2.1.4 discusses the ultra-dense deployment strategy in 5G.

2.1.1 5G Core Network Architecture - Service Based Approach

In the previous generation networks, 3GPP has used a point-to-point architecture. This means that different network functions would be connected to each other via standard interfaces that would allow multi-vendor networks. The issue with point-to-point architecture is that because of the many unique interfaces, it creates a web of dependencies between network functions. This results in great difficulty in modifying a network after it is deployed. Thus, if an operator decides to add a new function or upgrades an existing one, then it needs to reconfigure multiple adjacent functions and test them before they can be commercially used. But, in 5G, the operators are expected to serve a variety of use cases and offer diverse services. Thus, 5G networks need a more dynamic architecture that should be able to change on demand as per service requirements. The architecture for 5G systems is described by 3GPP in Technical Specifications (TS) 23.501 [9]. It defines the 5G architecture as service-based, meaning different network functions can be designed as an end-to-end service over standardized application programming interfaces (APIs) that could be deployed on a cloud infrastructure. It allows operators to add, remove or modify network functions and also create new service-specific paths on-demand by decoupling end-user service from the underlying network and platform infrastructure.

The non-roaming 5G system architecture is shown in Figure 2.1. Service-based interfaces are used within the Control Plane (in the top half of the figure) and are named *'Nx'*, where x stands for the network function service it is exposed to. The figure also shows user plane interfaces that use reference point representation and are named *'N*'* where * is a standardized number representing the interface. For example, the *Access Management Function (AMF)* and *Session Management*

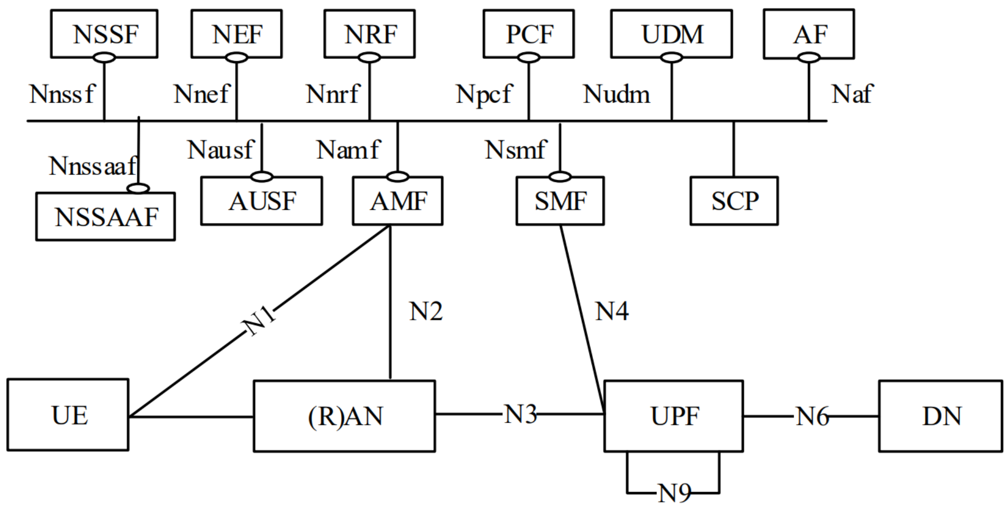


Figure 2.1: 5G System architecture *Source: 3GPP TS 23.501 v16.11.0* [9]

Function (SMF) connect to the user-plane nodes (in the bottom half of the figure) over *N1, N2,* and *N4* to manage subscribers' attachment, sessions, mobility, etc. In Figure 2.1, the *UE* and the *RAN* represent part of the radio access network. The *DN* represents the external data network or the internet. Some of the other major components of the 5G core architecture are listed below with a brief description of their functions:

- **Access and Mobility Management Function (AMF)**: AMF supports termination of Non-Access Stratum (NAS) signaling, NAS ciphering and integrity protection, registration management, connection management, mobility management, access authentication, and authorization, security context management.

- **Session Management function (SMF)**: SMF supports session management (session establishment, modification, release), UE IP address allocation and management, DHCP functions, termination of NAS signaling related to session management, downlink data notification, traffic steering configuration for UPF for proper traffic routing.

- **User plane function (UPF)**: UPF supports packet routing and forwarding, packet inspection, QoS handling, acts as an external protocol data unit (PDU) session point of interconnect to Data Network (DN), and is an anchor point for intra and inter-RAT mobility.

- **Policy Control Function (PCF)**: PCF supports a unified policy framework, providing policy rules to control plane functions and access subscription information for policy decisions in Unified Data Repository (UDR).

- **Authentication Server Function (AUSF)**: AUSF acts as an authentication server.

- **Unified Data Management (UDM)**: UDM supports the generation of Authentication and Key Agreement (AKA) credentials, user identification handling, access authorization, and subscription management.

- **Application Function (AF)**: AF supports application influence on traffic routing, accessing NEF, and interaction with a policy framework for policy control.

- **Network Exposure Function (NEF)**: NEF supports exposure of capabilities and events, secure provision of information from an external application to the 3GPP network, and translation of internal/external information. Thus, it acts as an API gateway that allows external users, such as enterprises or partner operators, the ability to monitor, provision, and enforce application policy, for users inside the operator network.

- **NF Repository function (NRF)**: NRF is a new functionality that provides registration and discovery functionality so that Network Functions (NF) can discover each other and communicate via APIs. NRF thus supports the service discovery function and maintains the NF profile and available NF instances.

- **Network Slice Selection Function (NSSF)**: NSSF supports selecting of the Network Slice instances to serve the UE, determining the allowed Network Slice Selection Assistance Information (NSSAI), determining the AMF set to be used to serve the UE.

The 5G core architecture is designed to make use of network functions virtualization (NFV) and software-defined networking (SDN) techniques, be cloud infrastructure friendly and use service-based interactions between control-plane functions. For this reason, some of the key 5GC design principles were:

1. **Control and user plane separation (CUPS)** to enable independent scalability and decoupled technical evolution. This will also support flexible deployments, such as at centralized and edge locations.

2. 5GC was designed to have a **modular function design** to allow the same functions with different requirements to have different modules.
3. **Minimize dependencies** between the Access Network (AN) and the Core Network (CN) to make it easier to integrate different 3GPP and non-3GPP access types.
4. Have a **unified authentication framework** for efficiency and offer services independent of access method.
5. Support **"stateless" network functions** with decoupled storage and compute resources like in cloud environments.
6. **Concurrent access** to local and centralized services to support access to low latency services hosted in edge data centers. A 5G network deployment could have multiple UPFs deployed remotely, whereas the control plane could be centralized.

It must be noted that the actual deployment and operation of the 5GC are independent of any specification and are not prescribed by the 3GPP. Operators and vendors have considerable freedom to implement the architecture in ways that are suited to the use cases or customers.

2.1.2 5G New Radio and NG-RAN Architecture

The system architecture for Next-Generation Radio Access Network (NG-RAN) is shown in Figure 2.2 and is described by 3GPP in Technical Specifications (TS) 38.300 [14]. The NG-RAN consists of the UE and the next-generation NodeB (gNB or ng-eNB). A gNB provides new radio (NR) user plane and control plane protocol terminations towards the UE, whereas an ng-eNB provides an E-UTRA user plane and control plane protocol terminations towards the UE. For our discussion, we only consider NG-RAN to consist of gNBs. The gNBs are interconnected by means of the Xn interface and are also connected by means of the NG interface to the 5G core (5GC). More specifically, gNBs are connected to the access and mobility management function (AMF) over the N2 interface and to the user plane function (UPF) over the N3 interface.
Some of the primary functions supported by the gNB are:

- Support Radio Resource Management functions like Radio Bearer Control, Radio Admission Control, Connection Mobility Control, Dynamic allocation of resources to UEs in both uplink and downlink (scheduling), etc.

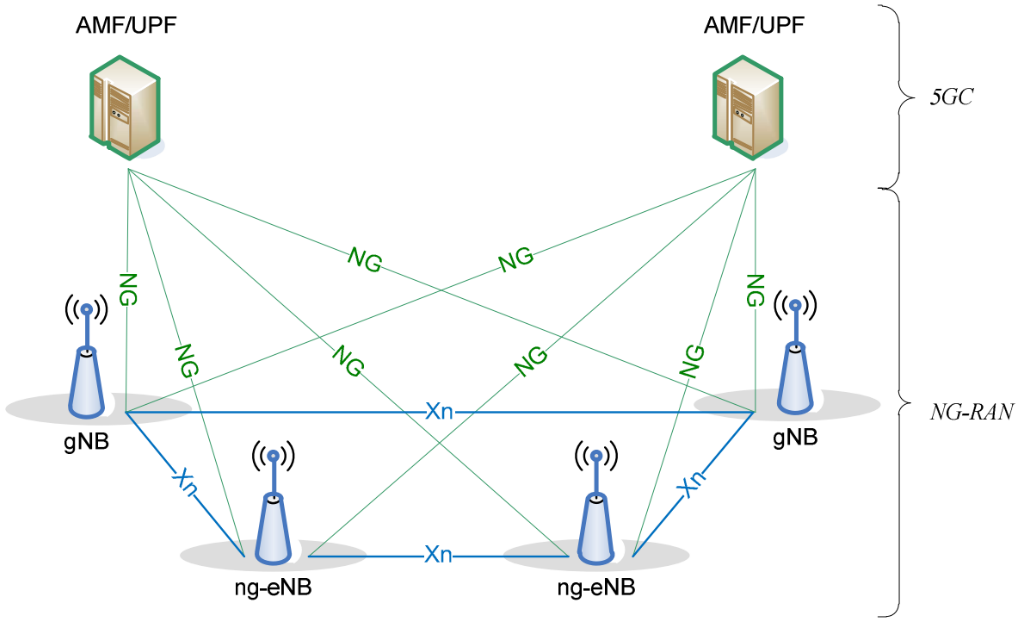


Figure 2.2: Next-Generation Radio Access Network (NG-RAN) architecture *Source: 3GPP TS 38.300 v16.4.0* [14]

- Selection of an AMF at UE attachment when UE has no previous AMF information.
- Routing of User Plane data towards UPF(s).
- Routing of Control Plane information towards AMF.
- Measurement and measurement reporting configuration for mobility and scheduling.
- Scheduling and transmission.
- Support of UEs in Radio Resource Control (RRC) Inactive state.

The radio protocol architecture can be divided into two parts User Plane Protocol Stack shown in Figure 2.3 and Control Plane Protocol Stack shown in Figure 2.4. The user plane protocol stack for NR consists of the physical (PHY) layer, medium access control (MAC) layer, radio link control (RLC) layer, Packet Data Convergence Protocol (PDCP) layer, and Service Data Adaptation Protocol (SDAP) layer. Each layer is responsible for performing a specific task. Some of the key functionalities of each layer are given as follows:

- **PHY**: The PHY layer is responsible for transmitting all information from MAC transport

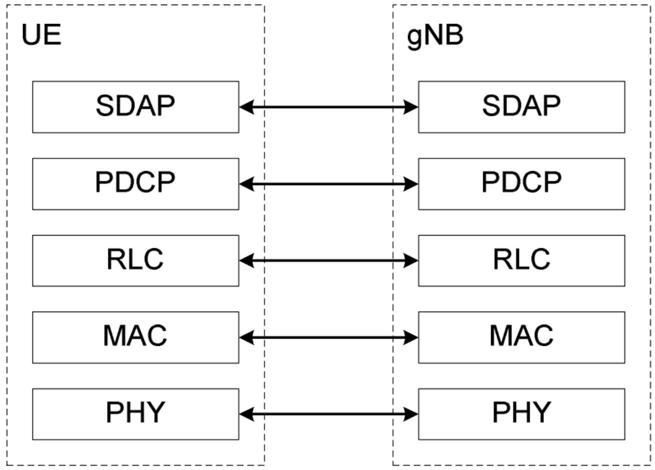


Figure 2.3: NG-RAN User Plane Protocol Stack *Source: 3GPP TS 38.300 v16.4.0* [14]

channels over the air interface and handles different functions such as power control, link adaptation, and cell search.

- **MAC**: The MAC layer provides the mapping between logical channels and transport channels and handles multiplexing/demultiplexing of RLC PDUs. It is also responsible for scheduling information reporting, error correction through HARQ, priority handling between UEs, transport format selection, etc.

- **RLC**: The main functions of the RLC layer include the transfer of upper layer PDUs according to transmission modes, i.e., Acknowledged mode (AM), Unacknowledged Mode (UM), or Transparent mode (TM). Other RLC layer functions include error correction using ARQ in AM mode, sequence numbering, segmentation to match the transmitted PDU size to the available radio resources, etc.

- **PDCP**: The PDCP layer handles the transfer of user data, header compression, sequence numbering, duplication detection, packet duplication, etc.

- **SDAP**: SDAP is a new layer added to the 5G stack that was not present in the LTE protocol stack. This layer mainly handles the mapping between the quality of service (QoS) flow and a data radio bearer.

The control plane protocol stack for NR consists of the PHY, MAC, RLC, PDCP, Radio

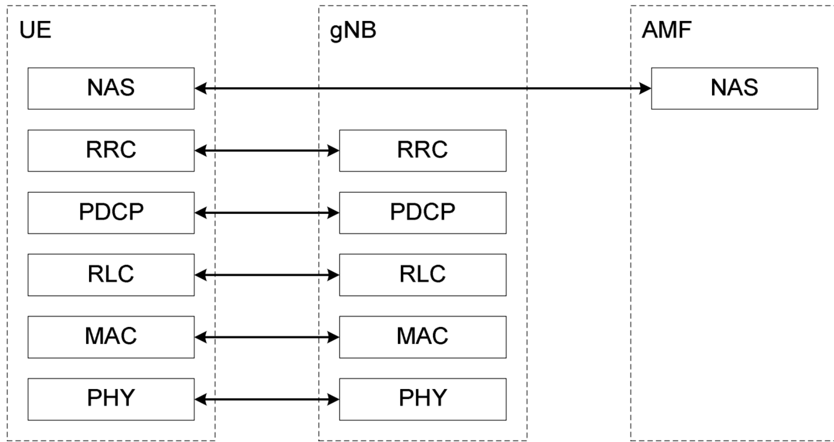


Figure 2.4: NG-RAN Control Plane Protocol Stack *Source: 3GPP TS 38.300 v16.4.0* [14]

Resource Control (RRC), and Non-Access Stratum (NAS) layers. The PHY, MAC, and RLC layers perform the same functions as for the user plane. The PDCP layer performs ciphering and integrity protection. The primary functions of the two new layers specific to the control plane are:

- **RRC**: The main functions of the RRC layer include establishment, configuration, maintenance, and release of data radio and signaling radio bearers. Addition, modification, and release of dual connectivity, broadcast of system information, mobility handling, etc.
- **NAS**: The NAS layer mainly handles connection and session management functions between the UE and the 5GC.

2.1.3 Standalone vs. Non-Standalone 5G Deployment

Although 5G has been standardized, operators have five different options when it comes to 5G deployment. Two network operators can choose to deploy 5G in very different ways. This choice of option could depend on the spectrum licensed to the operator, the geographic area they serve (terrain and user density), the capabilities of the equipment they use, and business factors. 3GPP has defined options covering both 4G and 5G technologies with respect to Radio Access Network (RAN) and Core Network (CN) in its specifications [6, 3]. These options can guide operators as they migrate from current 4G deployments to 5G deployments. It's expected that operators would first deploy 5G NR, let 4G RAN and 5G NR coexist with 4G EPC, and finally deploy 5GC.

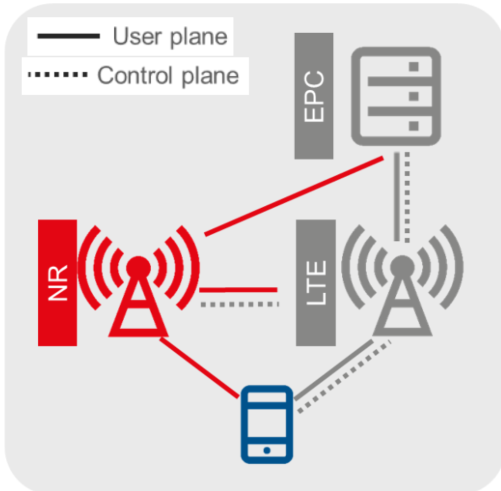
Table 2.1: Summary of 5G deployment options

Connectivity Option	Core Network	Master RAT	Secondary RAT	3GPP Term
Option 2	5GC	NR	-	NR
Option 3	EPC	LTE	NR	EN-DC
Option 4	5GC	NR	eLTE	NE-DC
Option 5	5GC	eLTE	-	eLTE
Option 7	5GC	eLTE	NR	NGEN-DC

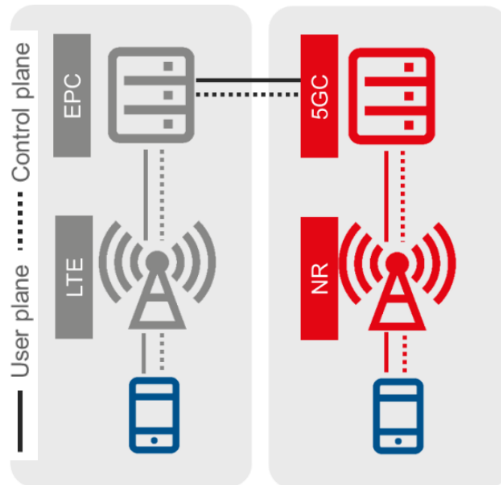
For 5G deployments, 4G RAN can be combined with 5G Core, or 5G NR can be combined with 4G EPC. Different combinations give rise to different 5G deployment options, as shown in Table 2.1. eLTE in the table represents an enhanced eNB radio technology that can directly communicate with the 5G core. Because of this flexibility, deployment scenarios are broadly classified into two categories:

- **Standalone (SA)**: SA deployment uses only one radio access technology, either LTE radio or 5G NR. Both control and user planes go through the same RAN element. Thus, deployment and network management are easier for operators. Inter-RAT handovers are needed for service continuity. Under SA, we have option 2 (5GC + 5G gNB), and option 5 (5GC + 4G ng-eNB).
- **Non-Standalone (NSA)**: In an NSA deployment, multiple radio access technologies are combined. The Control plane goes through what's called the master node, whereas the data plane is split across the master node and a secondary node. There's tight inter-working between 4G RAN and 5G NR. Under NSA, we have option 3 (EPC + 4G eNB master + 5G en-gNB secondary), option 4 (5GC + 5G gNB master + 4G ng-eNB secondary), and option 7 (5GC + 4g ng-eNB master + 5g gNB secondary).

The early deployments are adopting either non-standalone option 3 or standalone option 2, and the standardization of these two options has already been completed. The NSA option 3 leverages existing 4G deployment and can be brought to market quickly with minor modifications to the 4G network. This option also supports legacy 4G devices, and the 5G devices only need



(a) NSA deployment - Option 3



(b) SA deployment - Option 2

Figure 2.5: High-level architecture of non-standalone (NSA) and standalone (SA) deployments [25]

to support NR protocols, so the device can also be developed quickly. On the other hand, NSA option 3 does not introduce 5GC and, therefore, may not be optimized for new 5G use cases beyond mobile broadband. But, SA option 2 has no impact on LTE radio and can fully support all 5G use cases by enabling network slicing via cloud-native service-based architecture. However, this option requires both NR and 5GC, making time-to-market slower and deployment costs higher than that of NSA option 3. Furthermore, the devices would need to support NR and core network protocols, so it would take more time to develop devices. Finally, as the standalone 5G system would need to inter-work with EPS to ensure service continuity depending on coverage, the inter-working between EPC and 5GC would be necessary.

2.1.4 Ultra-Dense Network Deployment (UDN)

Most of the current 4G/LTE mobile network deployments are based on macro-cells. However, macro-cells that cover large geographical areas will struggle to deliver the dense coverage, low latency, and high bandwidth required by some 5G applications. Thus, 5G networks must utilize ultra-dense deployment of access points operating at higher mmWave frequency bands [36]. Ultra-Dense Network (UDN) is viewed as a key enabler in 5G [28] for providing very high data rates to users in indoor and outdoor scenarios. Compared with conventional cellular networks, UDNs are characterized by their high density of base stations (BSs). Although these BSs have limited coverage due to dense deployment, it also increases the probability of a line-of-sight (LoS) transmission. Millimeter-wave communications can be used for enhancing system capacity and simultaneously mitigating the interference levels in UDNs. As mmWave signals have higher propagation losses as compared to sub-6 GHz bands, they cause less interference with proper beamforming techniques. This is particularly important in the deployment of mmWave base stations that offer massive data rates but have a lower transmission range and are subject to blockage. Thus, this densification of base stations in UDN brings with it new opportunities but also introduces unique challenges in the case of mobile users. Technologies like carrier aggregation, multi-connectivity, and massive antennas would have to be used to overcome the higher path loss and blocking associated with such high frequencies. Thus, some of the key features of 5G and beyond 5G cellular networks would include heterogeneity of base stations (BSs), dense/ultra-dense nature of network deployment, multi-connectivity deployment, and diversified mobility patterns of users/devices and network nodes.

2.2 Mobility Challenges in Cellular Networks

For URLLC type of applications, mobility is a key requirement, together with latency and reliability performance. In 5G networks, 3GPP has defined the requirement that for general URLLC services, the target reliability be at least 99.999% and within a latency of 1 ms for a 32-byte packet. Further, NR needs to support mobility of up to 120 km/h for ordinary vehicles and up to 500 km/h for high-speed trains. Mobility performance is one of the most important aspects of wireless communications and becomes even more vital to URLLC type of applications. For many URLLC services, mobility is thus a key requirement, together with latency and reliability. The mobility performance-related KPIs in NR define mobility indicator as the maximum user speed at which a defined quality of service (QoS) can be achieved in km/h. Thus, the target for mobility in 5G networks would be 500 km/h. And, Mobility Interruption Time (MIT) indicator means the shortest time duration supported by the system during which a user terminal cannot exchange user plane packets with any base station during transitions. The target for mobility interruption time should be 0 ms for 5G.

2.2.1 Current Handover protocol

Handovers allow mobility in the network and let users move around freely and connect to any nearby cell that gives them the best service, better throughput, ultra-low latency, etc. Mobility is one of the prime factors to trigger a handover in any given network, but several other parameters like Received Signal Strength (RSSI), Signal to Interference and Noise Ratio (SINR), Fade Duration (FD), Quality of Service (QoS), shadowing, backhaul connectivity, network congestion, reliability, etc., may also result in a handover in any given network. A handover lets a user experience good service, often overcoming network latency and providing higher throughput.

In cellular networks, most handover decisions are based on SNR values. The network is designed to initiate a handover as soon as the SNR drops below a certain threshold. Further, in high mobility scenarios, the SNR changes are often quick, and handover needs to happen on time. Although 3GPP has specified a standard event list for handovers with an SNR threshold in their specifications for 5G, the values for the threshold are normally at the discretion of the service provider and can be altered as per the requirements of the user or the application. It has been seen that SNR/RSRP-based handover decisions have performed well, and thus, any enhancements or newly proposed handover techniques should also be based on SNR/RSRP values. In practical handover protocols, usually, the SNR must stay below the threshold for a predetermined time period, commonly known as the time-to-trigger (TTT). To avoid the ping-pong effect of a UE bouncing between base stations near a cell's edge, there often is some hysteresis for the given threshold. It is also common to have multiple rules working together for making handover decisions. For example, in a high-mobility scenario, the SNR could degrade rapidly, warranting low TTT values as compared to a low-mobility scenario.

As discussed earlier, future 5G networks would need ultra-dense deployments meaning more base stations (BS) within the same geographical region. This would shrink the footprint of each BS and would, increase the spatial spectral efficiency, and offer more capacity. However, the capacity gains offered by such network densification would have an adverse effect of increased handover (HO) rates. Many times such a negative impact on the network is overlooked in light of the improvements seen in capacity, throughput, etc. [18]. An increase in handovers would mean an increase in latency due to the additional time needed for communicating the handover-related control signals. In addition to the signaling overhead, the handover procedure interrupts the data flow to the user due to link termination with the serving base station and link establishment with the target base station. Increasing the handover rate increases the frequency of such undesirable interruptions as well as the associated latency. This may, to some extent, diminish or can even nullify the foreseen network densification gains.

In general, handover is performed in three phases: preparation, execution, and completion. In an intra-NR RAN handover, the preparation and execution phase of the handover procedure is performed without the involvement of the 5GC, i.e., preparation messages are directly exchanged between the gNBs [14]. The release of the resources at the source gNB during the handover completion phase is triggered by the target gNB. Figure 2.6 depicts the basic handover scenario where neither the AMF nor the UPF changes.

As part of the handover initiation, the user reports reference signal measurements from neighboring base stations to the serving base station. For instance, the signal measurement report could include, but not be limited to, reference signal received power (RSRP) and reference signal received quality (RSRQ). In the handover preparation phase, signaling is exchanged between the serving base station, the target base station, and the admission controller. The admission controller makes a decision about the initiation of the handover based on network-defined handover criteria. If the criteria are met, in the execution phase, the user releases the serving base station and attempts to synchronize and access the target BS using the random access channel (RACH) procedure. Upon synchronization with the target base station, the UE sends the confirmation message to notify the network that the radio resource control (RRC) handover is completed. In the final completion phase, the target gNB sends a message to AMF to trigger 5GC to switch the data path toward the target gNB and to establish a control interface towards the target gNB. Any new or buffered data that is pending to be delivered to the UE at the source gNB after the handover request is acknowledged at the end of the preparation phase is sent to the target gNB via the Xn interface for delivery. The target gNB buffers any new data destined for the UE until the path switch request is completed and a UE context is created at the target gNB. After this point, the UPF sends all UE-bound data to the target gNB, and the UE context in the source gNB is released, marking the end of the completion phase and the handover.

The handover procedure just discussed involves signaling overhead between the user, involved base stations, and the core network. This increases the latency, interrupts the data flow, and decreases the throughput of the mobile user. The rate at which such interruptions happen is a

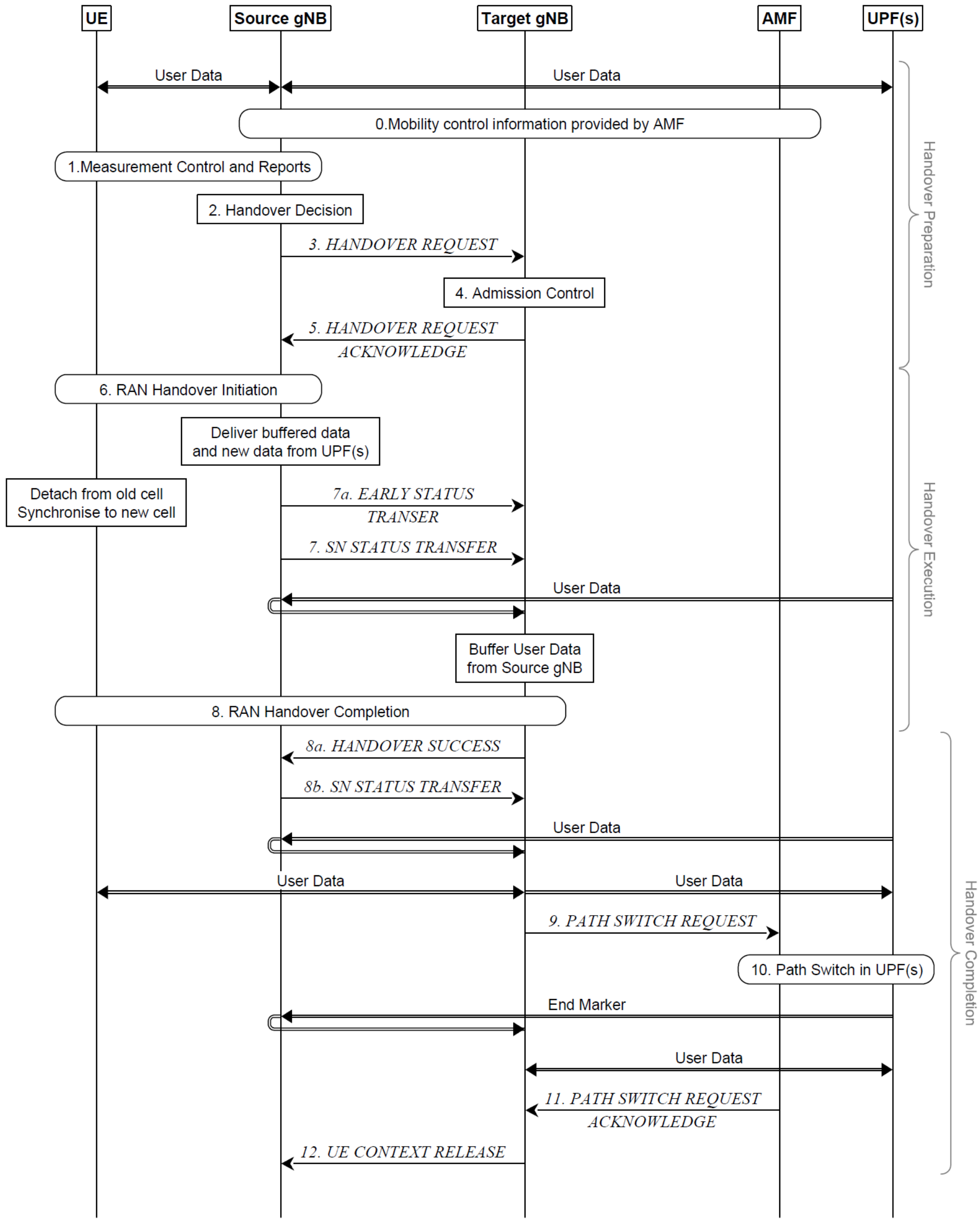


Figure 2.6: Intra-AMF/UPF Handover in 5G NR *Source: 3GPP TS 38.300 v16.8.0* [14]

function of the base station intensity and user velocity. The duration of each interruption is termed the handover delay and is measured from the beginning of the initiation phase to the end of the execution phase. Thus, at high velocities and/or dense cellular environments, it is important to have few handover-related delays or decrease the time of handovers.

2.2.2 Literature Survey on Handovers

Over time there have been many enhancements made to handover protocols to reduce the handover signaling time. Authors in [27] use a Software Defined Networking (SDN) and Network Function Virtualization (NFV) based enhanced mobility management unit implementation to enhance the signaling of the handover methods proposed by 3GPP. The use of X2 links that connect the base stations directly has been exploited to reduce handover times, and also work has been done to utilize backhaul resources more efficiently. [51] proposes a novel key exchange and authentication protocol for performing fast handovers when there are intermediate mobile terminals used for the static merging of fronthaul and backhaul equipment to allow better possession of services. A smart combination of three different metrics, i.e., RSSI, data rate, and SINR, was proposed in [57] to help make more efficient handover decisions instead of using each component individually. Multi-connectivity approaches with SINR evaluation have been used by researchers to achieve high reliability [46, 55]. The authors also showed how single connectivity reliability deteriorated with mobility and proposed a multi-connectivity concept as a solution for mobility-related radio link failures and throughput degradation of cell edge users. However, the concept used in these works relies on the fact that the transmissions from cooperating cells are coordinated for both data and control signals using a soft combining approach at the physical layer, more popularly known as Co-ordinated Multipoint (CoMP) joint transmission technique. The Multi-connectivity technique used in the dissertation work focuses on using two parallel network stacks at the UE with a single PDCP layer that is responsible for duplication removal. This technique allows flexibility and decreases complexity at the physical layer considerably.

2.2.3 Previous Work on Reducing Handover Latency for URLLC

A lot of URLLC work has been dedicated to lower-layer procedures, such as channel coding, Hybrid Automatic Repeat Request (HARQ), or scheduling improvements. In previous generations, packet reliability requirements are traditionally satisfied using Automatic Repeat Request (ARQ) and hybrid ARQ (HARQ) retransmission techniques at the radio link control (RLC) and medium access control (MAC) layers, respectively. In 5G NR, flexible numerology has been introduced to vary the sub-carrier spacing, and mini-slots can be used to enable ultra-reliable low-latency communication (URLLC) type applications. However, in NR, when it comes to handover principles and procedures, they are mostly the same as the previous generation [14]. Hence the issues with large handover interruption time and mobility robustness remain in NR as well. 3GPP has suggested a few enhancements like Make-Before-Break handover and RACH-less handover in protocol to decrease the interruption time in NR. Also, conditional handover is introduced to improve mobility robustness.

In Make-Before-Break handover (MBB HO), the interruption time can be reduced by not releasing the connection to the source gNB until the initial uplink (UL) transmission to the target gNB is established. The UE performs simultaneous reception of synchronization signals from a target cell while doing Tx/Rx operation in the serving cell. MBB HO can potentially reduce up to 35 ms of interruption time. However, in MBB HO, the timing of data forwarding from the source gNB to the target gNB could be an issue. The optimal time to start data forwarding is when UE performs the first UL transmission to the target gNB, and any time point earlier or later would lead to an increase in handover interruption time [10].

In RACH-less handover, the Random Access (RA) procedure to connect to the target gNB is skipped to save interruption time. This is done by synchronizing the base stations to avoid any timing advance and by transmitting the UL grant during the HO command. This can save up to 8.5 ms, but the timing of the UL grant could be an issue in such handovers. Ideally, the UL grant should be given right before the UE sends the HO complete message to the target gNB, but if it is sent earlier, then UL grant resources are wasted, and if the UL grant is delayed, then it will add to the interruption time. It is also possible to have MBB HO and RACH-less HO simultaneously.

In Conditional Handover (CHO), the UE would not immediately execute the HO after receiving the HO command. The source gNB decides when to execute HO based on a configured condition. When the condition is met, the source gNB receives the measurement reports from the UE and starts HO preparation procedures with the target gNB. After this, the UE receives an early HO command from the source gNB, which it saves for later. When the HO execution condition is met at the target gNB, the UE executes the previously received HO command. The HO preparation event happens when the target gNB SNR is slightly better than the source gNB SNR and thus helps reduce handover failure rates. The HO execution happens only after the SNR of the target gNB has gotten even better and thus reduces the ping-pong rate [48]. The UE also notifies the source gNB of the timing of an HO execution with an HO Indication message after synchronizing to the target cell to avoid the source gNB guessing when to perform an HO execution. The interruption time for conditional handover is similar to that of MBB HO.

2.3 Simulation Environment

5G networks offer a wide range of services, and their performance depends on a wide range of factors, like the deployment design, architecture, spectrum, front-haul, back-haul network, core services, etc. To be able to study the collective impact of all or some of these cellular network components, it is critical to use a platform capable of performing an end-to-end analysis. Given that there are many infrastructure complexities associated with building a real-world cellular network for testing. Discrete-event network simulators offer a far better and more scalable alternative to perform the analysis of these complex networks and study the development of novel protocols.

Wireless network simulators include physical-level simulators which are only suitable for physical-layer measurements, such as the signal-to-interference-plus-noise ratio (SINR) or spectral efficiency. On the other hand, link-level simulators many times do not even consider modeling layers above the MAC. Thus, to analyze system properties like end-to-end latency, throughput, packet loss, etc., application-level performance must be considered, which is only possible through *end-to-end simulators*. End-to-end network simulators can perform full-stack simulations using models for all the layers of the protocol stack, network equipment, and application logic. This ability to simulate the whole network stack plays a crucial role in understanding many new features of 5G networks.

The two most widely used network simulation frameworks are OMNeT++ and ns-3, which allow users to develop their own model libraries. Apart from Simu5G, which is based on the OMNeT++ framework and discussed in more detail in section 2.3.1, the other popular end-to-end simulation tools for 5G networks are based on ns-3. The 5G-LENA [49] model library uses the ns-3 framework and builds upon the LENA (LTE-EPC Network Simulator) 4G LTE library [19]. The 5G-LENA library mainly focuses on implementing the MAC and PHY layers of the 5G network stack. The ns-3 mmWave module was developed before 3GPP 5G standards were finalized and thus remains non-compliant with current standards. Both 5G-LENA and mmWave modules only support the Time Division Duplexing (TDD) mode. In this work, we used Simu5G [43], which is a 3GPP compliant 5G network simulator that implements the 5G New Radio User Plane Simulation Model with a fully customizable and modular design.

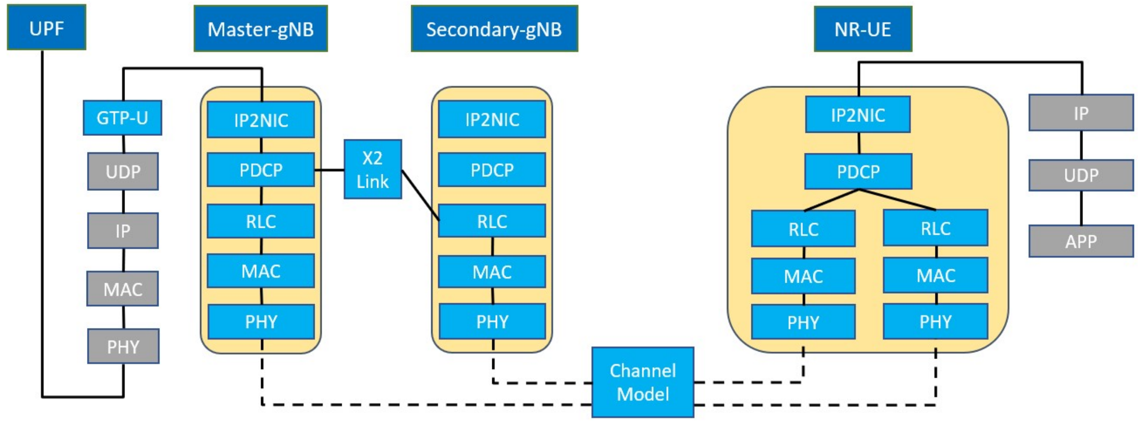


Figure 2.7: Simu5G sub-module structure

2.3.1 Simu5G Network Simulator

Simu5G is a 5G New Radio User Plane Simulation Model for OMNeT++ (Objective Modular Network Testbed in C++) simulation framework using the INET library. It is built upon OMNeT++ [47], which is a widely used discrete-event simulation framework that can be used to model wired and wireless networks, among others. The basic building blocks in the OMNeT++ framework are called *modules*. Modules can be simple, or they can be combined to create more complex compound modules. Modules can be linked through their interfaces called *gates* with links known as *connections*. Modules communicate among each other using *messages*, and simple modules are programmed to exhibit a specific behavior on receipt of these messages. The OMNeT++ model behavior is programmed in C++. It uses Network Description Language (NED) to define the modules along with their gates and connections. The parameter values needed to initialize a model are defined in an initialization (INI) file. It also provides an Eclipse-based Integrated Development Environment (IDE) to facilitate editing and debugging. Simu5G also uses the INET model library in OMNeT++, which implements networking-related models for various communication protocols, network nodes, connections, etc. This allows Simu5G to simulate highly complex end-to-end 5G network system scenarios for developing new modules and implementing new algorithms [29].

2.3.1.1 5G RAN Module Implementation

In Simu5G, the functionalities of layer 1 and layer 2 of the RAN are implemented as a stack of four protocols, which are the Packet Data Convergence Protocol (PDCP) layer, Radio Link

Control (RLC) layer, Media Access Control (MAC) layer, and Physical (PHY) layer, all combined to form the Network Interface Card. Simu5G simulates the NR capabilities of the 5G RAN through two main compound modules, *NrUe*, and *gNodeB*. Figure 2.7 shows the submodule structure of the two modules in an NR-DC deployment using Simu5G. The *gNodeB* module consists of four submodules *NrPdcp*, *NrRlc*, *NrMac*, and *NrPhy*, each representing a layer in the protocol stack along with an *Ip2Nic* submodule that deals with the IP (Internet Protocol) packets between the UPF and the PDCP layer. The *gNodeB* module also contains an *X2Manager* module responsible for forwarding IP packets on the X2/Xn link towards the secondary node when multi-connectivity is implemented. The PDCP layer on the master gNB is responsible for deciding if a packet will be sent to the lower RLC layer on the local stack or forwarded through the X2 link. The forwarded packet is then received by the RLC layer of the secondary gNB, after which the packet continues to flow through the lower layers toward the UE.

The *NrUe* module, on the other hand, implements all the layers of the protocol stack from the physical layer up to the application layer. It uses the INET library to implement the UDP and IP protocol layers in the stack. To implement dual connectivity, the *NrUe* has a dual stack with two PHY, two MAC, and two RLC layer submodules. *NrUe* has a single *NrPdcp* module responsible for reordering the incoming RLC PDUs (Protocol Data Units) from the dual stack and sending them to the higher layers.

2.3.1.2 5G Core and Other supporting Modules

As part of the 5G core implementation, Simu5G provides a *UPF* module responsible for handling IP packet flow between the data network and the gNBs through GTP (GPRS Tunnelling Protocol) tunnels. The *UPF* module is directly connected to a gNB resulting in the standalone architecture deployment scenario, and each gNB is connected with its adjacent gNB through an X2/Xn link. The standalone deployment scenario was extended to implement the NR-DC deployment in this work. During handovers, the source gNB buffers the received packets and sends them to the destination gNB through these X2/Xn links. The native version of Simu5G was extended for conducting this study to include gNB to gNB handovers in single connectivity and multi-connectivity architectures.

Simu5G implements a *Binder* module that is visible to every other node in the simulation and stores information about them, such as references to nodes. It can be used to locate, for instance, the interfering gNBs in order to compute the inter-cell interference perceived by a UE in its serving cell. Simu5G also offers a *carrierAggregation* module that can be used for communication on multiple carrier component frequencies. Using the parameters in the NED and INI files, we can choose the carrier frequencies, bandwidths, number of resource blocks, etc. Through this module, we can also adjust the numerology that varies the subcarrier spacing and the slot duration of 5G signals.

2.3.2 Channel Model

In Simu5G when any application generates a packet and sends it to the cellular NIC implemented, it flows down the protocol stack at the sender through the PDCP and RLC layers to the MAC layer encapsulated as a MAC PDU. When this MAC PDU is sent from a sender to a receiver, an OMNeT++ message is exchanged between them. On receipt of the message, the receiver node applies a channel model to compute the received power. The channel model has been configured to incorporate fading, shadowing, pathloss, etc., and can be made arbitrarily complex. From the received power, the receiver node computes the SINR, by querying the *Binder* module to know which other nodes were interfering with the same resources. Then, it leverages Block Error Rate (BLER) curves to compute the reception probability for each resource block (RB) used in the ongoing transmission. This allows Simu5G to translate an SINR and a transmission format to a probability of correct reception of the entire MAC PDU. The above modeling reduces the computational complexity of the decoding operation, thus decreasing the simulation running time while preserving its correctness.

Further, in Simu5G, each MAC transport block (TB) is encapsulated within an *AirFrame* message and sent to the destination module, which applies the model of the air channel to decide whether the *AirFrame* was received successfully or not. Since a MAC TB is associated with a given component carrier (CC), the corresponding *AirFrame* is subjected to channel effects (e.g., path loss, shadowing, etc.) that depend on that CC. This means that different channel models have to be applied to compute the SINR on the receiving side. For this reason, in Simu5G, each gNB/UE is equipped with a vector of *channelModel* modules, and each of them is associated with one of the CCs available in the *carrierAggregation* module.

To be able to study and develop novel network architectures and protocols, it is critical to have an accurate and reliable performance evaluation of the complex 5G network. It is important that channel models used for simulating radio access networks characterize their behavior precisely.

Standardization bodies like the 3GPP alliance have worked on developing such models and evaluation methodologies. TR 36.873 presents a study on a 3D channel model for LTE and covers the modeling of the physical layer of both the mobile UE and Access Network of 3GPP systems [1]. However, for the deployment of 5G networks, 3GPP NR specifications are being followed, which are the first to utilize mmWave communications. 3GPP NR can operate on a wide range of the spectrum, which includes the traditional sub-6 GHz bands as well as the mmWave frequency bands. These standards also support high mobility in devices that can communicate while moving at different speeds of up to 500 km/h. 3GPP in TR 38.901 [5] describes the channel modeling framework for frequencies from 0.5 GHz up to 100 GHz. The model also characterizes various channel conditions, propagation loss, and small-scale fading due to the effect of Doppler and multipath.

As discussed earlier in Simu5G, physical layer transmission characteristics are modeled via realistic and customizable channel models. The Simu5G release version has the *ChannelModel* module that uses the TR 36.873 specification, which is applicable only for sub-6 GHz channels. To accurately evaluate our proposed system, channel models from TR 38.901 have been incorporated into Simu5G in this work. This allows us to use Simu5G to study the performance of 5G networks using multi-connectivity and conduct a systematic evaluation of the system based on real application scenarios.

2.4 Mobility Enhancement using Multi-Connectivity in 5G Networks

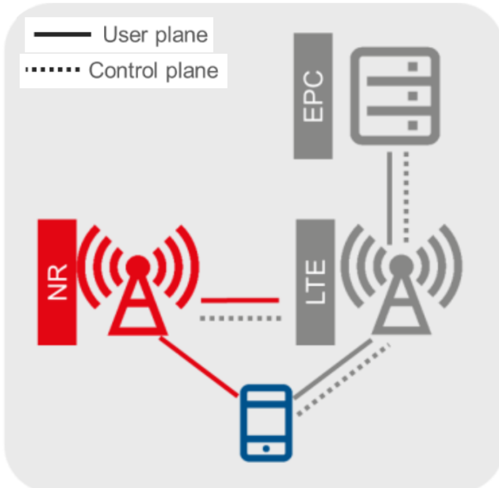
2.4.1 Multi-Connectivity Architecture for 5G Networks

Simultaneous connectivity to multiple radio access technologies or multiple bands represents an important solution for ongoing 5G deployments. It enhances the system performance by providing multiple simultaneous links between the user equipment (UE) and the base stations (BS) for data transfer and also eliminates the time needed to deal with frequent handoffs, link establishment, etc. Multi-connectivity has been widely recognized as one of the key features for improving the performance of next-generation wireless networks. The latest 3GPP standards [11] define Multi-Radio Dual Connectivity (MR-DC) as an architecture where a UE is configured to utilize resources provided by two different nodes connected via non-ideal backhaul, with one providing NR access and the other one providing either E-UTRA or NR access. One node acts as the master node (MN), and the other as the secondary node (SN). Suer et al. in [53] provide an overview of multi-connectivity and its potential to achieve reliable low-latency communication.

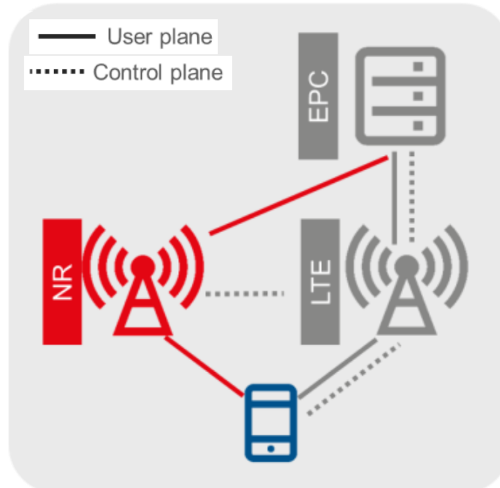
In multi-connectivity, we have two uncorrelated paths that separate out in the transmitter and get combined at the receiver. This separation and combining of paths could happen at different layers in the RAN protocol stack. On the Physical Layer (PHY), the different approaches are combined at the signal or symbol level, while for higher-layer approaches like on MAC Layer and above, path combining happens at the packet level. The physical layer multi-connectivity approaches aim at increasing the signal-to-interference-plus-noise ratio (SINR) at the receiver with coordinated multi-point (CoMP) like techniques [26] or by improving the bit error rate (BER) for a certain SINR by combining multiple paths. Authors in [53] briefly discuss the three combining schemes of selection combining (SC), maximum ratio combining (MRC), and joint decoding (JD) that could be used for PHY layer MC. For higher-layer approaches, techniques like MP-TCP could be used on Transport Layer, or Dual Connectivity at PDCP Layer can be implemented. Different scheduling techniques can be used for multipath on the Packet Level. First, there is the packet splitting (PS) or split bearer (SB) technique, where the payload is split into fragments, and a fragment is sent over each path. Sometimes entire packets could be marked by higher layers and would be split at the packet level without fragmentation. The marker would then decide which path the packet would take. In packet splitting receiving node would wait for all fragments to arrive before it would reassemble and forward the packet to the application. In such cases, by reducing the payload size per path, the transmission latency is reduced. The next packet-level MC technique is load balancing (LB), where the packets are distributed over all available paths. This reduces the load per path and thus reduces the access and queuing latency. Finally, there is the packet duplication (PD) technique which we discuss in more detail in Section 2.5. Here, the payload is duplicated and sent over multiple paths, and on the receiver side, only the first arriving copy of the packet is accepted and forwarded to the application. This technique could result in latency reduction and increases reliability.

2.4.2 Multi-Connectivity Deployment Strategies in 5G Networks

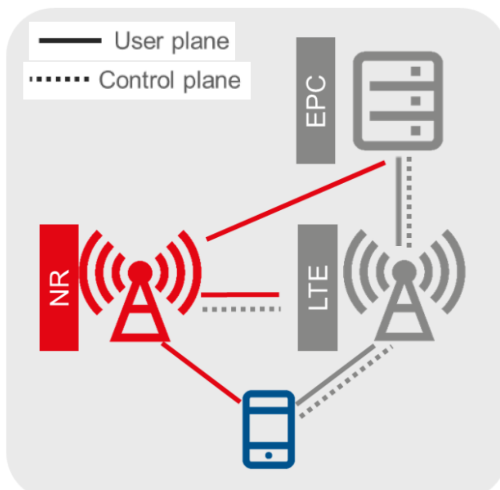
As discussed earlier in Section 2.1.3, there are three dual-connectivity options that 3GPP specified. Out of the three, two options need 5G core implementation and enhanced eNBs. The simpler and more popular commercial deployment option for operators hence has been option 3 or



(a) NSA deployment - Option 3



(b) NSA deployment - Option 3a



(c) NSA deployment - Option 3x

Figure 2.8: High-level architecture of non-standalone (NSA) Option 3 deployment variants [25]

EN-DC. For the deployment of EN-DC, the existing 4G network needs to support dual connectivity between E-UTRAN (LTE) and NR. This enhancement allows users to use radio resources provided by both 4G and 5G. Typically the 4G radio will be used to carry control signaling, while NR and/or LTE will be used for user data. Three variants of the NSA Option 3 solution have been defined, each producing a different impact on the LTE network. In all option 3 variants, the control signaling bearer connecting the core is terminated at the Master node, i.e., the LTE eNB. The variants thus differ in their data bearer configurations and are shown in Figure 2.8.

Thus, the standardized NSA EPC networking architecture or EN-DC includes Option 3, Option 3a, and Option 3x. In the Option 3 networking mode shown in Figure 2.8a, the X2 interface carries a large amount of user-plane traffic between eNB and gNB. Thus to be able to handle this traffic, the core network would need a significant increase in the bandwidth of the S1-U interface between eNB and EPC. In the Option 3a deployment mode shown in Figure 2.8b, there is only control plane traffic in the X2 interface, which is usually very small and hence not very suitable. Finally, there is the Option 3x networking mode shown in Figure 2.8c, where there is little LTE user plane traffic in the X2 interface. Option 3x puts the least amount of pressure on the existing core network. As a result, Option 3x has become the mainstream choice of operators deploying NSA models for 5G networks. As it uses 4G as the anchor for the control plane, it can provide good service continuity and flexibility to support rapid network construction in the initial stage of 5G deployment.

Initial 5G deployment used the non-standalone architecture where E-UTRA-NR Dual Connectivity (EN-DC) was used to improve the user throughput while maintaining good coverage. It did so by leveraging the already existing LTE infrastructure with innovative technologies deployed in 5G-NR. Multi-connectivity, specifically EN-DC, thus played an important role in ensuring that the high throughput benefits of 5G are available to a maximum number of users while ensuring connectivity during mobility scenarios. However, rapid 5G deployment is expected in the upcoming years, with more base stations having mmWave capabilities and the development of 5G core services. Moving towards a standalone 5G architecture enables us to leverage attributes of 5G, such as high throughput and ultra-reliable low latency communication. However, using the FR2 band alone could significantly impact the reliability of the network. Thus, the 3GPP standards propose the New Radio Dual Connectivity (NR-DC) feature containing high band NR small cells overlaid by sub-6 GHz FR1 band 5G macrocells for effective deployment. Furthermore, with the evolution of 5G services, more and more standalone 5G architectures are being deployed. Hence, in this study, we focus on NR-NR dual connectivity (NR-DC), in which the UE is connected simultaneously to two gNBs, with one acting as the master gNB and the other as the secondary gNB. NR-DC approach has been studied in more detail by the authors in [39].

2.5 Improving Reliability in Mobility Scenarios using Multi-Connectivity with Packet Duplication

This section discusses how we can improve the performance of multi-connectivity type deployments to satisfy the requirements of URLLC use cases. To improve reliability in UDN-type deployments and in mobility scenarios, it has been proposed that Packet Duplication in a multi-connectivity deployment could be particularly beneficial. It reduces the need for frequent handovers in mobility scenarios and can improve transmission robustness by minimizing the probability of link failures, especially at mmWave frequencies. The improvement in reliability of a system using packet duplication can be proven theoretically by using reliability theory from systems engineering [31]. The use of multiple redundant subsystems that operate independently and in parallel can be shown to theoretically increase the reliability of a system. If the reliability of i^{th} subsystem is assumed to be R_i, then the overall system reliability R can be calculated as follows:

$$R = 1 - \prod_{i=1}^{N} (1 - R_i)$$

where N is the number of independent uncorrelated subsystems. In the case of wireless communication, these systems would correspond to transmission links or channels. In a multi-connectivity setup, each parallel path from the transmitter to the receiver would be an uncorrelated system. Thus, to increase the reliability of the overall system, we could either improve the reliability of each individual uncorrelated link or increase the number of such uncorrelated links operating in parallel.

2.5.1 Reliability Enhancement using Packet Duplication

In LTE standards, dual connectivity (DC) design supported the split bearer model where data is split at the packet data convergence protocol (PDCP) layer, transmitted using two radio paths, and combined again at the receiver PDCP layer. Thus, it can be used to increase the throughput and perform load balancing based on the radio link quality. The latest 5G standards expand this DC functionality for reliability enhancement by allowing data or packet duplication (PD) in which the same data packet is independently transmitted through two different base stations [15]. The most obvious upside of using packet duplication for improving reliability is that it comes with no additional latency cost. Standards have suggested using Automatic Repeat Request (ARQ) in RLC

Acknowledged mode (RLC AM) and Hybrid Automatic Repeat Request (HARQ) in the MAC layer to ensure reliable communication. However, both these techniques introduce retransmissions and hence increase the overall latency of the system. Furthermore, packet duplication can be performed with minimal computational effort as compared to other techniques like Co-ordinated Multi-Point (CoMP) transmissions defined for LTE [24]. In this dissertation, the packet duplication function is incorporated at the PDCP layer and performs the duplication of protocol data units (PDUs). The duplicate PDUs are detected based on common sequence numbers and removed at the receiving PDCP entity in the UE.

In this dissertation work, we focus on a multi-connectivity approach with two active connections in the downlink direction, i.e., dual-connectivity towards terminals from two serving base stations. Authors in [50, 16] analyze in detail the PD technique and the related changes made to the 3GPP standards to satisfy URLLC requirements. There could possibly be many different implementations of reliability-oriented dual-connectivity, depending on the layer where we decide to per from the duplication. For example, data can be duplicated at the physical layer by using CoMP or Multi Transmission, and Reception (Multi-TRP) based approaches like joint transmission (JT). In Physical layer duplication, the transmitters must coordinate their transmission, and on the receiver side, the UE would combine the received packet at the physical layer. The rest of the layers stay agnostic to multi-connectivity. Such a high degree of coordination between the two active base stations and tight synchronization between transmission-reception points (TRPs) requires a high-speed backhaul connection, increasing the cost of network deployment.

On the other hand, in higher layer duplication, the packet is duplicated at the PDCP layer or above (transport layer in MP-TCP or Application). Each packet then passes through the lower layers of its base station and reaches the UE. In this case, compared to PHY layer duplication, no coordination is needed, and the packet can eventually be transmitted at different time instants, over different frequency resources, and with different physical layer parameters such as modulation and coding scheme (MCS). On the receiver side, only the first version of the arriving packet is passed on to the higher layer, and other duplicates are dropped. Thus performing packet duplication at higher layers has more relaxed requirements in terms of backhaul connection since the duplicated packets are not to be transmitted simultaneously in a synchronized manner. DC-based PDCP data duplication in downlink is also studied using an analytical model in [22]. The authors also highlight the advantages of PD in terms of packet reliability with no additional latency when compared to other techniques like Hybrid Automatic Repeat Request (HARQ), Co-Ordinated Multi-Point (CoMP) transmissions, and Multiple Transmission-Reception-Point (multi-TRP) schemes, etc.

2.5.2 Optimizing Packet Duplication for Efficient Radio Resource Utilization

The reliability gain offered by packet duplication can be offset by the link capacity reduction due to the transmission of redundant packets. Packet duplication increases the radio resource utilization, which could potentially outweigh its benefits. Thus, it is critical to consider enhancements to improve resource efficiency when implementing duplication. Authors in [37] perform a study on a reliability-oriented dual connectivity solution for URLLC applications using packet duplication and present its impact in terms of reduced network spectral efficiency. However, the study considers an idealized scenario that does not consider variables like the selection of nodes, queuing delays at high network load, etc. In this dissertation, we present results using an end-to-end simulation of a 5G network that considers all facets of a communication network. A recent ns-3 simulation-based study of Multi-Connectivity schemes with packet duplication for low latency applications was performed in [54]. However, the authors perform the simulations on a Wi-Fi network based on IEEE 802.11ac standards. The results of packet duplication and its optimization using UE-reported power are presented in this study which focuses on NR-DC with packet duplication in a highly realistic 5G network.

2.6 Single Grant Multiple Uplink Scheme for Performance Improvement in Mobility Scenarios

As discussed in this dissertation, multi-connectivity with packet duplication [41, 40] involving simultaneous connectivity to multiple radio access technologies or bands provides a way to increase reliability for ongoing 5G deployments [53]. But, multi-connectivity schemes also increase the user equipment (UE) complexity as multiple stacks must operate in parallel at the physical (PHY), medium access control (MAC), and radio link control (RLC) layers. In the uplink direction, this also leads to higher power consumption at the user equipment (UE) [38] as two independent transmissions must happen to the dual-connected base stations. With newer 5G use cases that demand high uplink throughputs, it becomes inevitable to explore flexible network architecture to meet the stringent requirements of 5G systems. Traditionally, each newer generation of cellular protocol standards has emphasized improving the downlink throughput compared to the uplink. Achieving such high levels of throughput in the uplink, along with high reliability, is an open research question. Thus, it is critical to explore multiple uplink schemes that will address these needs of the 5G systems.

2.6.1 Uplink transmission in 5G systems

In cellular networks, the UE and the base station are in constant interaction with each other. For successful communication, a UE must be always in sync or in complete synchronization with respect to the base station or gNB. In the downlink, this synchronization process helps UE detect the radio boundary, or the exact timing when a radio frame would start, and the OFDM symbol boundary, or the exact timing when an OFDM symbol starts. This process is done by detecting and analyzing the Synchronization Signal Block, which is part of the Physical Broadcast channel or PBCH. However, the uplink synchronization process is used by the UE to figure out the exact timing when it should send uplink data using the Physical Uplink Shared Channel (PUSCH) and Physical Uplink Control Channel (PUCCH). Usually, a gNB in the network is handling multiple UEs and the network has to ensure that the uplink signal from every UE should be aligned with the common receiver timer of the network. Thus the network has to adjust the UE's Tx timing or uplink timing of each UE, and this is usually done using the Random Access procedure or the RACH process.

In cellular networks, once the UE is synchronized with the gNB, it uses a Scheduling Request (SR) to signal to the network that it wants to send uplink data. Thus, SR is a special physical layer message for UE to ask the network to send a UL Grant so that UE can transmit PUSCH. Even though an SR message is a kind of physical layer message, it is controlled by the MAC layer process. Once SR is transmitted and gNB receives it, gNB should send UL Grant and UE has to send PUSCH in response to the UL Grant. The timing among SR, UL Grant, and PUSCH varies on whether it is FDD or TDD. The allocation of resources in the UL grant to the UE by the network is determined based on the amount of data the UE has indicated it wants to send and the number of resources the network is able to spare for the transmission. Buffer Status Report or BSR is the MAC layer control element that carries the information on how much data is in the UE buffer to be sent out. This process allows the network to allocate UL resources (UL Grant) only when UE has something to transmit and avoids allocating too many UL resources (more than what UE needs), thereby leading to the wastage of resources. However, when we consider multi-connectivity with packet duplication on the uplink, the conventional procedure would require individual UL grants to be sent to the UE for each connected gNB. This is despite the fact that these UL grants are essentially for transmitting the exact same data over to the network.

There has been a lot of research done to improve UL transmission in 5G networks. 5G standards allow the network the option to have UL configured-grant (CG) transmission, where a UE can be configured with periodic transmission resources by the base station so that the UE can transmit data in these resources without scheduling request (SR) and UL grant [13]. But this is only practical when the communication is very UL intensive as compared to DL. With mMTC type of applications, there would be an increase in IoT networks. To support the sporadic traffic generated from massive IoT networks, the authors in [30] consider Grant-free multiple access (GFMA) protocols. In GFMA protocol, each device transmits data packets without a grant from a base station via pre-reserved uplink resources. However, this type of transmission works when Tx are far apart in time, otherwise, this leads to inherent collisions when multiple UEs try to simultaneously access the same frequency resources [45] and thus needs multi-packet reception (MPR) capability at the base station.

Having multiple UL grants to do uplink transmission in multi-connectivity type deployment also leads to extra usage of radio resources. Even though the same packet is transmitted, the multiple UL grants would be configured with different resource blocks far apart in time and frequency. The strategy we consider allows us to use a single UL grant that is distributed across multiple gNB's distributed units (DUs) and connected to a central unit. This also allows us to have more centralized control of the radio resources that would be allocated to these DUs. This can be paralleled to a cell-free arrangement of network deployment. There has been some previous work done in this direction, and Cell-free massive multiple-input/multiple-output (MIMO) is considered a promising technology. But this increases detection complexity, hardware implementation, etc. as discussed by authors in [17]. Our implementation focuses on shifting this complexity to a more capable and powerful central unit and deploying a software-based solution. Reusing the same transmission can improve the spectral efficiency of the system. Some other ways of improving upon this efficiency have been discussed, one of them being Non-orthogonal multiple access (NOMA), which can improve spectral efficiency and reduce latency. But NOMA has reliability issues that limit its usage in URLLC type of applications [58]. Additionally, limited research studies consider achieving high throughput and reliable low-latency communication for mobility scenarios. In our work, we consider these scenarios, and the proposed technique increases the probability of successful transmission and improves the throughput, reliability, and latency.

In the past, there have been works focusing on performing multiple uplink transmissions to improve performance. One of these works focused on improving the throughput of the wireless communication link by using uplink CoMP macro-diversity reception techniques. These techniques were applied for better channel state information (CSI) measurement and interference cancellation and worked by combining signals of jointly received UE transmitted signals at multiple geographically separated antennas [42]. The collaboration of cells in such a CoMP technique can be implemented as either intra-site or inter-site. Inter-site coordination is performed using cross-site cooperation of the signals that were jointly received at different sites through the Xn interface. In such a CoMP architecture, the serving cell is responsible for performing the radio resource management (RRM) and, based on the UE measurements, takes the decision to do macro-diversity reception. However, such a CoMP feature needs advanced processing capabilities at the lower layers of the base stations for combining the signals. Our proposed architecture intends to keep the base physical layer calculations and operations to a minimum to support deployment in remote radio units (RRUs) of modern-day distributed base stations.

2.6.2 Uplink Enhancement Technologies from 3GPP

A variety of uplink enhancement technologies have been defined in the latest 3GPP releases, like increasing the transmit power of the terminal, introducing long-PUCCH (Physical Uplink Control Channel), and leveraging multiple frequency bands for uplink enhancement. Dual-Connectivity in 5G has been proposed as one of the methods to deliver better uplink capacity and reliability. Using EN-DC with non-standalone architecture does improve the 5G uplink throughput compared to 4G, but it is still lower than the 5G standalone architecture. Moreover, due to limitations like antenna design complexity and low transmission power of UE, 5G terminals can generally have only two transmission channels for uplink. Thus, with EN-DC, there is only one transmission channel for NR, which makes dual-stream NR transmission on uplink difficult. Similarly, UE can also establish dual connections with two NR base stations, better known as NR-DC architecture, as discussed in more detail by authors in [39]. However, this would require the UE protocol stack to be modified to have a dual NR-capable stack increasing its complexity.

Carrier Aggregation (CA) technology has been used in the downlink to improve downlink throughput commercially in the past. It can also be used to improve uplink throughput by aggregating spectrum resources so that they can use multiple transmission channels of UE at the same time. NR CA [12] has been included since 3GPP Rel-15. Intra-band CA can aggregate multiple carriers in the same operating band and improve the user uplink experience. However, the throughput of inter-band CA that aggregates the carriers of different operating bands can be limited in many cases by the number of transmission channels of terminals. Additionally, CA technology can only be used to improve the capacity and latency of the system but would not improve reliability.

Supplementary uplink (SUL) was introduced in 3GPP Rel-15 to extend the uplink coverage by providing an additional uplink channel [14]. In SUL technology, along with an NR frequency band for downlink, two uplink frequency bands are configured in the same cell, one is an NR frequency band, and the other is the SUL frequency band, usually in the sub-3 GHz band. In SUL, the UE can dynamically select either the NR or SUL band for UL data transmission and uses the SUL carrier for transmitting data only when the UE is moving beyond the uplink coverage of the NR carrier. SUL technology improves the uplink coverage by using sub-3 GHz bands for uplink transmission but is obtained by sharing the spectrum with a 4G network. Therefore, 5G must be co-sited with 4G, which limits the flexibility of 5G deployment and brings new problems to network deployment.

To improve the uplink capacity and spectrum resource utilization efficiency in multi-carriers networking, uplink Tx Switching was introduced in 3GPP Rel-16 [13]. It uses one transmitting channel either for carrier one or carrier two and uses the other transmitting channel exclusively for carrier two. By using uplink Tx Switching, the uplink performance of all the above enhancement technologies can be improved. However, supporting technologies like dual connection, SUL, uplink Tx Switching, etc., add to the complexity of the devices and raises their cost and power usage. The technique suggested in this work improves the uplink performance with almost no change to UE complexity.

Chapter 3

Mobility Enhancement using Multi-Connectivity in 5G Networks

In this chapter, we explore how the 5G and beyond 5G network architecture would be impacted by mobility. We propose and evaluate the new architecture under various mobility scenarios to compare its performance with the current architecture. Section 3.1 describes the problem this chapter focuses on. The experimental methodology with the testing and simulation setup is explained in Section 3.2. Section 3.3 presents the experimental design used for our evaluation along with the results.

3.1 Mobility Issues in Current Architecture

5G networks and beyond are key to meeting the exponentially increasing demands of next-generation services for high throughputs and ultra-reliable low-latency communication. However, under mobility scenarios, all cellular networks are subject to handovers and demonstrate increased latency due to the additional time needed to execute them. Additionally, high mobility further adds to the latency because of the network overhead associated with control signals related to handover management.

Ultra-dense networking (UDN) is a promising technique for deploying 5G networks and is characterized by their high density of base stations (BSs). Although these BSs have limited coverage due to dense deployment, it also increases the probability of a line-of-sight (LoS) transmission. Millimeter-wave communications can be used to achieve high data rates and simultaneously mitigate the interference levels in UDNs. However, mmWave links have a lower transmission range and are subject to blockage. In mobility scenarios, the densification of base stations in UDN introduces challenges like repeated handovers that result in increased latency. Deploying small cells along with macro base stations is another potential solution for increasing traffic, but such an architecture leads to low cell resource utilization and adds to infrastructure deployment costs due to an imbalanced network load.

Multi-connectivity, on the other hand, is seen as a key approach to improving the reliability of next-generation wireless systems. Simultaneous connectivity to multiple radio access technologies or multiple bands represents an important solution for ongoing 5G deployments. Multi-connectivity enhances the system performance by providing multiple radio connections. When used in a split bearer model, the data is transmitted using two radio paths, thus, increasing the throughput of the system. Initial 5G deployment used the non-standalone architecture [11] where E-UTRA-NR Dual Connectivity (EN-DC) was used to improve the user throughput while maintaining good coverage. However, in EN-DC, all the control signaling is handled by the LTE RAN connected to the 4G core. This limits the implementation of many new 5G features that could potentially improve the network performance. 3GPP standards also propose the New Radio Dual Connectivity (NR-DC) feature containing high band NR small cells overlaid by sub-6 GHz FR1 band 5G macro-cells connected to the 5G core for effective deployment. Hence, the UE is connected simultaneously to two gNBs, with the sub-6 gNB acting as the master Node and the mmWave gNB acting as the secondary node. Thus, potentially when the mmWave base stations go out of range, the packets are momentarily delivered using the sub-6 GHz link rather than getting dropped and improving the reliability along with the throughput as a result.

Hence, this chapter tries to study the problem of increased latency and low reliability associated with mobility. We first analyze a standard standalone 5G UDN deployment operating at 28 GHz mmWave range. This chapter then proposes multi-connectivity as a potential solution to increasing the reliability and improving the latency of 5G networks. In this regard, we implement and evaluate the NR-DC architecture.

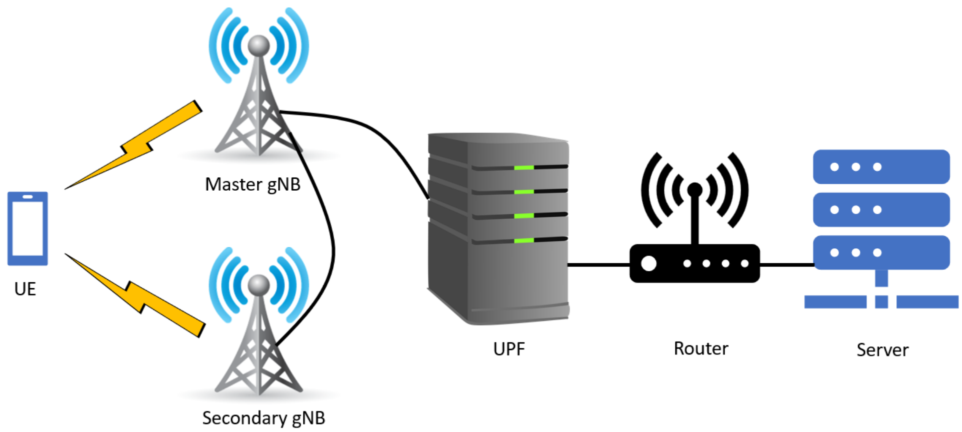


Figure 3.1: System Model for 5G NR-DC Deployment

3.2 Multi-Connectivity Environment and Simulation Methodology

Our proposed multi-connectivity architecture for an NR-DC deployment provides the UE with both low-band and high-band channels simultaneously, as shown in Figure 3.1. We deploy an NR-DC architecture with the master node as a 5G NR macrocell operating in the sub-6 GHz frequency and the secondary node as a 5G NR small cell operating in the mmWave frequency. The UE is capable of simultaneously connecting to both the master node and the secondary node. The master nodes are connected to the core network and facilitate the dual connectivity configuration based on the reported UE channel measurements. NR-DC can assure higher data rates, higher reliability, and lower delays, which are critical for emerging applications.

To simulate and study the performance of such a system, we use Simu5G, a 5G simulation library based on the OMNeT++ framework. Figure 3.2 shows the NR-DC deployment using Simu5G. Using Simu5G, we simulate the data plane of an end-to-end standalone 5G network that transmits packets from a remote server via the User Plane Function (UPF) in the 5G core to a UE that is connected to a gNB. The Simu5G model was extended to allow NR-NR Dual Connectivity (NR-DC) deployment, where two 5G gNBs could coexist. In the extended NR-DC configuration, one of the gNB works as a Secondary Node (SN) for another gNB, which acts as Master Node (MN) and is connected to the 5G Core Network. The two gNBs are connected through the X2 interface, and all traffic needs to go through the master gNB. The data packets that are destined to a UE being served

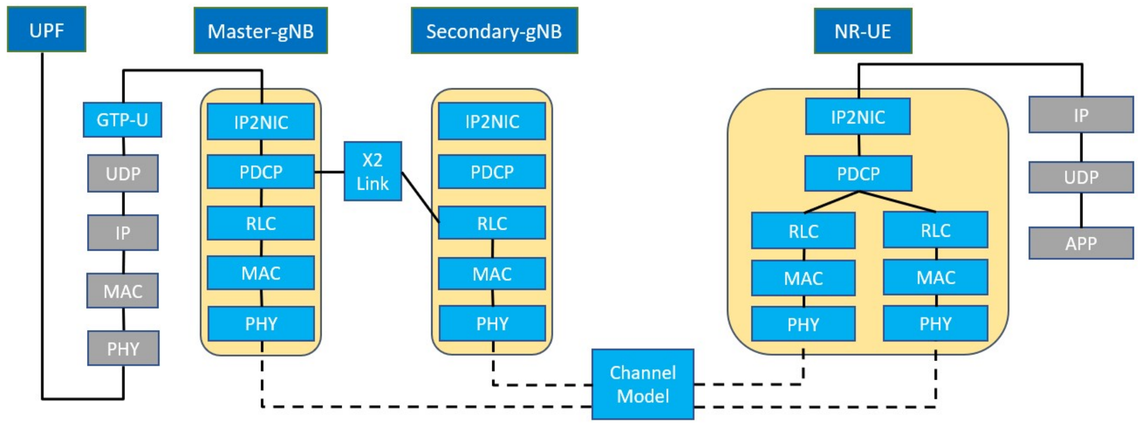


Figure 3.2: Simu5G NR-DC Deployment

by the master gNB in the Master Cell Group (MCG), use the MCG bearer and follow the usual NR protocol stack. On the other hand, data packets that are destined to a UE served by the secondary gNB in the Secondary Cell Group (SCG) use the SCG bearer and are first routed to the NR PDCP entity at the master gNB and then transferred to its peering NR RLC entity in the secondary gNB, via the X2/Xn interface. As per the 3GPP standards, Simu5G also supports Split Bearers (SBs), where incoming data packets are distributed between the MCG and the SCG bearers. With this feature, data belonging to the same connection can be delivered either through the master gNB or the secondary gNB. In our split bearer implementation, data packets are split equally among the two bearers i.e., half of the packet will travel through the master gNB, and the other half would use the secondary gNB. The UE side has a dual NR stack with two instances of the PHY, MAC, and RLC layers. The common PDCP layer at the UE side is responsible for reordering the PDUs coming from the two RLC layers.

We also modify the channel model from Simu5G based on the 3GPP study of the channel model for frequencies from 0.5 to 100 GHz. An NR-NR handover is also introduced by enhancing the existing LTE handover functionality of Simu5G to account for gNB to gNB handover in standalone 5G deployment. The motivation behind this work is to better understand network behavior for the specialized use case where the requirement is to achieve low latency and high reliability, similar to URLLC use cases, along with achieving high throughputs. This type of use case is becoming more and more common with modern applications. Further, this dissertation focuses on evaluating network performance in mobility scenarios from two aspects. The first aspect looks at the UE performance in a city-like environment, and the second looks at how the network scales when the number of UEs is increased. In this respect, we test for up to 64 UEs based on the fact that authors in [37] discuss 3GPP 2A deployment scenarios that average around 30 UEs per macro-cell area. The next section thus goes into the details of our experimental setup to evaluate this and also presents our results.

3.3 Evaluation

To study the performance of NR-DC network architecture in a city-like environment under mobility, a systematic evaluation of the architecture was conducted. In our experiment, we consider downlink transmission, where a server is connected to the UPF in a 5G standalone deployment via a router and sends data packets to mobile UE(s) in our system. The UE depending on the type of deployed architecture may be connected to one or more gNBs. We run a UDP type of application at the server, and a UDP client runs at the UE. The server application sends UDP data packets destined for the UE. Once these packets are routed to the UPF, the traffic filter deployed at the UPF looks at the source and destination addresses in the IP header to find out the IP of the destination UE. The packet is then encapsulated using the GTPU protocol and tunneled from the UPF to the gNB to which the destination UE is currently attached. The gNB, on receiving this packet, will transmit it to the UE over the air using the radio resources allocated to it. For our experiments, we vary the radio access network design and characteristics based on the network architecture that is being evaluated. We also vary the mobility pattern and speed of the UE based on the type of experiment we are simulating.

3.3.1 Performance Metrics

To measure the performance of the concerned network architecture, we primarily use the following three metrics:

1. **Throughput**: This metric refers to the ratio of data transmitted in a unit of time and is represented in bits per second or bps. Higher throughput would mean better network performance. Throughput is a critical measure for eMBB type of applications. But, it is also expected that throughput levels for a network remain consistent over various mobility scenarios for URLLC-type applications. In our experiment results, we present the accumulated average

throughput of all the successfully received packets and not the instantaneous throughputs of the received packets.

2. **Latency**: Network latency is the total time, usually measured in milliseconds, required for a server and a client to complete a network data exchange. In our experiments, we measure the one-way latency between the applications, i.e., the time taken for a packet to reach the client from the server. Lower latencies reflect better network performance and are a critical measure for URLLC-type applications.

3. **Reliability/Packet loss**: This metric is a measure of whether the delivery of data to intended recipients was successful or not. The reliability metric is complementary to the packet loss metric, which is a ratio of the number of packets that were dropped to the number of packets that were transmitted. They are usually measured in percentages. Good networks are expected to have high reliability or a low packet loss ratio. In our experiments, we present the packet loss ratio metric for evaluation.

3.3.2 Experimental Setup

In this chapter, we perform experiments to study the effect of multi-connectivity, especially NR-DC, on the network performance when compared to a UDN type of deployment. As we want to simulate a city-like environment, we consider a mobile UE with a mobility pattern that would be similar to a user moving around blocks in the downtown area of a city. For the UDN deployment to simulate the path loss and interference patterns that would be observed in a city downtown with tall buildings, we use the Urban Microcell path loss model from the 3GPP specifications. We also calculate and use line-of-sight probabilities in the simulation for better accuracy. We primarily run two experiment sets where the first one looks at the comparison between the UDN and NR-DC network's performance observed by a single UE moving in a pattern that simulates city grid-like patterns. The second set of experiments studies the impact of increasing load on the network by testing the network performance with multiple randomly moving UEs in a specific area.

In the UDN or dense network type deployment, we have multiple gNBs covering the simulation area, each operating in the 28 GHz range. The transmit power for each mmWave gNB is fixed at 35 dBm. In the NR-DC deployment, we deploy a particular number of master-secondary pairs across the simulation area. In NR-DC, the master gNB operates at 2 GHz in the sub-6 fre-

Table 3.1: Simulation parameters for UDN and NR-DC deployments used during optimization

Parameter	**Value**
mmWave Carrier Frequency	28 GHz
mmWave Bandwidth	100 MHz
Sub-6 GHz Carrier Frequency	2 GHz
Sub-6 GHz Bandwidth	20 MHz
UDP Packet Size	1024 B
UDP Packet Inter-arrival Time	1 ms and 10 ms

quency range and at a 49 dBm transmit power level, whereas the secondary gNB operates at 28 GHz in the mmWave range and transmits at a 35 dBm power. For an accurate comparison between scenarios, we use the same number of total gNBs in the NR-DC deployment as we do in the UDN scenario. For example, if the UDN experiment is conducted with six gNBs in the area, then the corresponding NR-DC deployment for comparison would have an arrangement with three NR-DC master-secondary gNB pairs giving a total of six gNBs. In other words, the Master Cell Group (MCG) consists of three gNBs, and the Secondary Cell Group (SCG) consists of three gNBs. The server UDP application generates packets of size 1024 bytes. We run experiments for two different packet inter-arrival times of 1 ms and 10 ms. The details of the simulation parameters that were used in the experiment are provided in Table 3.1. The packet size and packet inter-arrival times are chosen to evaluate the network performance for an application use case that needs low latency and reliable communication while providing good throughput. The 10 ms packet inter-arrival time allows a maximum throughput of around 0.8 Mbps with a packet size of 1024 B, whereas when the packet inter-arrival time is 1 ms, it increases the flow of data allowing the system to achieve a maximum throughput of nearly 8 Mbps.

As mentioned earlier, we broadly run two sets of experiments. Both sets of experimental results are compared by testing UDN deployment against NR-DC deployment in an environment that simulates the downtown of a city. It assumes tall buildings with streets running between them. The first set of experiments is conducted using a single UE that is moving along the city grid. The mobility pattern from OMNeT++ simulation library used for this experiment is *TurtleMobility*,

Table 3.2: gNB Coordinates for UDN deployment in 1000 m and 1500 m scenarios for 4 and 6 gNB cases

UDN Deployment				
Scenario	**1000 m x 1000 m**		**1500 m x 1500 m**	
gNB#	**4 gNBs**	**6 gNBs**	**4 gNBs**	**6 gNBs**
gNB1	(250 m, 250 m)	(166 m, 250 m)	(375 m, 375 m)	(250 m, 375 m)
gNB2	(250 m, 750 m)	(166 m, 750 m)	(375 m, 1125 m)	(250 m, 1125 m)
gNB3	(750 m, 250 m)	(500 m, 250 m)	(1125 m, 375 m)	(750 m, 375 m)
gNB4	(750 m, 750 m)	(500 m, 750 m)	(1125 m, 1125 m)	(750 m, 1125 m)
gNB5	NA	(833 m, 250 m)	NA	(1250 m, 375 m)
gNB6	NA	(833 m, 750 m)	NA	(1250 m, 1125 m)

where we start at a fixed point and move up to a given point at a random speed chosen uniformly between 10 m/s to 40 m/s in our case. On reaching this intermediate point, the UE in our simulation takes a 90-degree clockwise turn and moves to the next end-point, again with a uniformly selected random speed. This pattern repeats till the end of the simulation that is run for a fixed amount of time. We repeat each set of experiments ten times and average the results.

For the second set of experiments, we test the two deployments for the network load they can handle. We basically evaluate the network performance by slowly increasing the number of UEs present in the simulation. We start with 2 UEs and increase the number of UEs to 4, 8, 16, 32, and 64. In each of these experiments, the UEs are moving randomly using an OMNeT++ defined mobility pattern called *RandomWaypointMobility*. The path loss and interference models are still assumed to be that of a city-like environment. In this type of mobility, the simulation randomly selects the start point and an endpoint for the UE. The UE then travels to this endpoint at a speed that was randomly selected. On reaching the end-point, the model picks the next endpoint and another random speed to move the UE there. In our experiments, we pick the speed of the UE randomly using a uniformly distributed random variable between 1 m/s and 40 m/s.

Table 3.3: Master gNB (M-gNB) and Secondary gNB (S-gNB) Coordinates for NR-DC deployment in 1000 m and 1500 m scenarios for 4 and 6 gNB cases

NR-DC Deployment				
Scenario	**1000 m x 1000 m**		**1500 m x 1500 m**	
gNB#	**4 gNBs**	**6 gNBs**	**4 gNBs**	**6 gNBs**
M-gNB1	(250 m, 250 m)	(166 m, 250 m)	(375 m, 375 m)	(250 m, 375 m)
S-gNB1	(250 m, 251 m)	(166 m, 251 m)	(375 m, 376 m)	(250 m, 376 m)
M-gNB2	(750 m, 750 m)	(500 m, 750 m)	(1125 m, 1125 m)	(750 m, 1125 m)
S-gNB2	(750 m, 751 m)	(500 m, 751 m)	(1125 m, 1126 m)	(750 m, 1126 m)
M-gNB3	NA	(833 m, 250 m)	NA	(1250 m, 375 m)
S-gNB3	NA	(833 m, 251 m)	NA	(1250 m, 376 m)

Each experiment set tests the two network deployments in two different area scenarios. As discussed, the first set of experiments is conducted with a single UE, and the second set of experiments with Multiple UEs. Each set of experiments compares the UDN deployment with the NR-DC deployment. The first scenario considered has a 1000 m by 1000 m simulation area which we often refer to as the 1000 m scenario in this dissertation. And the second scenario considers a 1500 m by 1500 m simulation area and is referred to as the 1500 m scenario. In each scenario, we further consider two cases, one with four gNBs and the other with six gNBs. Table 3.2 and 3.3 gives the coordinates for the gNB deployment in these UDN and the NR-DC deployments, respectively. One of the major issues with the performance of UDN-type deployments is that they are subjected to repeated handovers. To evaluate how NR-DC improves upon the number of handovers happening in the system, we also keep track of the number of handovers in both UDN and NR-DC experiments.

3.3.3 Results for Single UE

The first set of experiments we conducted to show and compare the performance of mmWave-based UDN deployment against the NR-DC deployment. The NR-DC deployment for these exper-

Table 3.4: Results comparison for mmWave-based UDN and NR-DC deployments for a single UE in 1000 m scenario

Inter-arrival time		1 ms				10 ms			
Deployment Type		Throughput (Mbps)	Packet Loss (%)	Latency (ms)	No. of HOs	Throughput (Mbps)	Packet Loss (%)	Latency (ms)	No. of HOs
4 gNBs	**UDN**	4.77	41.83	116.62	4	0.75	8.29	9.35	3
	NR-DC	6.42	21.64	16.17	2	0.79	3.83	8.65	2
6 gNBs	**UDN**	5.83	28.85	34.38	6	0.77	6.44	8.78	6
	NR-DC	7.06	13.90	10.59	3	0.80	2.26	7.35	4

iments used a split bearer configuration that splits the incoming packets equally among the two gNBs. The three performance metrics of throughput, packet loss, and latency were measured in each case. We also keep a track of the handovers that the UE goes through during the simulation. We present these results in Table 3.4 for the 1000 m by 1000 m area simulation and in Table 3.5 for the 1500 m by 1500 m area simulation. Results are shown in each case for both 1 ms and 10 ms packet inter-arrival times. Also, we show results for two types of deployment strategies one with 4 gNBs and the other with 6 gNBs.

3.3.3.1 NR-DC performs better than mmWave-based UDN deployment

From all the results in Tables 3.4 and 3.5, we observe that NR-DC deployment performs consistently better than the mmWave-based UDN deployments. All the performance metrics are seen to have improved because of using NR-DC configuration as it uses both a low latency high throughput mmWave link as well as a reliable sub-6 GHz link. We observe that for the 1000 m scenario with 4 gNBs when packet inter-arrival time is 1 ms, the latency of the system is reduced from 116.62 ms in UDN deployment to 16.17 ms using NR-DC deployment. In the same scenario, the packet loss improves from 41.83% in UDN to 21.64% in NR-DC deployment. There is a corresponding improvement in throughput as well. We notice similar improvements in 6 gNB deployments as well.

From the results, we notice that as we increase the simulation area, we see a degradation

Table 3.5: Results comparison for mmWave-based UDN and NR-DC deployments for a single UE in 1500 m scenario

Inter-arrival time		1 ms				10 ms			
Deployment Type		Throughput (Mbps)	Packet Loss (%)	Latency (ms)	No. of HOs	Throughput (Mbps)	Packet Loss (%)	Latency (ms)	No. of HOs
4 gNBs	**UDN**	4.37	46.66	142.70	4	0.66	19.51	14.87	4
	NR-DC	6.35	22.51	23.61	3	0.78	5.68	13.31	2
6 gNBs	**UDN**	5.30	35.32	121.68	8	0.73	10.68	9.97	7
	NR-DC	6.55	20.08	19.61	3	0.78	5.65	9.12	4

in the network performance. This is true for both UDN and NR-DC type deployments, however, we also notice that the degradation is more significant in the case of UDN as compared to NR-DC deployments. For example, we notice that in the 1000 m scenario, the packet loss in the 4 gNB case for packet inter-arrival time of 10 ms is 8.29% for UDN deployment and 3.83% in the case of NR-DC deployment. However, when we look at similar metrics in the case of the 1500 m scenario, we see that the packet loss percentages in the case of UDN have gone up to 19.51%, a rise of little more than 11%, whereas in the case of NR-DC that loss percentage only went up-to 5.68%, a rise of only 1.85%. Similar trends are seen for all other scenarios as well. But we do see that these rise in the loss percentages and latency are more pronounced in the case of 4 gNBs as compared to 6 gNB deployments. The reason for such a rise is the propagation characteristics of wireless signals. Due to the consistent and reliable path loss attributes of the 2 GHz frequency range over long distances, we see that the performance is very well maintained even if we increase the simulation area to 1500 m. On the other hand, the limited propagation characteristic in the mmWave frequency range results in even poorer performance when the area is increased to 1500 m.

Also, it must be noted that in an NR-DC deployment, when the secondary mmWave cell goes out of coverage and detaches itself from the UE, the master starts to deliver all those packets by using its own resources. This allows for the packets to continue to be delivered rather than get dropped, as in the case of UDN-type deployments. This is the reason we see lower losses in NR-DC as compared to UDN deployments. Also, as in the 10 ms packet inter-arrival time case, the load on the network is lower the improvement seen is also less compared to the 1 ms case. This, however, also leads to the fact that we see more packets lost in NR-DC deployment when the packet inter-arrival time is smaller, like in the case of 1 ms. When there is a large amount of incoming packet flow, then the master struggles to schedule all the packets in its buffer using the limited radio resources that it has. This slowly leads to buffer overflows, and eventually, drops are seen, leading to packet loss.

3.3.3.2 Performance largely depends on Inter-gNB distance

As NR-DC deployment tries to utilize the best of both mmWave and sub-6 frequency ranges, the propagation characteristics of both frequency ranges have an impact on the network performance. However, as our design focuses on utilizing the high bandwidth and lower latency mmWave link, the performance is also heavily tied to the inter-gNB distance in the simulations. This trend is expected and observed in the UDN deployment, but we also see there is a similar exact trend for the NR-DC deployments as well. When we analyze the results from both the packet inter-arrival times, we see that the trend is similar. We see that the 4 gNB case in the 1500 m scenario has the worst performance, whereas the performance of the 6 gNB case in the 1000 m scenario results in the best performance. This is because of the inter-gNB distance difference in the two cases. The 4 gNB case in the 1500 m scenario has an inter-gNB distance of 750 m, and for the 6 gNB case in the 1000 m scenario, we observe a distance of 333 m and hence we see the difference in performance.

Likewise, if we compare the 6 gNB case in 1500 m and the 4 gNB case in 1000 m scenarios we observe very similar performance results as in both these scenarios the inter-gNB distance happens to be 500 m. For both packet inter-arrival times, we see results that performance results are very close to each other, but on closer examination, we notice that the 6 gNB case in 1500 m performs slightly better in terms of packet loss but worse in terms of latency when compared to the 4 gNB case in 1000 m scenario. The reason for this variation seems to be the availability of two extra gNBs in the 1500 m scenario. Although the result is the same inter-gNB distance, the two extra gNBs provide additional resources to handle the incoming load. As the experimental scenario is designed to have the UE move continuously in the area, the UE serving cell distribution will be uniform among the available base stations. This will lead to the packets being distributed uniformly among the base stations, with each gNB having its own MAC buffer to schedule packets this eventually results in more packets getting delivered and lower packet loss. This also, however, means that some packets have to stay longer in the buffer waiting to get delivered. This is why we see lower packet loss but higher latency in more gNB cases.

3.3.3.3 Adding more gNBs improves the performance

Another conclusion we could derive from the results is that adding more gNBs to a deployment scenario improves the performance of the NR-DC deployments. This is expected because mmWave base stations play a big role in the performance of our NR-DC deployments, and as more gNBs are added, the inter-gNB distances decrease. The mmWave frequency range propagation characteristics result in a significant performance decrease when the inter-gNB distances are larger than 300 m. We observe that for 10 ms packet inter-arrival time in the 1000 m scenario, 4 gNBs case results in a packet loss of 3.83% and latency of 8.65 ms, but when we increase the number of deployed gNBs in the area to 6, we see an improvement in performance to 2.26% packet loss and decrease in latency to 7.35 ms. The same observation is seen for all the other cases as well. Thus the reason why the 6 gNB case in the 1000 m scenario gives the best network performance is that it has the lowest inter-gNB distance of the test scenarios. Although we could start to see a decline in this performance if we start to increase the inter-gNB distance indefinitely. Eventually, we would see the performance improvements would stagnate, and they could get worse if the gNBs are too closely located because of unnecessary handovers in such scenarios.

3.3.3.4 Network load can impact the performance significantly

A major factor that determines the network performance is the resources allocated to it. Factors like spectrum, bandwidth, buffer sizes, etc. all play a major role in determining this performance. Propagation characteristics at different frequencies are something we have already discussed. Similarly how much bandwidth is allocated to each gNB would decide the amount of data that can be scheduled in each transmission time interval. This would eventually mean how fast the gNB is able to clear its buffer and have an impact on the latency of the packet. This also determines how quickly the MAC buffer can be cleared by scheduling all the packets that get into it. If the bandwidth is low, then the buffers will start filling up, and a small buffer would lead to packet drops and poor performance.

However, all these factors come into play only if we have enough packets to stress the system to reach that point. If the data flow is too slow, then a smaller bandwidth and low buffer size would be enough to schedule the transmission. Only when the packet inflow is faster, then the issues will be seen in the system. This is the reason we test the system under two load sizes. A packet inter-arrival time of 10 ms is the low load condition, and that of 1 ms is the larger load condition. We also see the same in the results, as there is a significant increase in the packet loss and latency numbers between the two cases. The ideal load conditions would lie somewhere in between these two cases. Increasing the bandwidth would lead to better performance, but we must keep in mind that radio resources are scarce in the sub-6 GHz range and very hard to come by. Thus allocating large resources all the time would be infeasible. Similarly having large buffers could also lead to issues like buffer bloat in the network and thus degradation in network performance. We see that for the packet inter-arrival time of 1 ms, we have significantly more throughput of 7.06 Mbps in the best-case scenario as compared to the 10 ms case of 0.80 Mbps. Although we do see that the bandwidth of 100 MHz leads to an increase in packet loss from 2.26% in the 10 ms case to 13.90% in the 1 ms packet inter-arrival time case.

3.3.3.5 Handovers performed decrease in NR-DC deployment

Another important aspect of UE mobility performance that we observe in the results is the improvement in the performed handovers. We see in almost all the cases see a decrease in required handovers while performance is maintained throughout. The number of handovers happening depends on the number of gNBs that have been deployed in that scenario. The 6 gNB cases have more handovers happening when compared to the 4 gNB cases. We see that for the 1500 m scenario in the 6 gNB case for packet inter-arrival time of 1 ms, there are 5 fewer handovers done in NR-DC deployment compared to the UDN deployment. The number of handovers is less as we have a wider coverage of the sub-6 GHz range.

3.3.4 Results for Multiple UEs

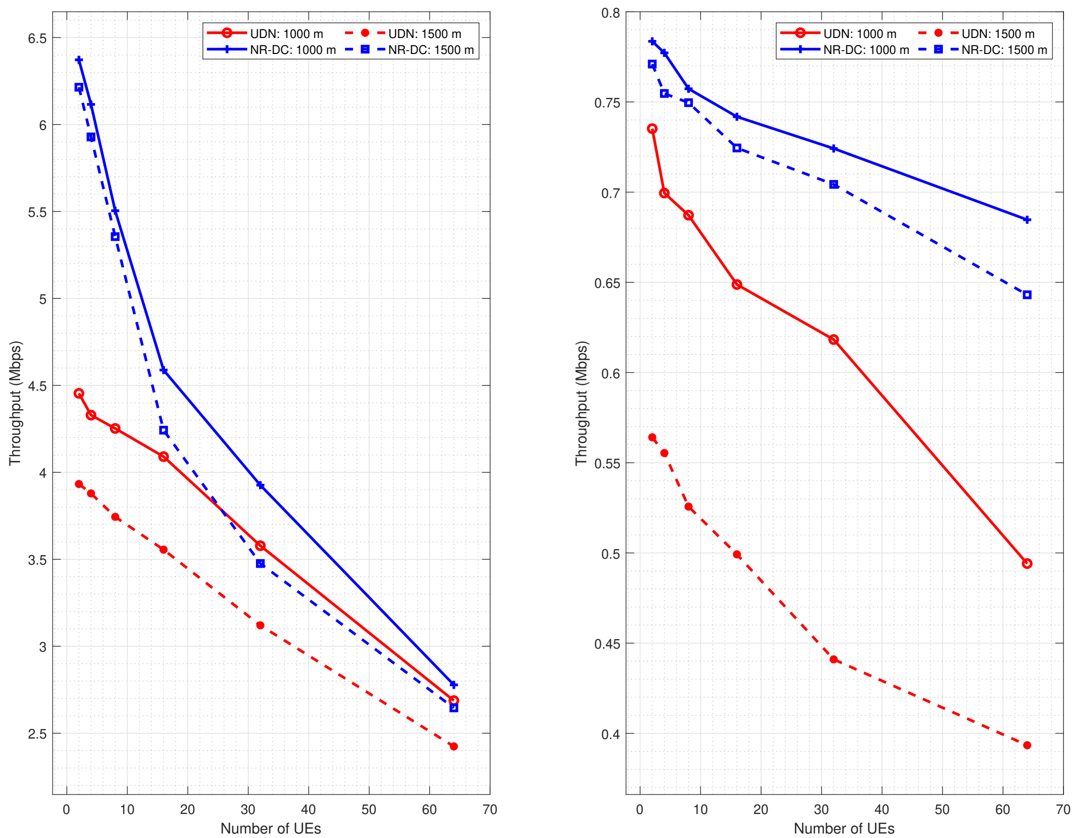
The second set of experiments we conducted was to show and compare the performance of the NR-DC deployment when the network load in terms of users increased. The NR-DC deployment used for these experiments again was a split bearer configuration that splits the incoming packets equally among the two gNBs. The three performance metrics of throughput, packet loss, and latency were measured in each case. In the case of latency, we report the average of the latency observed by all the UEs. We also show the minimum and maximum latency that is observed by a UE in that group. The throughput and packet loss shown are averaged across the UEs. We also keep a track of the handovers that the UEs go through during the simulation and present the total number of handovers that happened during the simulation period.

We present the throughput, packet loss, and latency results for multiple UEs in Figures 3.3, 3.4 and 3.5, respectively. The results are shown for the 1000 m by 1000 m area and 1500 m by 1500 m area simulation with both 1 ms and 10 ms packet inter-arrival times. The experiments conducted consider a 4 gNB mmWave-based UDN-type and NR-DC deployments. We also compare the latency results by showing how the minimum, maximum, and average latency of the system changes with the number of UEs in the system. Figures 3.6 and 3.7 show these results for the 1500 m and 1000 m cases respectively. Finally, we present the result for the number of handovers taking place in the system in Figure 3.9.

3.3.4.1 NR-DC shows better performance

We observe that NR-DC continues to perform better than UDN in the case when there are multiple UEs in the system as well. We see in the case of packet inter-arrival time of 10 ms the NR-DC deployment always performs better than the mmWave-based UDN deployment in terms of having lower latency, lower packet loss, and better throughput. Similarly in the case of the packet inter-arrival time of 1 ms as well, the NR-DC deployment shows better performance than the UDN deployment in terms of packet loss and throughput. However, as we increase the number of UEs in the system we see the performance of the NR-DC deployment start to decay very fast and the performance is comparable to that of the UDN deployments. The latency results for NR-DC deployment in both 1000 m and 1500 m scenarios show worse results compared to UDN deployments for packet inter-arrival time of 1 ms when there are around 16 or more UEs in the system.

We observe that the packet loss results in NR-DC deployment are significantly better than UDN for most cases except in the case of packet inter-arrival time of 1 ms with a higher number of UEs where they are comparable to each other. For example, in the 1000 m scenario, the loss is 22.21% in NR-DC compared to 45.62% in UDN deployment when there are two UEs in the system. Thus giving almost a 50% improvement in packet loss performance. However, these numbers for packet loss rise to 66.09% in NR-DC deployments and 67.19% in UDN deployment for 64 UEs in the system. Thus, the improvement in the system decreases to only 1.64% as compared to 50%. The reasons for this and the latency increase mentioned earlier are discussed in more detail in the



(a) Packet inter-arrival time of 1 ms

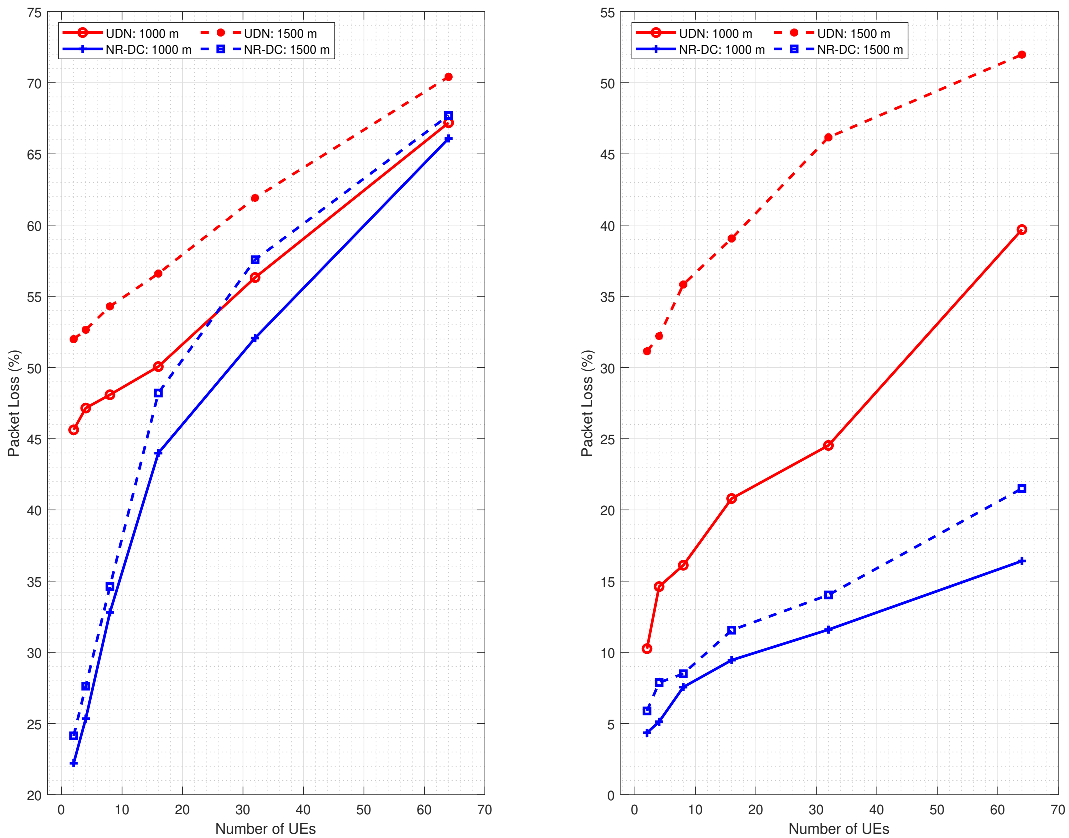
(b) Packet inter-arrival time of 10 ms

Figure 3.3: Throughput comparison of the mmWave-based UDN vs. NR-DC deployment for the 1000 m and 1500 m scenarios

upcoming paragraphs.

3.3.4.2 Performance decreases as the number of UEs served increases

We observe a constant decrease in all three performance statistics as we increase the number of UEs being served in the network. We observe a similar trend for both UDN-type and NR-DC deployment and across packet inter-arrival times. It is the range of the variation that differs in each case. For example, in NR-DC deployment for packet inter-arrival time of 10 ms we observe the best performance of 9.26 ms latency and 4.35% packet loss for the 2 UE systems in the 1000 m scenario. These results shoot up to 105.19 ms and 16.41% when the number of UEs served is increased to 64.



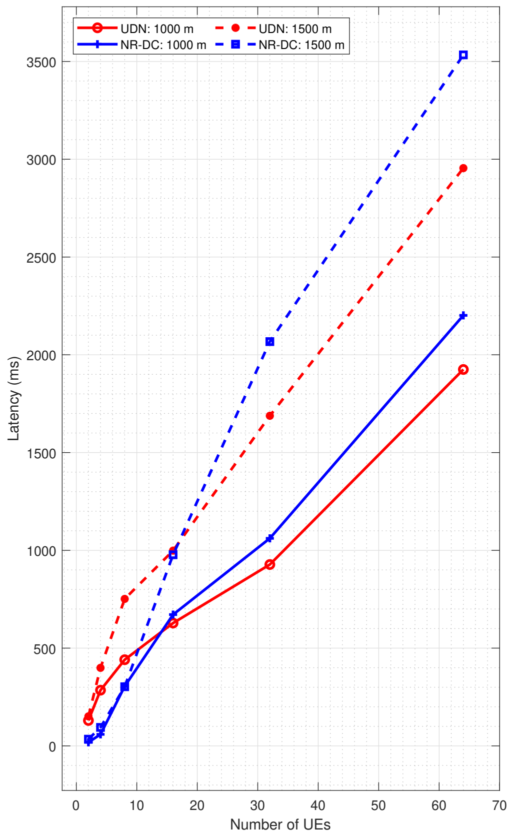
(a) Packet inter-arrival time of 1 ms

(b) Packet inter-arrival time of 10 ms

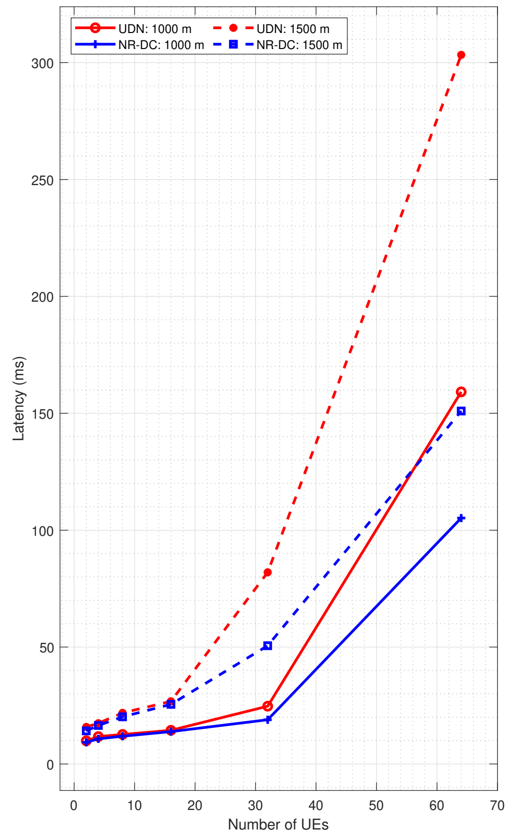
Figure 3.4: Packet loss comparison of the mmWave-based UDN vs. NR-DC deployment for the 1000 m and 1500 m scenarios

Whereas for the packet inter-arrival time of 1 ms in the same 1000 m scenario we see the lowest latency is 18.20 ms and packet loss is 22.21% for 2 UE. This increases to 2201.63 ms latency and 66.09% packet loss when we deploy 64 UEs.

However, the performance decrease is not exactly uniform in the case of NR-DC as we observe in the case of UDN deployment. Especially, in the case of packet inter-arrival time of 1 ms we see a very rapid decay in performance with an increase in UE. If we look at Figure 3.5a we can see a steady uniform rise in packet loss numbers of UDN deployment where as there is a sharp rise in the numbers of NR-DC deployment. Similar, observations are also seen in the case of latency and throughput as well. We also see this trend with the minimum and maximum latency results.



(a) Packet inter-arrival time of 1 ms



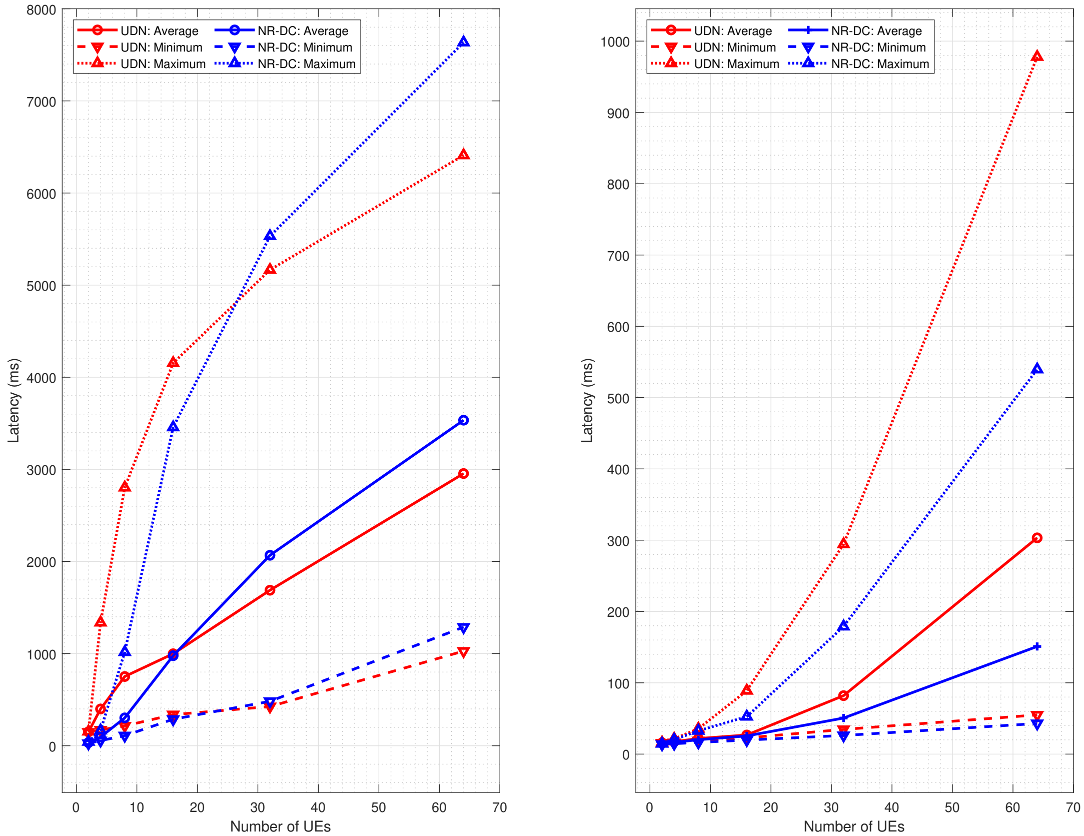
(b) Packet inter-arrival time of 10 ms

Figure 3.5: Average latency comparison of the mmWave-based UDN vs. NR-DC deployment for the 1000 m and 1500 m scenarios

However, the rise in minimum latency is very small from 2 UEs to 64 UEs where as the maximum latency in the system rises very fast. This can be better observed from Figures 3.6 and 3.7.

3.3.4.3 NR-DC latency for a large number of UEs increases rapidly

As discussed earlier the performance of NR-DC deployment degrades rapidly as the number of UEs in the system increases. Although the NR-DC performance in terms of packet loss stays better than UDN, the latency of the NR-DC system becomes larger than UDN deployments when there are 16 or more UEs in the system. The reason for such a behavior is the propagation properties and available bandwidth of the sub-6 GHz range. In NR-DC when mmWave base stations go out of



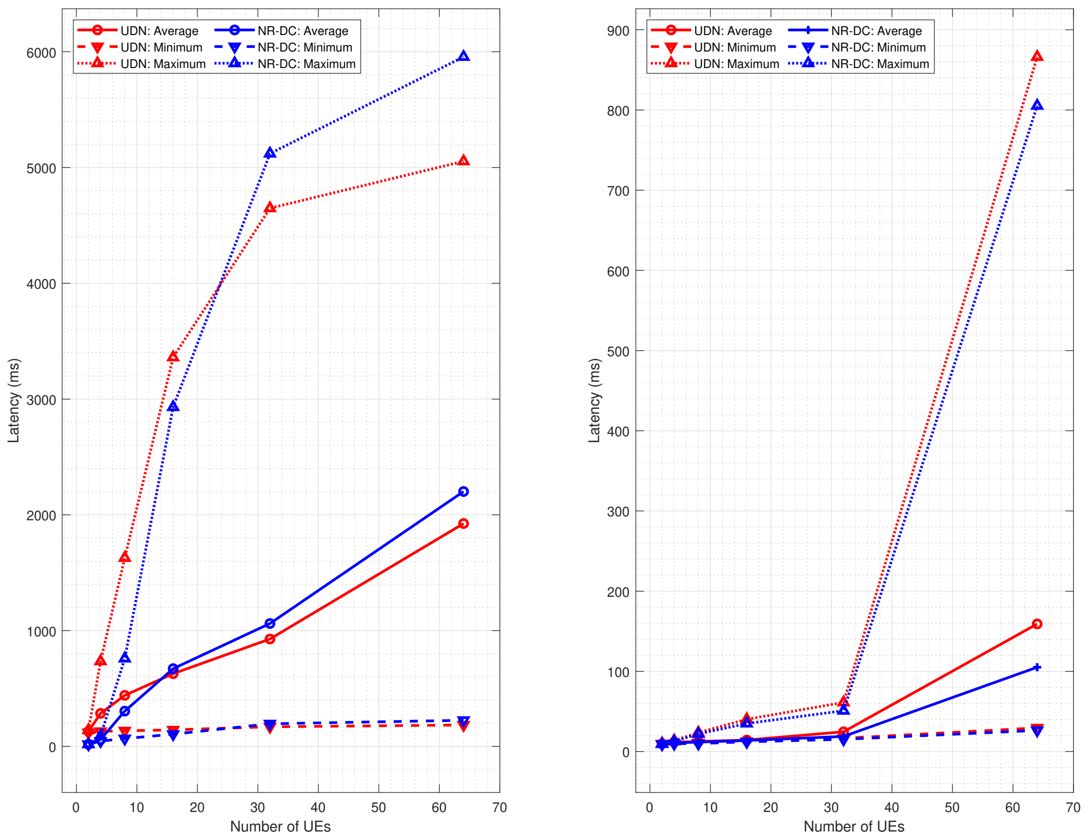
(a) Packet inter-arrival time of 1 ms

(b) Packet inter-arrival time of 10 ms

Figure 3.6: Average, minimum, and maximum latency results of the mmWave-based UDN vs. NR-DC deployment for the 1500 m scenario

range and communication falls back to sub-6 GHz base stations the connection to the UE improves in terms of quality. But, we have very limited bandwidth available at this range and thus sub-6 gNBs can only schedule a limited amount of data every TTI, which is much lesser than mmWave gNBs. Thus, packets stay longer in the buffer as the network slowly tries to deliver them, but when the number of UEs increases the load is too high to sustain good performance. This is the reason we observe high latency but as packets still are getting delivered more than the UDN deployments we see smaller packet loss percentages.

The effect we discussed above is very clearly seen by looking in Figures 3.6b and 3.7b at the maximum latency values for packet inter-arrival time of 10 ms case in both deployments. We



(a) Packet inter-arrival time of 1 ms

(b) Packet inter-arrival time of 10 ms

Figure 3.7: Average, minimum, and maximum latency results of the mmWave-based UDN vs. NR-DC deployment for the 1000 m scenario

can observe a sudden rise in the max latency from 51.00 ms in the 32 UE case to 805.28 ms in the 64 UE case for the 1000 m deployment scenario. A similar rise is seen in the 1500 m deployment scenario as well. This rise shows how the packets are getting delivered but significantly delayed. Although the maximum latency in UDN increases steadily, we see a sudden inflection point in the case of NR-DC where we see a sudden rise in maximum latency. This pushes the average latency results up and seems to happen around 8 UEs for a packet inter-arrival time of 1 ms and 32 UEs for a packet inter-arrival time of 10 ms.

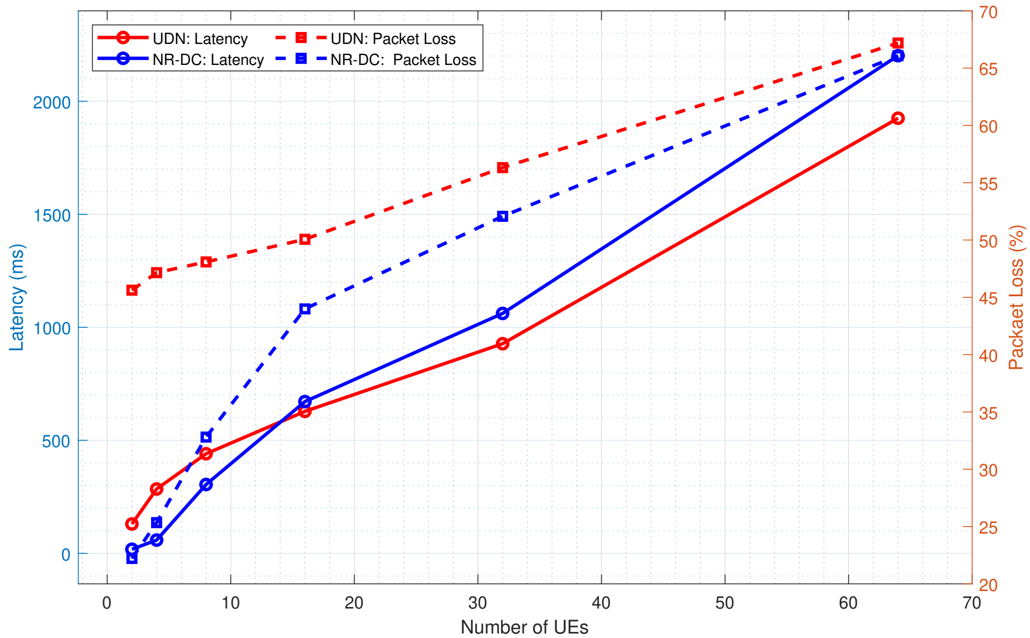


Figure 3.8: Comparison of average Latency and Packet loss results between NR-DC and UDN deployments for 1000 m scenario and packet inter-arrival time of 1 ms

3.3.4.4 NR-DC performance decreases with increase in network load

As we saw in the single UE experiments it is very important that we test our system under various load conditions. Although, increasing the number of UEs served in the system increases the amount of data that the system needs to handle but varying the packet inter-arrival time adds another dimension to the problem. With faster-arriving packets, it's critical that the network is able to deliver them equally fast, or else the network buffers can become saturated very quickly leading to packet drops.

When the number of UEs is large in the case of packet inter-arrival time of 1 ms the NR-DC deployment latency is high as compared to UDN because NR-DC has better coverage but lower bandwidth. As the MAC buffers are always full, even though the packets are getting delivered throughout the latency starts increasing. This is because the packets now sit longer in the buffers to be delivered. But, we also note that because they are delivered the packet loss stays better in NR-DC than UDN even for larger load conditions of packet inter-arrival time of 1 ms. This can be clearly seen in Figure 3.8 that shows this comparison. We plot the packet loss and average latency in the same graph with two different y-axes on either side. We also observe that the latency for packet inter-arrival time of 1 ms is larger than the latency when packet inter-arrival time is 10 ms in the UDN scenario as mmWave coverage is poor and a lot of packets need to be delivered and the MAC buffers are mostly full. This is the reason packet loss also increases when a packet inter-arrival time of 1 ms is used.

3.3.4.5 Performance degrades with larger inter-gNB distance

Another important observation we can make from these results is that the performance for multiple UEs also improves as the inter-gNB distance is decreased. We observe that the results of the performance metrics are all better for the 1000 m scenario than the 1500 m scenario as the inter-gNB distance is smaller in the previous case. We can observe this trend clearly in Figures 3.3, 3.4 and 3.5. Thus the effect of mmWave propagation characteristics has a clear and obvious effect on the performance of NR-DC deployments as well, but sub-6 GHz communication is able to handle poor coverage situations under low loads, thereby improving the performance considerably.

3.3.4.6 NR-DC deployment significantly decreases the number of handovers in the system

In terms of handovers, we observe a similar number of handovers for packet inter-arrival time of 1 ms and 10 ms cases as would be expected for each scenario. The number of handovers happening in each scenario for different packet inter-arrival times and areas are shown in Figure 3.9. We see more handovers in the 1500 m scenario as there would be a larger variation in RSSI and SINR. We observe that there is a considerable reduction in the number of handovers happening in the NR-DC deployment as compared to the UDN deployment. Handovers in the NR-DC deployment are almost half of that in the UDN scenario as we have master plus secondary pairs and there is handover happening from sub-6 Master gNB to another sub-6 master gNB compared to handovers between mmWave gNBs in UDN deployment.

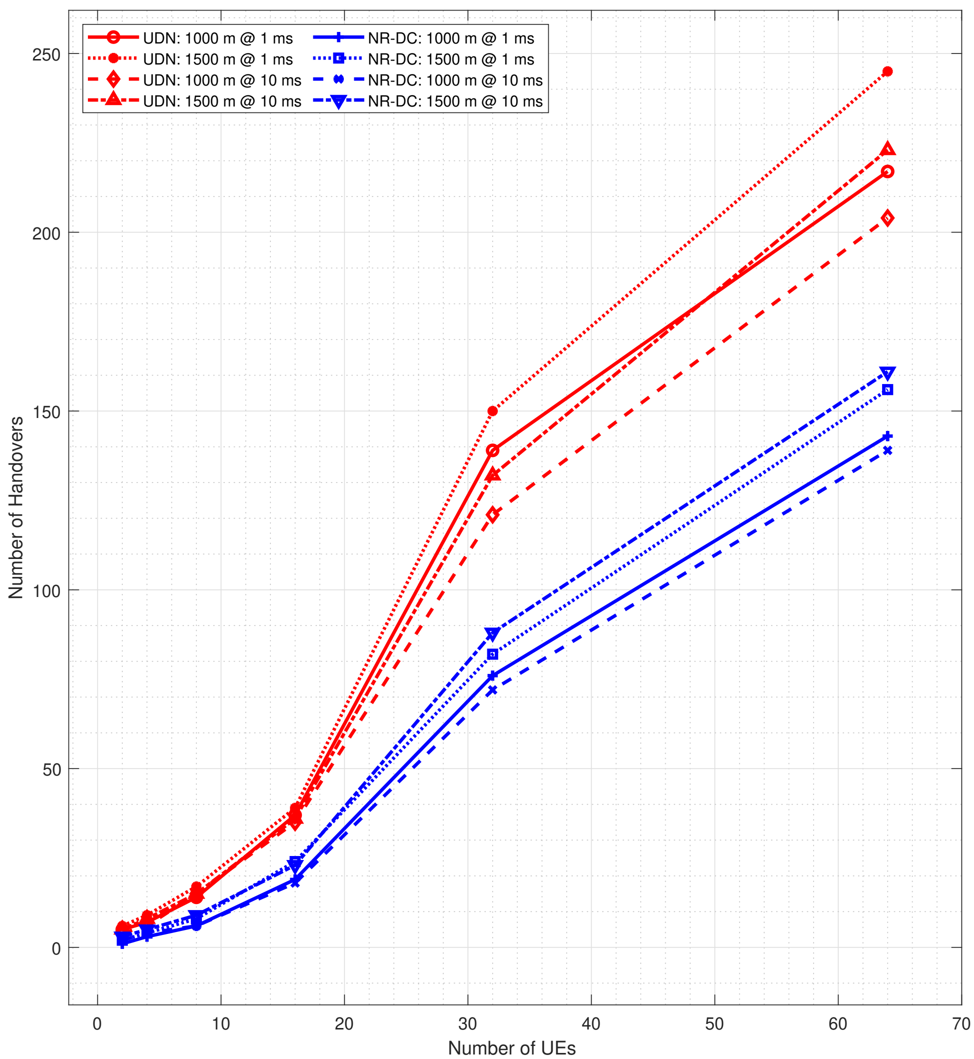


Figure 3.9: Comparison of the total number of handovers in NR-DC and UDN deployments

Chapter 4

Improving reliability in mobility scenarios using Multi-Connectivity with Packet Duplication

In this chapter, we explore a novel technique employed with multi-connectivity architecture that supports enhanced mobility in 5G networks. We evaluate this technique and compare it with the present-day solution using simulations. We also analyze ways to improve radio resource utilization with the proposed technique. Section 4.1 takes a closer look at the Multi-Connectivity architecture discussed in the previous chapter and identifies areas for performance improvements. The implementation methodology and simulation details are presented in Section 4.2. Section 4.3 discusses the experimental evaluation strategy and also the results we obtain from our simulations.

4.1 Performance Improvement in Mobile URLLC Applications

As discussed earlier, URLLC use cases require a paradigm shift from current cellular network standards to contend with the stringent requirements with complex trade-offs. Incrementally improving the existing or traditional approaches to satisfy the URLLC requirements could prove exceptionally challenging and expensive resource usage-wise. For URLLC use cases, more emphasis is given to the delivery of critical packets with high resiliency to variations in channel and load conditions rather than high throughput. These requirements may sometime vary depending on the specific use case. In general, 3GPP requires URLLC transmissions in the 5G NR to be within a maximum transmission time of 1 ms on the user plane (UP) and achieve target reliability of 99.999% for a packet size of 32 bytes [7]. But, in another related use case, the packet size can be up to 300 bytes, while the latency and reliability targets can range between 1 ms to 15 ms and 99.999% to 99.9999999%, respectively. Thus, we see similar areas could have very different sets of requirements and the network has to be dynamic enough to handle all these use cases.

The dynamic nature of wireless links has made techniques like error detection, error correction, retransmissions, etc an integral part of wireless networks. In previous generations, packet reliability requirements are traditionally satisfied using automatic repeat request (ARQ) and hybrid ARQ (HARQ) retransmission techniques at the radio link control (RLC) and medium access control (MAC) sub-layers [10]. Although these techniques no doubt have improved the reliability of transmission to some extent but the retransmissions involved push the latency results to well exceed the 1 ms limit requirement from the specifications. Hence, it is necessary to consider a fundamentally new technique that would be effective, practical, and yet fit easily with the current RAN architecture.

To address these challenges, the research community is considering adopting packet duplication (PD) of both data plane and control plane packets as a fundamental technique in 5G NR. In this dissertation chapter, we discuss a simulated experimental framework that implements packet duplication to analyze its impact on the 5G RAN architecture for supporting URLLC. Packet duplication can theoretically satisfy the latency as well as reliability requirements without adding to the complexity of the system. As discussed earlier, meeting URLLC requirements for the network is an even bigger task when we add UE mobility to the equation, but packet duplication can improve transmission robustness during mobility and against radio link failures. As a result, we analyze the packet duplication technique under various mobility scenarios and provide its performance results. However, it can also be observed that the improvement in reliability comes at the cost of reduced network spectral efficiency. Since packet duplication using dual-connectivity mode utilizes the resources of two gNBs to serve one UE, the total radio resources needed for communication are also doubled. Thus to improve the radio resource utilization when packet duplication is used, we also propose enhancements to the packet duplication feature based on the received signal strength indicator (RSSI) level reported by the UE for the attached gNBs and based on the UE position and mobility.

4.2 Packet Duplication Strategy and Simulation Methodology

The multi-connectivity technique has been extended as a reliability enhancement solution using packet duplication by 3GPP in 5G NR standards [11]. In this study, packet duplication is performed at the PDCP layer and undergoes independent processing at each base station. This requires lesser coordination between TRPs, leading to lower network deployment and computational costs. The duplicated packets may be transmitted at different time instants, using different radio resources, and with different modulation and coding schemes. We also note that the lower layer procedures like HARQ retransmissions, fragmentation, reassembly, etc. all still take place at each gNB and UE stack as the two transmissions are independent of each other. Packet duplication provides high reliability with no additional latency when compared to other techniques like HARQ, CoMP transmissions, multi-TRP schemes, etc.

The improvement in reliability of the system using packet duplication, however, comes at the cost of higher radio resource usage from the transmission of redundant packets. We utilize double the radio resources when sending two duplicated instances of the same data message on different links as compared to just sending one instance using a single link. This raises the question of how much of the radio spectrum and resources must the network dedicate to achieve what level of reliability. And, how much loss can the network allow to save on the radio resource usage? Thus, finding the correct balance between reliability and radio resource utilization in these cellular networks remains an open research question. Such balance can only be attained by dynamically controlling the decision to perform packet duplication. We must decide based on channel conditions and other UE parameters if certain scenarios are unfavorable for the network and start using packet duplication only in such cases. Thus, packet duplication becomes critical to enhance the system for achieving high reliability along with good radio resource utilization. In this work, we propose an enhancement to the packet duplication feature for efficient radio resource utilization by looking into different UE parameters. Parameters like the distance of the UE from gNB, the velocity of a mobile UE, or the UE-reported RSSI levels for the gNBs are some of the parameters that could help us take such decisions.

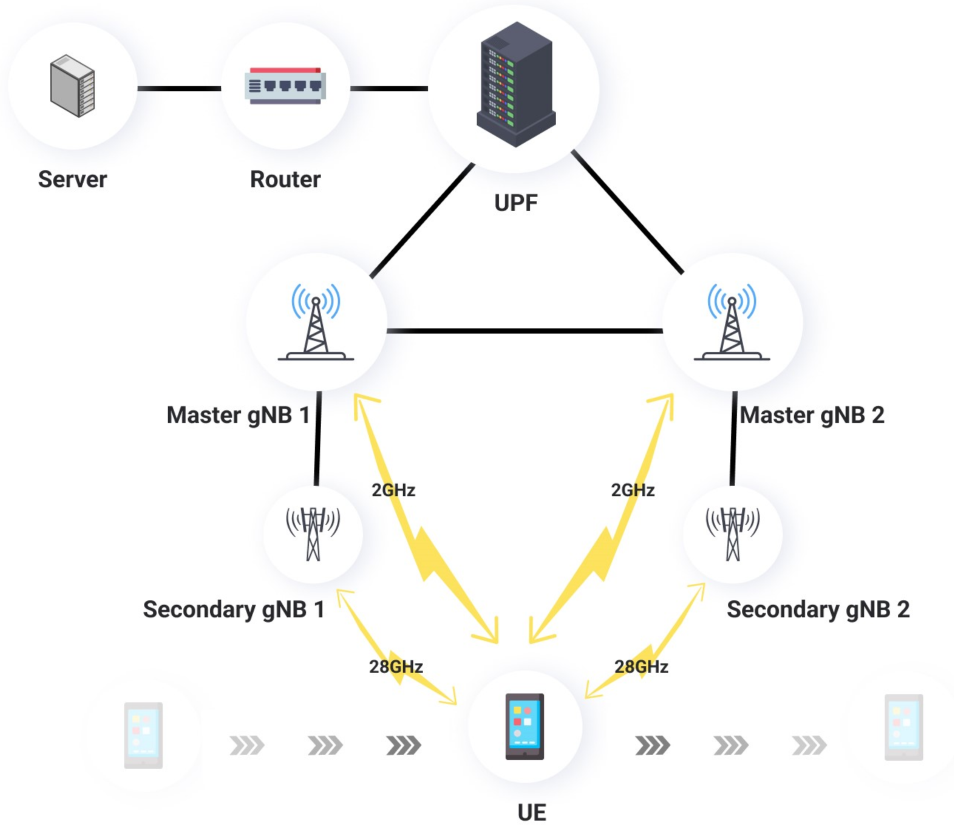


Figure 4.1: System Model for 5G NR-DC Deployment with Packet Duplication

The system model of an NR-DC deployment is shown in Figure 4.1. For this study, we use the OMNeT++ framework and the Simu5G simulation model library to implement the proposed NR-DC architecture that supports packet duplication and its optimization. We deploy a dual connectivity architecture (NR-DC) with the master gNB as a 5G NR macro cell operating in the sub-6 GHz frequency and the secondary gNB as a 5G NR small cell operating in mmWave frequency. The UE is capable of simultaneously connecting to both the master and secondary gNB, which are interconnected via an Xn interface. The master gNBs are connected to the core network and facilitate the secondary gNB selection and DC configuration based on the reported UE channel measurements.

As part of the 5G core implementation, Simu5G provides a UPF module responsible for forwarding IP packets between the data network and the gNBs through the GTP (GPRS Tunnelling Protocol) tunnels. The UPF module can be directly connected to the gNBs, and each gNB is connected with its adjacent gNB through an X2 link. In the case of mobile UEs, when handovers

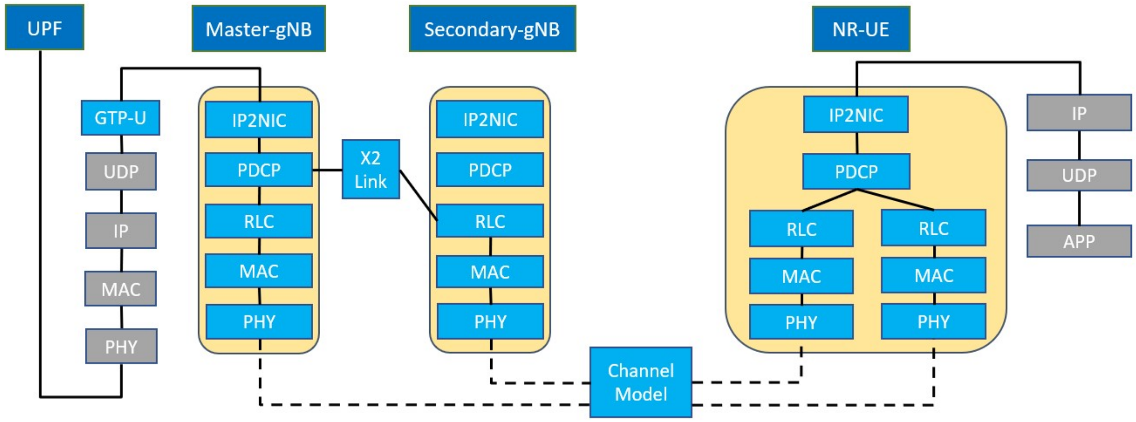


Figure 4.2: Simu5G Architecture for NR-DC Deployment with Packet Duplication

must be executed between gNBs, the source gNB buffers the received packets and sends them to the destination gNB through these X2 links. The native version of Simu5G was extended for conducting this study to include gNB to gNB handovers in single connectivity and DC architectures. As experiments in this study consider mmWave communications, the latest channel models specified in TR 38.901 [5] are also implemented.

Figure 4.2. shows the submodule structure of NR-DC deployment in Simu5G. The 5G NR in Simu5G is simulated in two main compound modules called *NrUe* and *gNodeB*. The functionalities of the RAN are implemented as a stack of the Packet Data Convergence Protocol (PDCP) layer, Radio Link Control (RLC) layer, Media Access Control (MAC) layer, and Physical (PHY) layer. The *gNodeB* module also contains an *X2Manager* module responsible for forwarding IP packets on the X2 link toward the secondary node when multi-connectivity is implemented. Apart from PDU duplication, the PDCP layer on the master gNB is also responsible for deciding if a packet will be sent to the lower RLC layer on the local stack or forwarded through the X2 link. The forwarded packet is then received by the RLC layer of the secondary gNB, after which the packet continues to flow through the lower layers toward the UE.

The *NrUe* module implements the full protocol stack from the physical layer up to the application layer. Along with the Simu5G RAN layers, it uses the INET library to implement the UDP and IP layers in the stack. The dual connectivity feature in the *NrUe* is implemented using a dual stack with two PHY, two MAC, and two RLC layer submodules. *NrUe* has a single *NrPdcp* module responsible for reordering and removing any incoming duplicate PDUs from the dual-stack

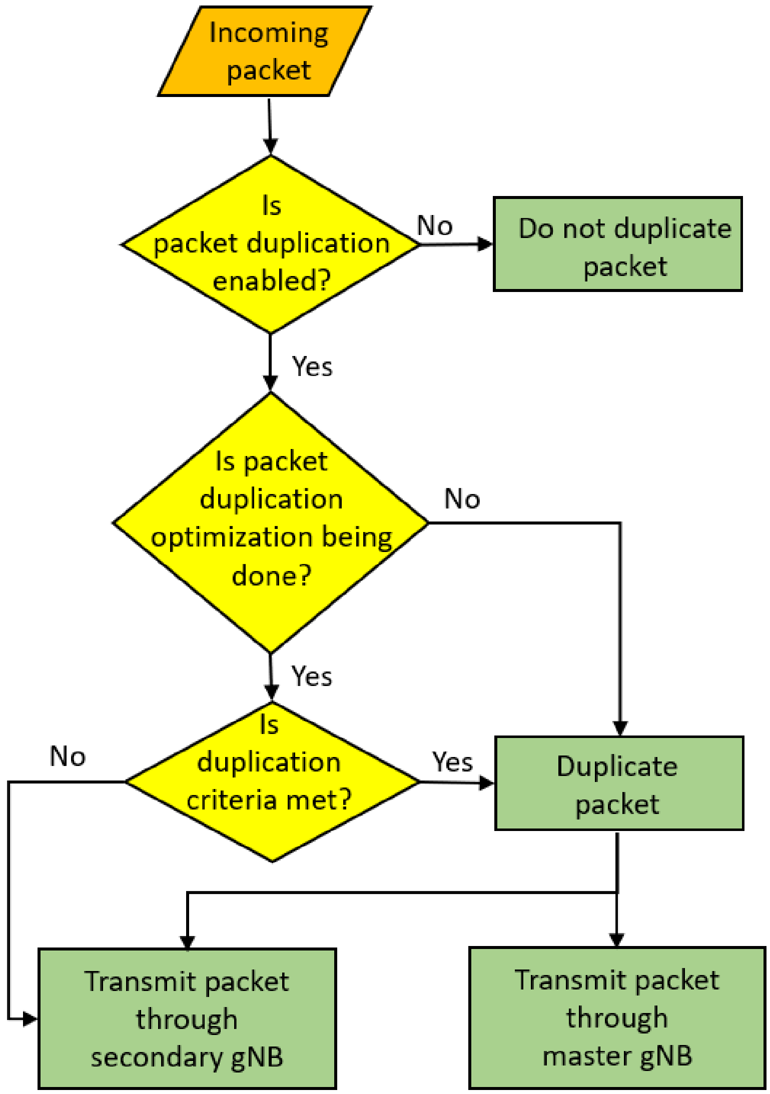


Figure 4.3: Packet duplication process at the sender

and sending them to the higher layers. The *carrierAggregation* module in Simu5G can be used for simulating communication on multiple carrier component frequencies and varying the numerology.

Figures 4.3 and 4.4 show the packet duplication process flow at the sender and receiver sides, respectively, along with the optimization process. The sender node implements the optimization decision block that improves the radio resource utilization. The PDCP layer at the sender node is modified to duplicate packets when needed, and in the receiver, it is added with the intelligence to detect duplicate packets and discard them. When an application generates a packet destined for a

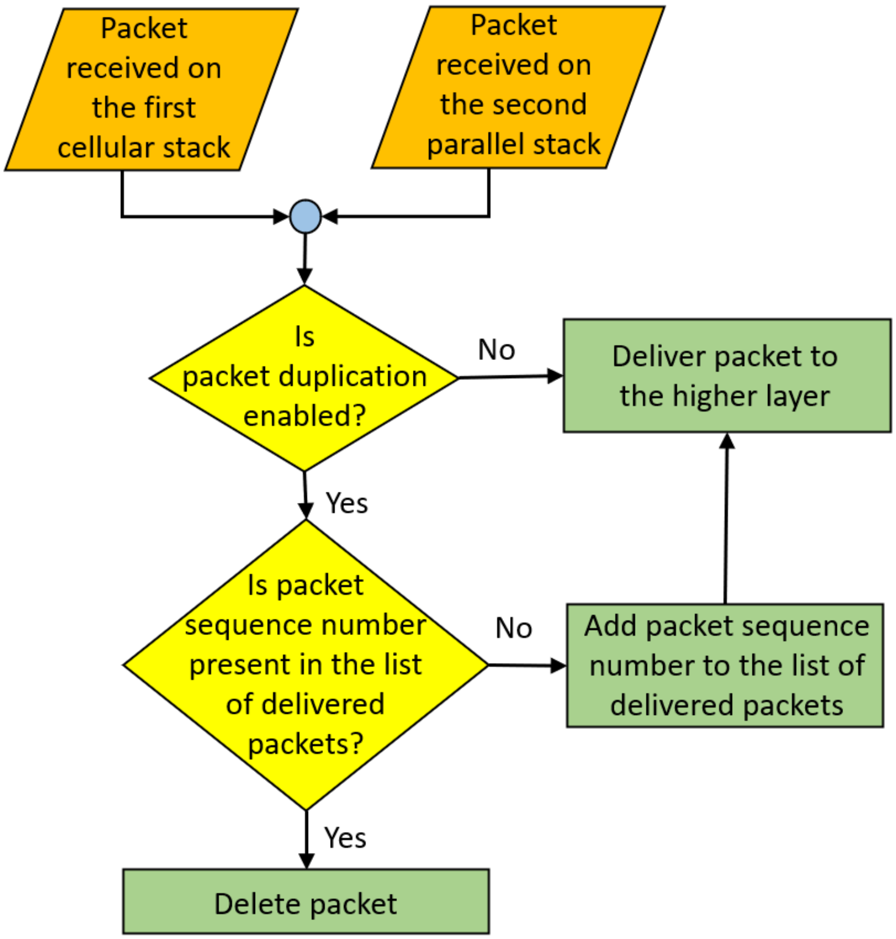


Figure 4.4: Packet duplication process at the receiver

UE connected to our network, it is first routed to the UPF module. From there, the UPF sends the packet to one of the master gNB in the network based on which gNB is serving the destination UE currently. When a packet from the 5G core network arrives at the master gNB, the very first decision the gNB takes is to check if packet duplication is enabled. This decision taken by the network of whether to perform packet duplication can potentially be based on many factors, like the type of data traffic, network capacity, UE location, and UE eligibility based on policy, among others. Applications that generate URLLC-type packets are prime candidates for packet duplication as they are both time sensitive and need guaranteed delivery. If, however, there is no duplication enabled, then the gNB will handle the packet like any other packet and passes it down its own protocol stack for transmission.

If packet duplication is enabled in the network, the gNB next checks whether packet duplication optimization is being done. If not, it blindly duplicates every packet it receives and passes a copy for transmission to the secondary gNB. However, if optimization is being performed, then gNB looks at the duplication parameter being used and compares it with the threshold to decide if should perform duplication or not. In this work, we consider three types of packet duplication optimization. The distance of the UE from the gNB is the first parameter that is considered for optimization, and packet duplication is done if the distance exceeds the distance threshold. For the second case, the velocity of the UE is used for optimization, and only if the velocity exceeds the velocity threshold packet duplication is done. The UE periodically reports to the gNB the RSSI that it has measured from the gNB to take decisions like handover etc. In the third case, this received RSSI reported by the UE is used by the gNB for taking the optimization decision, and packet duplication is done if the RSSI is below the RSSI threshold. The master gNB takes the decisions based on a pre-set threshold. If the duplication criteria is not met, then it is assumed that the link between the secondary gNB and the UE is good enough for data transmission. If no duplication is done, then our design uses only the high throughput, low latency mmWave link at the secondary node (SN), assuming the link is strong and can sustain the transmissions.

But if the duplication criteria are met, the master gNB starts duplicating the received PDUs and sends the PDUs using both master gNB and secondary gNB. When we use packet duplication for increasing reliability, the PDCP layer of the master gNB duplicates the incoming PDUs destined for the user and sends them to the secondary gNB via the X2/Xn interface, along with sending one of the incoming PDUs to its own RLC layer. The duplicated PDUs go through independent scheduling and transmissions at each node.

At the receiver side, a UE that has the dual-connectivity capability must be informed whether packet duplication is being used by the network during the PDU session establishment. When active, the PDCP layer in the UE is responsible for in-order delivery and forwards the first successfully received PDU to the higher layers while discarding any duplicates that arrive later. The duplication process is implemented using PDCP sequence numbers where the duplicate of a PDU carries the same sequence number as the original. The receiving PDCP entity uses these sequence numbers to detect and remove any duplicates.

4.3 Evaluation

A systematic evaluation was conducted to study the performance of the packet duplication technique in multi-connectivity NR-DC architecture. In our experiment, we consider downlink transmission, where a server is connected to the UPF in SA deployment via a router and sends data packets to a mobile UE in our system. The server and the UE are running an application that sends fixed-size packets at regular packet inter-arrival times. In the NR-DC deployment, the master gNBs operate at a carrier frequency of 2 GHz to improve the system's reliability, especially for mobility scenarios. Additionally, secondary gNBs operating at a mmWave carrier frequency of 28 GHz are used to provide a high bandwidth link. The NR-DC deployment used is specifically designed to initiate handovers during the simulation and also simulates areas without mmWave coverage to test reliability performance. The radio access network design and other network parameters are kept constant across experiments.

4.3.1 Performance Metrics

To ensure that we truly observe the benefits of implementing packet duplication in the network, we need to track the latency and reliability metrics. As a result, we measure the end-to-end application-level latency and the packet loss ratio observed at the UE to measure the performance of the packet duplication technique. We also measure the end-to-end throughput of the system to ensure that there is no drop in it because of packet duplication. To evaluate the packet duplication optimization feature and measure the improvement in radio resource utilization, we use the packet duplication ratio (PDR) as a metric. PDR is measured as the percentage of the number of extra packets that the master and secondary gNBs transmit to the number of packets that the server generated for transmission. This means that when no optimization is used, i.e., all packets are duplicated, then the PDR is 100%. As the duplication threshold is changed, the PDR will change. For example, if we decrease the RSSI threshold, it means we now take a less conservative approach by not duplicating some packets at higher RSSI levels, and hence the PDR decreases. Similarly, if the distance threshold is increased, the PDR will decrease as we do not duplicate when UE is closer to the gNB, assuming the signal will be stronger there. Thus, by looking at the PDR metric, we can directly observe the correlation between packet duplication and the performance of the network.

4.3.2 Experimental Setup

In this chapter, we perform experiments to study the effect of packet duplication in an NR-DC type deployment on network performance. We again want to simulate a city-like environment, we consider a mobile UE with a mobility pattern that would be similar to a user moving around blocks in the downtown area of a city. For the NR-DC deployment to simulate the path loss and interference patterns that would be observed in a city downtown with tall buildings, we use the Urban Microcell path loss model from the 3GPP specifications for secondary gNBs operating in the mmWave frequency range and the Urban Macrocell path loss model for master gNBs operating in the sub-6 GHz range. We accordingly calculate and use line-of-sight probabilities in the simulation for better accuracy. We primarily run two experiment sets where the first one looks at the comparison between the effect of different types of packet duplication optimization on the network's performance as observed by a single UE moving in a pattern that simulates city grid-like patterns. The second set of experiments studies the impact of increasing load on the same network by testing the network performance with multiple randomly moving UEs in a specific area.

In the NR-DC deployment, we deploy two master-secondary pairs across the simulation area. In NR-DC deployment, the master gNB operates at 2 GHz in the sub-6 frequency range and at a 49 dBm transmit power level, whereas the secondary gNB operates at 28 GHz in the mmWave range and transmits at a 35 dBm power. For an accurate comparison across experiment scenarios, we use the same number of gNBs with the same placements in the NR-DC deployment. So every experiment has the Master Cell Group (MCG) consisting of three gNBs and the Secondary Cell Group (SCG) with three gNBs. The server VoIP application generates packets of size 1024 bytes. We run experiments for packet inter-arrival times of 10 ms. The details of the simulation parameters that were used in the experiment are provided in Table 4.1. The packet size and packet inter-arrival times are chosen to evaluate the network performance for an application use case that needs low latency and reliable communication. The 10 ms packet inter-arrival time combined with the radio resources allows us to measure the range of performance the network can deliver, especially in the multiple UE scenario. By using these parameters with a small number of UEs, the network is able to show a good level of performance, but as the number of UEs is increased, we can also observe the saturated network behavior in a situation where resources are overloaded.

As mentioned earlier, we broadly run two sets of experiments. Both sets of experimental

Table 4.1: Simulation parameters for UDN and NR-DC deployments used during optimization

Parameter	**Value**
mmWave Carrier Frequency	28 GHz
mmWave Bandwidth	100 MHz
Sub-6 GHz Carrier Frequency	2 GHz
Sub-6 GHz Bandwidth	20 MHz
UDP Packet Size	1024 B
UDP Packet Inter-arrival Time	10 ms

results are compared by testing different packet optimization techniques in an NR-DC deployment for an environment that simulates the downtown of a city. It assumes tall buildings with streets running between them. The first set of experiments is conducted using a single UE that is moving along the city grid. The mobility pattern used for this experiment is *TurtleMobility*, where we start at a fixed point and move up to a given point at a random speed chosen uniformly between 10 m/s to 40 m/s. On reaching this intermediate point, the UE takes a 90-degree clockwise turn and moves to the next end-point, again with a uniformly selected random speed. This pattern repeats till the end of the simulation that is run for a fixed amount of time. We repeat each set of experiments ten times and average the results. For the second set of experiments, we test each strategy for the network load they can handle. We basically evaluate the network performance by increasing the number of UEs present in the simulation. We start with 2 UEs and increase the number of UEs to 4, 8, 16, 32, and 64. In each of these experiments, the UEs are moving randomly using a mobility pattern called *RandomWaypointMobility*. The path loss and interference models are still assumed to be that of a city-like environment. In this type of mobility, the simulation randomly selects the start point and an endpoint for the UE. The UE then travels to this endpoint at a speed that was randomly selected. On reaching the end-point, the model picks the next endpoint and another random speed to move the UE there. In our experiments, we pick the speed of the UE randomly using a uniformly distributed random variable between 1 m/s and 40 m/s.

Each experiment set tests the two network deployments in two different area scenarios. The first set of experiments is conducted with a single UE and the second set of experiments with

Table 4.2: Master gNB (M-gNB) and Secondary gNB (S-gNB) Coordinates for NR-DC deployment in 1000 m and 1500 m scenarios for 4 gNBs

NR-DC Packet Duplication Deployment		
Scenario	**1000 m x 1000 m**	**1500 m x 1500 m**
M-gNB1	(100 m, 500 m)	(100 m, 750 m)
S-gNB1	(100 m, 501 m)	(100 m, 751 m)
M-gNB2	(900 m, 500 m)	(1400 m, 750 m)
S-gNB2	(900 m, 501 m)	(1400 m, 751 m)

Multiple UEs. The first scenario considered has a 1000 m by 1000 m simulation area, and the second scenario considers a 1500 m by 1500 m simulation area. In each scenario, we consider four gNBs i.e. two master and two secondary gNBs as shown in Table 4.2 that gives the coordinates for the gNB deployment in the NR-DC deployments scenarios. Both sets are run for each scenario and for the three optimization methods. Apart from testing the case when full duplication is done, we run experiments with the following thresholds. In the single UE set for RSSI-based optimization, the RSSI thresholds were taken as 10, 20, 30, 40, and 50. Similarly, the velocity thresholds were taken as 15, 25, 35, and 45 m/s. The distance thresholds for the single UE case with *TurtleMobility* pattern were chosen based on the resulting UE-gNB distances for that simulation area. Thus for the 1000 m scenario, where the minimum UE-gNB distance was 100 m and the maximum UE-gNB distance came out to be 500 m, the distance thresholds of 150, 200, 250, 300, 350, 400, 450, and 500 m were chosen. Whereas for the 1500 m scenario, 300, 350, 400, 450, 500, 550, 650, 700, and 750 m were taken as the distance thresholds as the minimum UE-gNB distance, in this case, was 250 m and the maximum UE-gNB distance was 763 m. On the other hand for the multiple UE experiments with *RandomWayPoint* mobility to reduce the complexity and run time we chose the threshold a little more conservatively. For RSSI-based optimization, the RSSI thresholds were taken as 10, 20, 30, 40, and 50. Similarly, the velocity thresholds were taken as 5, 15, 25, and 35 m/s. The distance thresholds were chosen based on the area in use, thus for the 1000 m scenario 150, 250, and 350 m were taken as the thresholds. Whereas for the 1500 m scenario 150, 250, 350, 450, 550, and 650 m were taken as the thresholds.

Table 4.3: Performance comparison for distance-based optimization of packet duplication in a 1000 m scenario with 2 pairs of Master and Secondary gNBs in an NR-DC deployment for a single UE

Distance-Based Optimization				
Distance Threshold	**PDR (%)**	**Latency (ms)**	**Packet Loss (%)**	**Throughput (Mbps)**
Full Duplication	100.00	4.71	0.00	0.82
150	81.45	5.07	0.00	0.82
200	73.02	5.33	0.01	0.82
250	69.01	5.43	0.01	0.82
300	59.30	5.81	0.04	0.82
350	35.97	6.53	0.26	0.82
400	30.52	6.70	0.36	0.82
450	10.58	7.65	1.32	0.81
500	0.00	8.64	3.81	0.79

4.3.3 Results for single UE

The first set of experimental results is presented for the single UE case. The goal of these experiments is to understand the effects of packet duplication and the optimization techniques on the system without pushing it to the limits. We try to analyze the similarities and differences in deploying different types of optimizations. With a single UE system, we explore a wider range of thresholds for each optimization technique to better understand the variations in performance caused by different duplication percentages.

4.3.3.1 Distance-based Optimization

The results for distance-based packet optimization are shown in Table 4.3 and 4.4 for the 1000 m-by-1000 m simulation scenario and 1500 m-by-1500 m scenario, respectively. We present the values for packet duplication ratio (PDR), latency, packet loss, and throughput for different distance thresholds. For the 1000 m scenario as discussed earlier 500 m is the maximum UE-gNB

Table 4.4: Performance comparison for distance-based optimization of packet duplication in a 1500 m scenario with 2 pairs of Master and Secondary gNBs in an NR-DC deployment for a single UE

Distance-Based Optimization				
Distance Threshold	**PDR (%)**	**Latency (ms)**	**Packet Loss (%)**	**Throughput (Mbps)**
Full Duplication	100.00	4.95	0.00	0.82
300	79.83	5.85	0.06	0.82
350	77.26	6.03	0.15	0.82
400	63.65	6.86	0.42	0.82
450	59.09	7.16	0.73	0.82
500	46.40	8.01	1.92	0.81
550	39.37	8.53	2.62	0.80
600	29.83	9.14	2.80	0.80
650	24.69	9.62	2.89	0.80
700	15.30	10.55	3.61	0.79
750	3.34	12.85	5.21	0.78

distance possible in the simulation. As a result, we see that at the distance threshold of 500 m, the PDR value is 0. On the other hand, the minimum UE-gNB distance possible is 100 m in the simulation and hence at the distance threshold of 150 m we see 81% duplication has happened. We also present the results for the case when full duplication is being performed leading to a PDR of 100%. Similarly, for the 1500 m scenario as discussed we started at the distance threshold of 300 m as the threshold of 250 m would have resulted in full duplication or PDR of 100%. And, as the maximum distance possible is 763 m we observe at a distance threshold of 750 m only 3.34% packets are being duplicated.

Using distance-based optimization we can clearly improve in terms of the number of packets being duplicated in the system. As distance thresholds are increased we see PDR decreases whereas the packet loss and latency increase with it. As the location of the UE is a very important charac-

Table 4.5: Performance comparison for velocity-based optimization of packet duplication in a 1000 m scenario with 2 pairs of Master and Secondary gNBs in an NR-DC deployment for a single UE

Velocity-Based Optimization				
Velocity Threshold	**PDR (%)**	**Latency (ms)**	**Packet Loss (%)**	**Throughput (Mbps)**
Full Duplication	100.00	4.71	0.00	0.82
15	77.63	5.01	0.13	0.82
20	34.70	6.37	2.14	0.80
25	16.30	7.03	2.79	0.80
30	7.55	7.74	3.11	0.80
35	3.16	8.18	3.41	0.79
40	0.18	8.78	3.87	0.79

teristic that is constantly tracked by the network for many other purposes like handovers, tracking area updates, etc., we can easily use this information to calculate the distance between the UE and the gNB. Thus, this UE-gNB distance can be a very simple real-time optimization criterion that could lead to significant performance improvements.

4.3.3.2 Velocity-based Optimization

We present the results of velocity-based optimization in Tables 4.5 and 4.6 for the 1000 m and 1500 m scenarios, respectively. We observe result trends that are similar to what we saw in distance-based optimization. With an increase in PDR as the velocity thresholds are decreased we observe that the packet loss and latency decrease as well. In the 1500 m scenario with a threshold of 15 m/s we are duplicating 91.66% of the packets and we see it results in only 0.28% packet loss. However, we do see a slight increase in the average latency of the packets in the order of nearly 0.41 ms. Similarly, at a velocity threshold of 40 m/s, we only duplicate around 2.87% of the packets and that results in the worst latency and packet loss among all thresholds. These results also are very close and similar to the NR-DC Split-Bearer case. We also observe that the 1000 m scenario gives

Table 4.6: Performance comparison for velocity-based optimization of packet duplication in a 1500 m scenario with 2 pairs of Master and Secondary gNBs in an NR-DC deployment for a single UE

Velocity-Based Optimization				
Velocity Threshold	**PDR (%)**	**Latency (ms)**	**Packet Loss (%)**	**Throughput (Mbps)**
Full Duplication	100.00	4.95	0.00	0.82
15	91.66	5.36	0.28	0.82
20	60.75	6.53	0.75	0.82
25	32.60	8.13	2.44	0.80
30	19.15	10.15	3.68	0.79
35	11.44	11.33	5.00	0.78
40	2.87	12.94	5.83	0.77

lower losses and smaller latency for similar PDR when compared to the 1500 m scenario.

Although velocity is not a UE characteristic that is fundamentally measured by the network, it is a parameter that can be very easily derived from remembering the UE's past locations and times when it was present there. Both this information could be readily made available to the network to calculate the velocity of the UE thus, make an optimization decision in almost real-time. Although the signal path loss is more closely related to the UE-gNB distance and the velocity only has an impact on path loss in terms of causing a Doppler shift, the velocity-based optimization performs significantly well. One of the reasons for such performance could be that the UEs that are moving at a high velocity are more likely to move out of coverage very soon and the link could degrade quite rapidly. Hence, it is a good policy to duplicate these packets for performance improvement.

4.3.3.3 RSSI-based Optimization

Tables 4.7 and 4.8 show the results of 1000 m and 1500 m simulation scenarios with RSSI-based packet duplication. We again observe that PDR decreases with the RSSI threshold and thus latency and packet loss both increase. We see that in the larger simulation area scenario

Table 4.7: Performance comparison for RSSI-based optimization of packet duplication in a 1000 m scenario with 2 pairs of Master and Secondary gNBs in an NR-DC deployment for a single UE

RSSI-Based Optimization				
RSSI Threshold	**PDR (%)**	**Latency (ms)**	**Packet Loss (%)**	**Throughput (Mbps)**
Full Duplication	100.00	4.71	0.00	0.82
50	61.17	5.24	0.18	0.82
40	20.42	6.55	2.10	0.80
30	2.03	7.69	2.63	0.80
20	0.40	8.16	3.28	0.80
10	0.00	8.65	3.96	0.79

Table 4.8: Performance comparison for RSSI-based optimization of packet duplication in a 1500 m scenario with 2 pairs of Master and Secondary gNBs in an NR-DC deployment for a single UE

RSSI-Based Optimization				
RSSI Threshold	**PDR (%)**	**Latency (ms)**	**Packet Loss (%)**	**Throughput (Mbps)**
Full Duplication	100.00	4.95	0.00	0.82
50	84.39	5.40	0.52	0.82
40	47.61	6.70	2.66	0.80
30	9.62	9.35	3.71	0.79
20	1.60	10.77	4.98	0.78
10	0.00	13.39	5.71	0.77

similar thresholds result in a larger duplication ratio, as expected and thus assist in performance improvement in such cases. Again at maximum thresholds results are very similar to the NR-DC split-bearer case and full duplication again results in 0 loss and low latency. We also observe that the 1500 m scenario shows higher losses and larger latency for similar PDR when compared to the 1000 m scenario.

RSSI, as a UE metric, is the most accurate representation of the channel quality and hence closely determines if the packet will be transmitted successfully. But it is not readily available as the other metrics like distance or readily calculated by the network like velocity. RSSI would be measured by the UE and reported to the network in frequent intervals leading to additional overhead and delay. However, by using RSSI-based optimization we can improve the radio resource efficiency with minimal impact on performance. RSSI threshold of 50 dBm duplicates 40% fewer packets at the cost of adding only 0.53 ms latency and 0.18% packet loss.

4.3.3.4 Packet Duplication improves the performance of the network

We observe from the results shown for all the optimization techniques that the packet loss and latency results are lower when some duplication is being done compared to the results we saw in the case of NR-DC with split-bearer. This is because we are selectively duplicating the packet based on UE metrics like RSSI, distance, etc., rather than blindly splitting the incoming packets. By sending the packets that are more likely to be transmitted correctly through the mmWave links we decrease the average latency of the packets. Further, by duplicating the packets when there is a chance that their transmission on the mmWave link might fail we also improve upon the packet loss as the sub-6 GHz link would be more reliable.

4.3.3.5 No duplication performance is comparable to NR-DC split-bearer

At zero PDR or when no duplication is being done we observe the latencies are similar to what was seen in the case of NR-DC with Split bearer. With more duplication of packets, we observe improvement in the latency as well as the packet loss and reach zero packet loss when full duplication is done. The reason we see zero PDR performance comparable to split bearer is that, in our implementation at zero PDR, all incoming packets would be sent to the secondary mmWave link with more available resources. However, when the link is poor and there is no connection between the UE and the mmWave gNB, all packets are transmitted via the sub-6 master gNB. This allows for an effective way for the network to fall back on the more resilient backup link. We hence still notice losses and latency comparable to NR-DC split bearer but slightly better than when we blindly performed a 50-50 split of packets. At lower loads, this would not impact the latency but increases the amount of data the master gNB is responsible for scheduling. Thus, as long as the load is manageable we see lower latency from the sub-6 GHz links and the packet loss percentages are also low.

4.3.3.6 Similar PDR values for distance-based optimization results in lower packet loss but higher latency

We plot the results of latency versus the PDR in Figure 4.5 for both 1000 m and 1500 m scenarios for all three optimization techniques. Although we observe a similar trend in PDR versus the latency and packet loss results for both distance and RSSI thresholds, we also see that when using distance-based optimization, for similar PDR values we observe lower packet loss but higher latency when compared to RSSI-based optimization. Fundamentally, in an identical scenario, the packets that are being duplicated for RSSI-based optimization compared to distance-based optimization would not be the same. As discussed RSSI as an optimization metric incorporates many factors in the environment, of which UE-gNB distance is a very important one among them. When we use simple distance-based optimization we are ignoring many other factors that could affect the packet transmission like speed, surrounding buildings, environment, etc.

The trend of observing lower packet loss but higher latency in distance-based optimization results is most probably the result of using a delayed RSSI measurement. In the simulation, we measure RSSI every second and report it to the network for duplication decision-making. Any packet that comes in the next second will be duplicated based on this measurement which could have potentially changed. However, when we use distance as a criterion we get a more recent real-time location of the UE to take the duplication decision. Thus, with the distance-based method, we could potentially be duplicating packets that we should not as the link may be strong. On the other hand, with the RSSI-based method, we could have missed duplicating some packets that we should have. Thus leading to slightly higher packet loss. But the packets we send are mostly transmitted with low latency mmWave link that is sure to be in good condition thus decreasing the latency we see. This again is thus a trade-off we observe between doing a more accurate RSSI-based duplication versus a more real-time distance-based option that is less resource-consuming and faster. We must

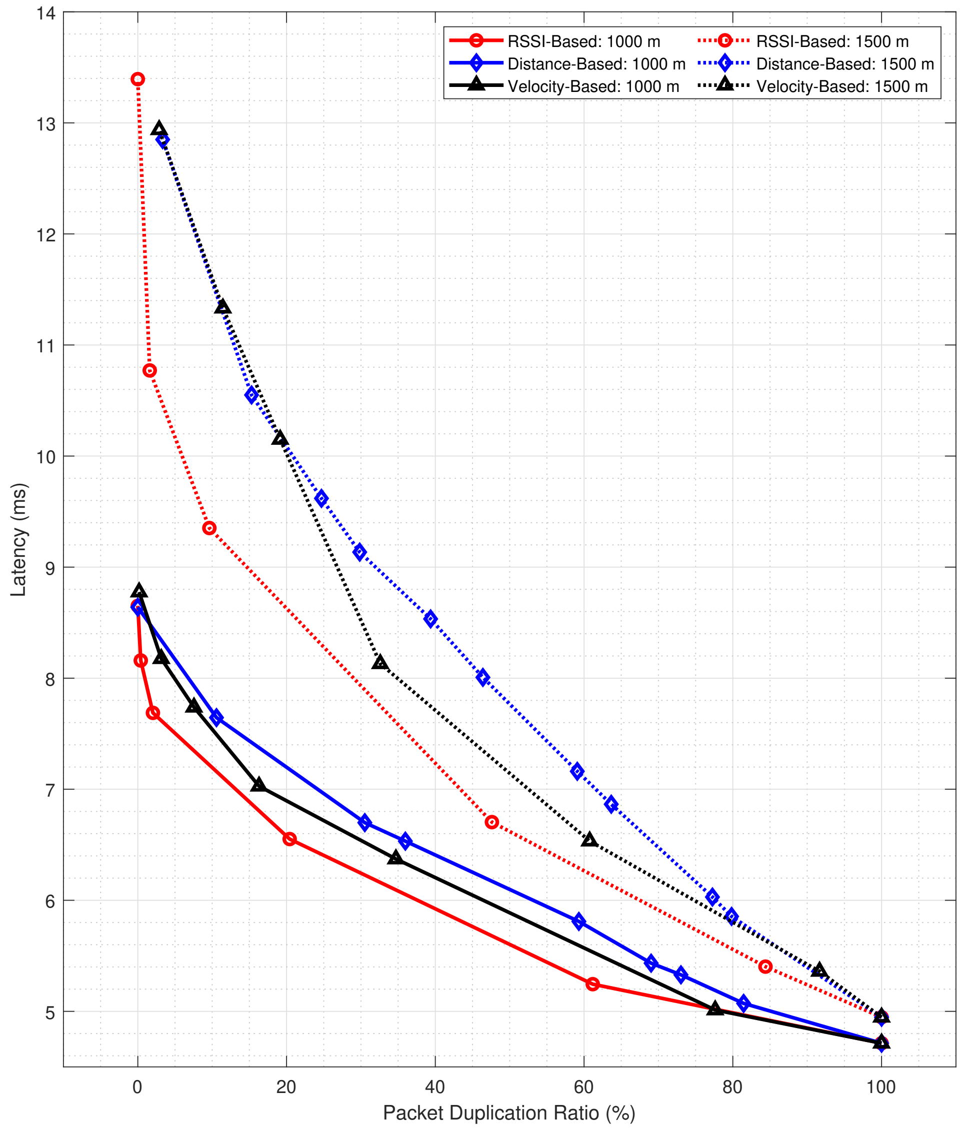


Figure 4.5: Latency vs. Packet Duplication ratio plots with single UE in the system for 1000 m and 1500 m scenarios

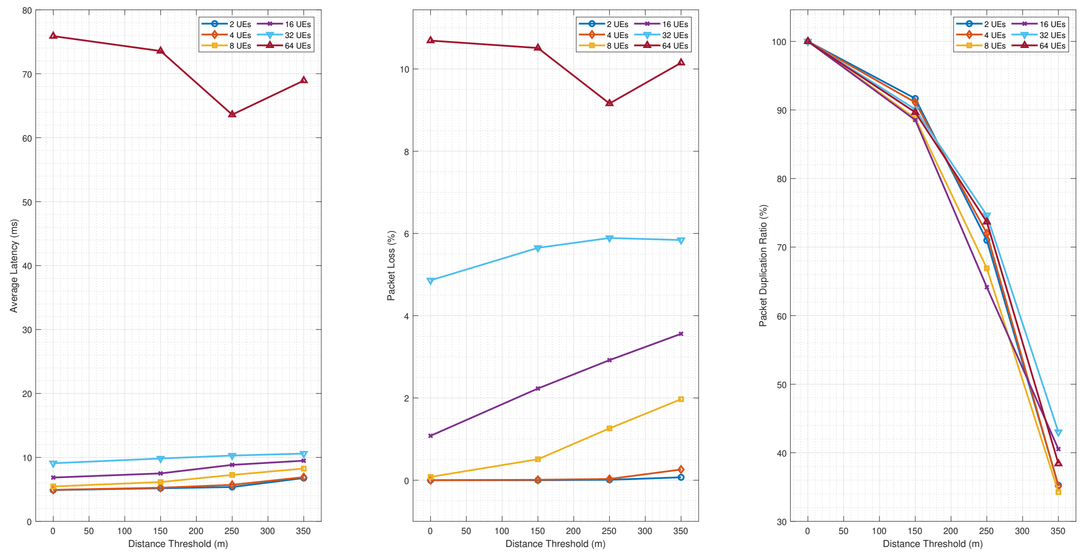
also note that reporting RSSI at smaller intervals would also increase the resource consumption at the UE for RSSI calculation and at the network level as well for communicating this information.

4.3.4 Results for multiple UEs

The next set of experiments was run to test the Packet Duplication system combined with the optimization techniques under larger load conditions with multiple UEs. These experiments allow us to observe the variation in the system performance as the load in the system is increased with multiple UE trying to communicate in the network. As performing packet duplication increases the amount of traffic the network would be handling, it is critical to test the limits of our system and study its behavior under high loads. We present the results of our experiments from using different forms of packet duplication optimizations with Multiple UEs and our observations in the next few paragraphs. From our observations of all the results, we notice that packet duplication and all the optimization techniques work as expected and the PDCP layers at the receiver and sender are able to handle and create the duplicates, respectively. It basically proves that the solution is scalable for up to multiple UEs. The performance and the network limits will depend on the resource allocation and strategy used.

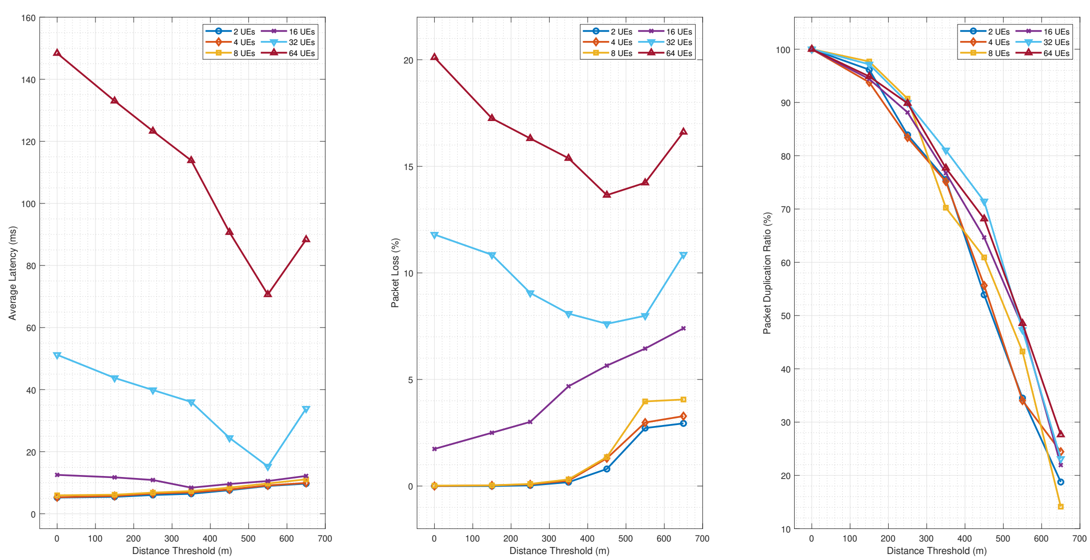
4.3.4.1 Distance-based Optimization

The results for latency, packet loss, and packet duplication ratio for the multiple UE experiments with Distance-based optimization are shown for the 1000 m and 1500 m scenarios in Figures 4.6 and 4.7, respectively. The result plots use the 0 m distance threshold on the x-axis to represent full duplication or when no form of optimization is utilized. As a result, we observe PDR of 100% in such cases, and we also observe minimum latency and zero packet loss. With full duplication for 2 UEs in the system, we are able to achieve an average latency of 4.891 ms in the 1000 m scenario and a slightly higher latency of 5.228 ms in the 1500 m scenario. We observe that the average latency and packet loss in both scenarios increase steadily as more UEs are introduced to the system. However, in the case of a large number of UEs in the system, we observe that full duplication does not result in the best results, and rather when selective duplication of specific packets is done based on the distance, we see an improvement in result. For 64 UE cases, we could achieve as low as 63.607 ms average latency for the 1000 m scenario with a PDR of 73.67% and 70.649 ms for the 1500 m scenario with a PDR of 48.54%. We also observe lower packet loss in these cases.



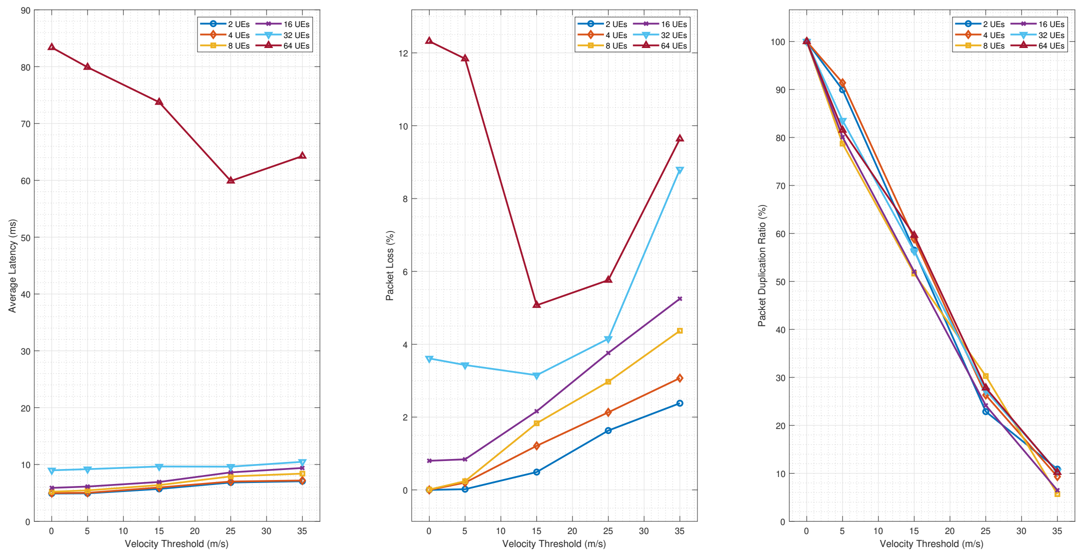
(a) Latency (b) Packet loss (c) Packet Duplication Ratio

Figure 4.6: Comparison of results for distance-based optimization in NR-DC PD with packet inter-arrival time of 10 ms for the 1000 m scenario (0 m on the x-axis represents full duplication)



(a) Latency (b) Packet loss (c) Packet Duplication Ratio

Figure 4.7: Comparison of results for distance-based optimization in NR-DC PD with packet inter-arrival time of 10 ms for the 1500 m scenario (0 m on the x-axis represents full duplication)



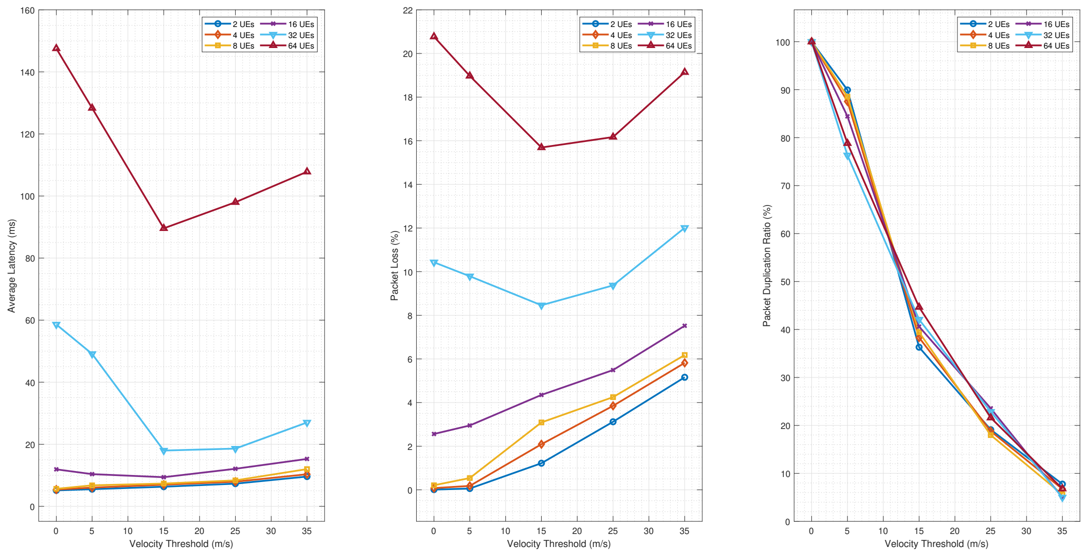
(a) Latency (b) Packet loss (c) Packet Duplication Ratio

Figure 4.8: Comparison of results for velocity-based optimization in NR-DC PD with packet inter-arrival time of 10 ms for the 1000 m scenario (0 m/s on the x-axis represents full duplication)

Thus, we observe that the performance in the case of a large number of UEs, given the limited network resources, can be improved when selective duplication is performed. The reason when there are a large number of UEs in the system, the full duplication and lower PDR cases give larger average latency is because of the extra load in the system. When we blindly duplicate, we overload the system unnecessarily, and when we do not duplicate enough, we experience issues related to mmWave propagation and limited sub-6 resources. Overall, the network resources fall short in managing the number of packets coming in. This is discussed in more detail later and better explained by looking at the comparison of minimum, average, and maximum latency trends in Figure 4.12 when the number of UEs in the system are increased.

4.3.4.2 Velocity-based Optimization

The velocity-based packet optimization results for the 1000 m and 1500 m are shown in Figures 4.8 and 4.9, respectively. We plot the average latency, packet loss, and packet duplication ratio values against the velocity threshold used in the experiments. For the purpose of representing the full duplication results on the plot 0 m/s velocity threshold was used on the x-axis. We again



(a) Latency (b) Packet loss (c) Packet Duplication Ratio

Figure 4.9: Comparison of results for velocity-based optimization in NR-DC PD with packet inter-arrival time of 10 ms for the 1500 m scenario (0 m/s on the x-axis represents full duplication)

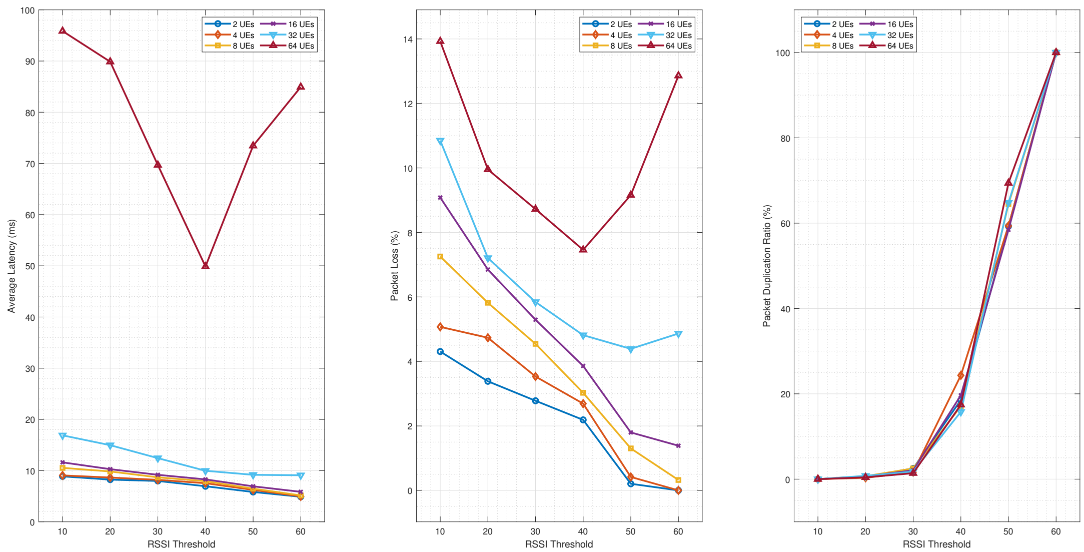
observe that by performing full duplication, we can achieve the best possible results when the number of UEs in the system is small and there are enough resources available to schedule the incoming traffic from these UEs. For example, in the result for velocity-based optimization experiments, we see that when full duplication is performed with 2 UEs in the system, we can achieve average latency of as low as 4.891 ms in the 1000 m scenario and 5.128 ms in the 1500 m scenario. We also get zero packet loss in these cases. As we start optimizing and only selectively duplicate packets based on velocity, we start seeing a rise in the value of latency as well as packet loss.

The value for these performance metrics also rises with the addition of more UEs in the system. Like in the case of distance-based optimization, we also see that in this case, as we have a large number of UEs in the system, like 32 and more UEs in the 1500 m scenario, the results for latency at full duplication start degrading very rapidly. From the detailed results, we observe that the increase in minimum latency with the increase in the number of UEs is much slower compared to the increase in the values of maximum latency. This results in an overall increase in average latency as seen in Figures 4.8a and 4.9a. This is also a reflection of the overall increase in the load of the system as more UE are added. By looking at the maximum latency trends versus the duplication ratio as the number of UEs is increased, we observe that there is a large rise in the maximum latency compared to the minimum latency. Thus, showing that the packets are backed up in the system due to large loads. This is also reflected in the packet loss numbers at a large number of UEs. We discuss these results in more detail later.

4.3.4.3 RSSI-based Optimization

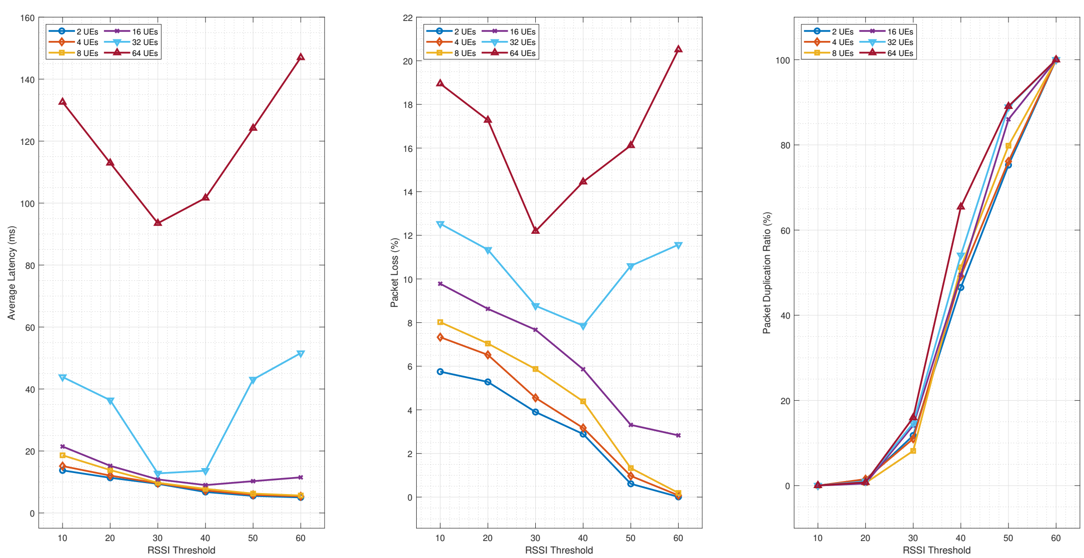
The performance results for RSSI-based optimization are presented in Figures 4.10 and 4.11 for the 1000 m and 1500 m scenarios, respectively. We again plot the average latency, packet loss, and packet duplication ratio against the RSSI threshold values used in the experiments. We notice that for RSSI-based optimization, as the thresholds are increased, we start to duplicate more and more packets. Also, for plotting the results, we use the RSSI 60 threshold to represent full duplication on the x-axis. With RSSI-based optimization and the way the experiments were designed and the threshold values assumed, we observe that for the 1000 m scenario, we do not duplicate any packets for an RSSI threshold of 10, and only around 60% of the packets are duplicated for the highest RSSI threshold of 50. Similarly, approximately 75% of the packets are duplicated for the highest threshold of 50, and almost no duplication happens for the RSSI threshold of 10 in the 1500 m scenario. The performance observed at the highest thresholds is comparable to full duplication even though we are able to save resources for almost 40% and 25% of the packets in these scenarios. This shows that RSSI-based optimization performs quite well.

For RSSI-based optimization at the lowest RSSI threshold of 10, we see a PDR of almost 0% meaning that there is very less amount of duplication happening in the system. When we compare these results with the results that we obtained from our experiments with multiple UEs with NR-DC split-bearer configuration, we observe that the performance of NR-DC PD at even 0% PDR is slightly better. For example, for the 1000 m scenario in the case of 2 UEs, we see NR-DC with PD but no duplication results in an average latency of 8.873 ms and packet loss of 4.31%. In comparison for the NR-DC split-bearer case for 2 UEs in the 1000 m scenario, we observed average latency of 9.26 ms and packet loss of 4.35%. Similarly, observations are also made for the 1500 m scenario where the 2 UE case NR-DC with packet duplication deployed but no duplication at RSSI threshold of 10 gives an average latency of 13.745 ms and packet loss of 5.75%. Whereas NR-DC split-bearer in comparison provides an average latency of 14.15 ms and packet loss of 5.88%. This improvement in performance at PDR 0 or no duplication is because of the policy of using mmWave first. In our



(a) Latency (b) Packet loss (c) Packet Duplication Ratio

Figure 4.10: Comparison of results for RSSI-based optimization in NR-DC PD with packet inter-arrival time of 10 ms for the 1000 m scenario (60 dBm on the x-axis represents full duplication)



(a) Latency (b) Packet loss (c) Packet Duplication Ratio

Figure 4.11: Comparison of results for RSSI-based optimization in NR-DC PD with packet inter-arrival time of 10 ms for the 1500 m scenario (60 dBm on the x-axis represents full duplication)

simulation, we only send the packets to sub-6 gNB if there is no mmWave gNB attached and use duplication when the link is bad instead of blindly splitting the packet between the two gNBs. As we increase the RSSI threshold we thus observe even better performance.

By looking at Figures 4.10c and 4.11c we also observe higher duplication percentages for similar RSSI thresholds in the 1500 m scenario compared to the 1000 m scenario. For example, in the case when there are 16 UEs in the system we see that for the 1000 m scenario, we duplicate around 60% and 20% of the packets for RSSI thresholds of 50 and 40, respectively. Whereas when we come to the 1500 m scenario because of the larger simulation area and a wider variation in the RSSI levels observed by the UE we see PDR percentages of around 85% and 50% for RSSI thresholds of 50 and 40, respectively in comparison. This allows us to set thresholds for optimization without worrying about the scenario we are going to operate in like in the case of distance-based optimization, where we add more thresholds when we move to a larger simulation area.

Another critical observation we notice, especially by looking closely at Figure 4.11c which gives the PDR results for a 1500 m scenario is that as the number of UE is increased we start to see more duplication happening at similar RSSI threshold levels. This seems to be due to the lowering of the RSSI levels observed at the UEs when there are more of them present in the system. This could be happening as a result of higher interference levels observed by the UEs in the system because of the many communications happening simultaneously.

4.3.4.4 Performance degrades with an increase in the number of UEs in the system

From the results, we observe that performing packet duplication can successfully improve the packet loss and latency results for multiple UEs as well. However, we also observe an increase in the values of packet loss and latency as we increase the number of UEs being served in the system. This is expected as with the number of UEs, more data needs to be scheduled in the system and the network resources are fixed across runs. This results in additional latency for packets that get delivered and in cases of scenarios where there is low to no coverage, we see some increase in dropped packets.

Similar to the case of NR-DC with split-bearer, in packet duplication as well we observe that when multiple UEs are present, there is a significant increase in latency and packet loss of the system. When there is a large number of UEs present in the system and given that network resources available in the simulation are limited, there is a shortage of available radio resources to schedule the amount of data that is coming in. As a result, we see a saturation of MAC buffers that leads to an increase in packet loss and latency.

4.3.4.5 Full Duplication leads to rapid degradation in performance as UEs in the system increase

In the case of packet duplication when no optimization is done, and we perform full duplication of packets, we observe a decrease in performance as the number of UEs in the system is increased, as can be seen in Figure 4.12. For example, in the 1500 m scenario with RSSI-based packet duplication, we see average latency of 11.49 ms with full duplication compared to 21.448 ms in the case of 0% PDR or no duplication at 10 dBm RSSI threshold when there are 16 UEs in the system. However, when this number increases to 32 UEs, we see that the average latency for full duplication exceeds that of the case when PDR is around 0% at the RSSI threshold of 10 dBm by almost 8 ms. This trend continues as the number of UEs is increased in the system further to 64. We observe a similar trend for the packet loss metric as well. The same observation is also made for other optimization techniques as well.

This phenomenon is observed because at higher load scenarios like the case when there is a large number of UEs in the system, more duplication means that the load is significantly increased. These extra packets that need to be handled in the system lead to higher loss and latency. We also observe that this effect is more pronounced in the 1500 m scenario as compared to the 1000 m scenarios, where the results for full duplication versus no duplication stay more comparable to each other for 32 UEs and get worse for 64 UEs. The reason here is that the simulation area is smaller with lesser low mmWave-coverage areas. Thus more packets can be delivered using the larger resources of the mmWave bands, and less amount of fallback is there onto the sub-6 GHz bands with limited resources.

4.3.4.6 Maximum Latency provides a good measure of the system load

As discussed in the earlier paragraphs, the observed high packet loss and latency for 32 and 64 UEs in the 1500 m scenario and for 64 UEs in the 1000 m scenario is primarily because of the large network load. This extra data load in the system is evident by looking at the results of the maximum latency we get in the system. The minimum latency of the system, however, increases steadily as the number of UEs increase. Thus average latency and packet loss percentages we see

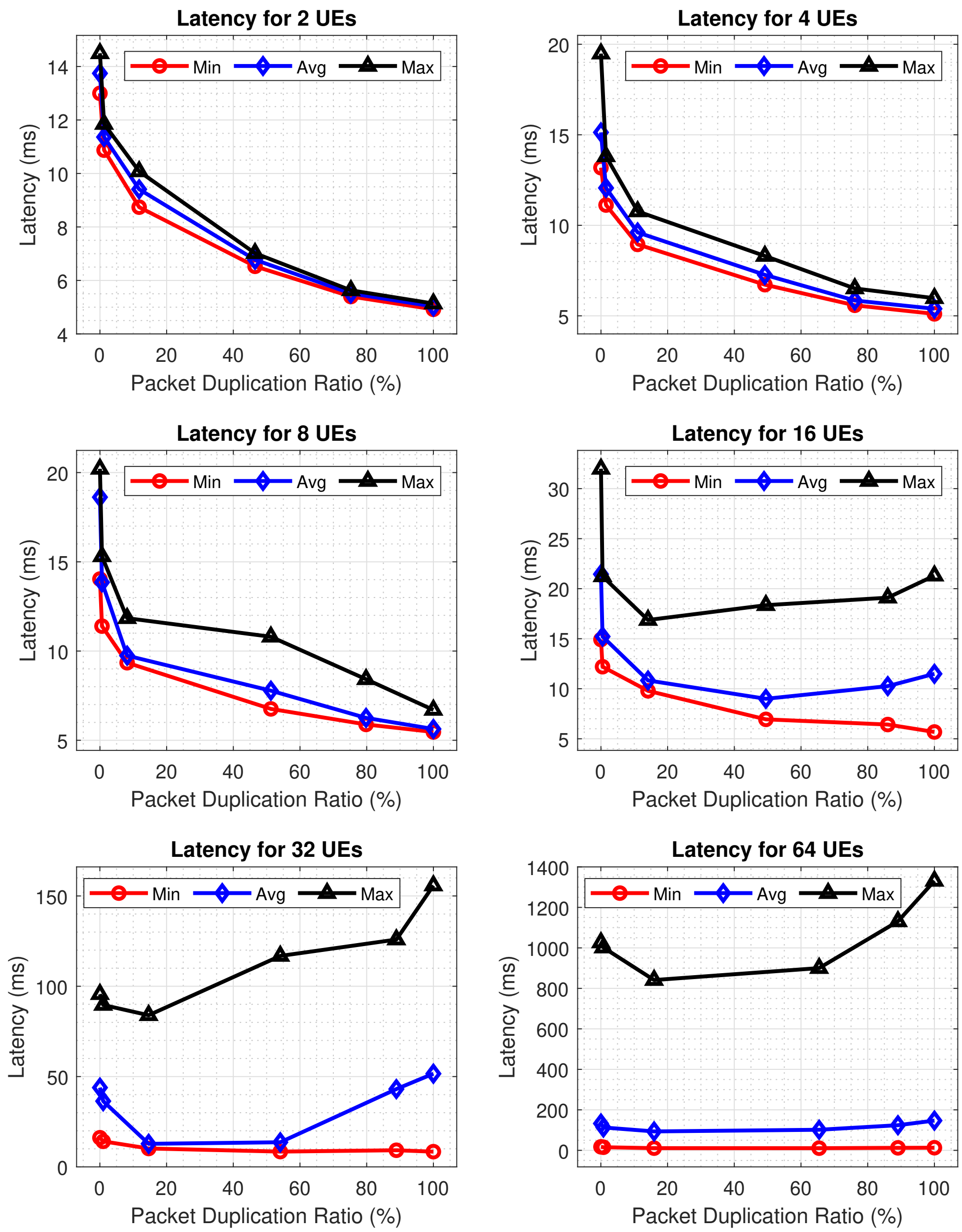


Figure 4.12: Latency vs. Packet Duplication Ratio plots with different numbers of UEs for RSSI-based optimization of packet duplication in the 1500 m scenario

can be perceived as the measure of network load. As the packet that arrives in the simulation are periodic in nature and arrive every 10 ms, some of the UE's packet is regularly scheduled ahead of the others. This leads to a steady increase in the minimum latency. However, the maximum latency is directly correlated to the time it takes for packets of a UE to be delivered successfully, and the longer the packets sit in the buffer because of congestion, the larger the maximum latency gets. We thus see a large spike or rise in the maximum latency results as the load is increased by adding more UEs in the system.

As observed earlier at high loads like in the case of 64 UEs, we noticed that PDR 0% results in a high packet loss and latency as packets coming from 64 UEs have to be scheduled using resources of only two secondary gNBs operating in mmWave range with low coverage. The performance slowly improves as more duplication is done. But when full duplication is done or even at larger duplication percentages, we again start to see a drop in performance in terms of increasing packet loss and latency. This is again because, with duplication, the load of the system significantly increases, leading to more drops and higher delays. Thus there needs to be a right balance between the threshold to be used, the network load, the radio resources available, and the performance requirements of the system.

Chapter 5

Single Grant Multiple Uplink Scheme for Performance Improvement in Mobility Scenarios

In this chapter, we analyze the current standards and architecture for uplink cellular communication and identify bottlenecks to mitigating strategies. Thus, we explore a strategy that is based on a single grant multiple uplink schemes for performance improvement in mobility scenarios in 5G networks. Section 5.1 looks into the major issues in uplink cellular transmission and identifies areas for improvement in performance. The implementation methodology and simulation details are presented in Section 5.2. Section 5.3 presents the evaluation of our novel technique in terms of the utilized performance metrics and discusses some of the key observations.

5.1 Performance Improvement in Mobility Scenarios in Uplink Scheme

Modern-day applications such as autonomous driving, live video, etc., are much more uplink intensive as compared to the traditional use cases, which have been more downlink intensive. The 5G and beyond networks need to meet the stringent demands of throughput, latency, and reliability for the uplink traffic at mobile user devices. But, it has been seen that the uplink performance is limited because of many design and architectural decisions that have been carried forward from the previous cellular generations, and thus it is paramount to improve 5G uplink performance.

5G cellular communication uses a grant-based (GB) scheduling for the uplink, where the UE in connected mode with the base station must request access to radio resources through Scheduling Requests (SR) [34]. The base stations then allocate the radio resources and notify them to the UE through UL grant messages. The UEs transmit uplink data only after receiving these grant messages, which causes delays in starting the uplink data transmission and thus, cannot meet the URLLC latency requirements [33] in the uplink direction. In grant-free (GF) scheduling, the network pre-configures the radio resources for the UEs and the UE can transmit data as soon as they have something to transmit. This enables the networks to meet the URLLC requirement, improves the energy consumption of the UEs, and reduces their complexity [20]. However, it also could lead to the wastage of radio resources as they could go unused if there is no data with the UE to perform uplink transmission. In another modification to this strategy, the BS can decide whether resources are dedicated to specific UEs or are shared by a group of UEs [32] to improve radio resource efficiency. But, this could lead to resource contention and contention-based transmission suffers from poor reliability due to collisions with other neighboring UEs transmitting simultaneously over the shared resource.

Thus, it is difficult to meet the stringent latency requirement of URLLC applications, and there is a critical need to satisfy these requirements. One of the standards-based techniques to improve transmission reliability is HARQ retransmission and authors in [56] propose a modification to this. HARQ retransmissions are done when there is a Negative ACKnowledgement (NACK) resulting from a failure in the previous transmission. But, this increases the latency as the base station first receives the packet for detection and then issues the feedback. Before performing retransmission, the UE needs to wait for the feedback which depends on the HARQ round-trip-time (RTT), that is the time duration from the beginning of the transmission until processing the feedback. In the proposed K-repetition scheme, a pre-defined number of consecutive replicas of the same packet are transmitted without waiting for the feedback, and then the BS performs soft combining of these repetitions to improve the reliability [44], but these techniques also increase the latency and lead to wastage of resources.

As discussed in earlier chapters, even though millimeter wave bands have large spectrum availability that results in higher capacity, these bands suffer from high propagation loss and are subject to blockages by building materials, foliage, the human body, etc. These challenges place limitations on mmWave deployments. We have shown that multi-connectivity provides a way to increase reliability in the downlink for ongoing 5G deployments. However, multi-connectivity schemes also somewhat increase the UE complexity as multiple stacks must operate in parallel at the PHY, MAC, and RLC layers. It also leads to higher power consumption at the UE as two independent transmissions must happen to the dual-connected base stations. Hence, there is a critical need to explore flexible network architecture to meet the stringent requirements of 5G systems in the uplink direction. We thus propose a novel single grant-based multiple uplink scheme for performance improvement in mobility scenarios. In this technique, the UE performs only a single transmission using an omnidirectional or multi-directional antenna on a common channel, and every base station that can receive this signal would accept and process it. In a dense deployment, with multiple mmWave base stations around the UE, this further removes the need to perform frequent handovers and allows high mobility. This also reduces the need to perform multiple HARQ retransmission, thus reducing the round trip time and latency. By allowing multiple copies to be processed by several centrally connected gNBs, we essentially create multiple copies of the uplink packet without spending valuable UE resources. The duplicate removal of redundant packets is performed at the PDCP layer of the central unit. The implementation of this technique makes use of the 5G CU-DU split architecture [2], which is discussed in more detail in the following Section 5.1.1.

5.1.1 CU-DU Architecture

The logical 5G CU-DU architecture of a gNB, as shown in Figure 5.1, consists of a Central Unit (CU) and one or more Distributed Units (DUs) called gNB-Central Unit and gNB-Distributed Units, respectively. The fronthaul (F1) interfaces can be split into two categories F1-C carrying the control signals and F1-U responsible for the data packets, providing control plane and user plane connectivity, respectively. The central unit is a logical node that includes the gNB functions like transfer of user data, mobility control, radio access network sharing, positioning, session management, etc., except those functions, are allocated exclusively to the DU. CU controls the operation of DUs over the fronthaul interface. On the other hand, Distributed Unit is a logical node that includes a subset of the gNB functions, depending on the functional split option, and its operation is controlled by the CU.

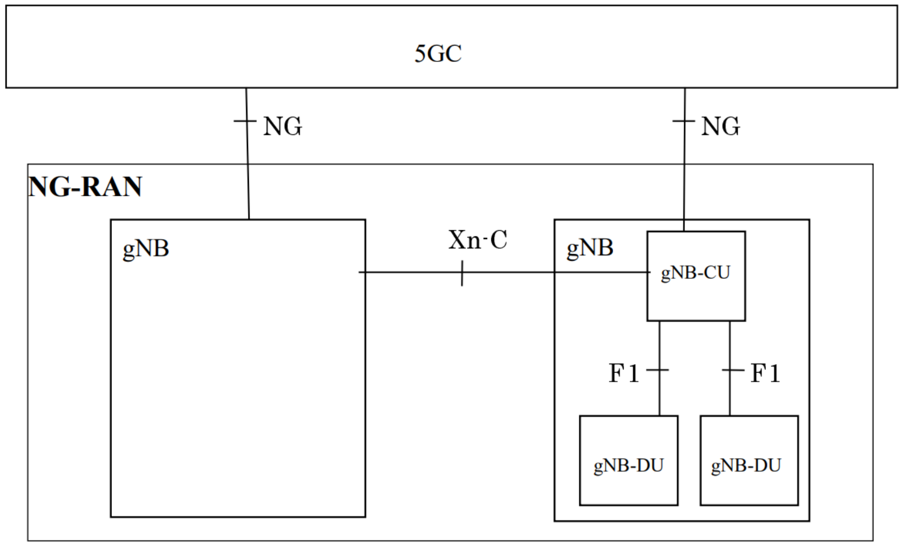


Figure 5.1: Overall Architecture of NG-RAN CU-DU Split [2]

The NG and Xn-C interfaces in Figure 5.1 for a gNB terminate in the gNB-CU [2]. The internal structure of the gNB is not visible to the core network and other RAN nodes, so the gNB-CU and connected gNB-DUs are only visible to other gNBs and the 5GC as any other gNB. The F1 interface supports signaling exchange and data transmission between the endpoints, separates the radio network layer and transport network layer, and enables the exchange of UE-associated and non-UE-associated signaling. Based on scenarios and performance requirements, RAN functions can be optimized to be present at different locations. Thus, the gNB-CU can be further separated into its control plane (CP) and user plane (UP) parts which are the gNB-CU-CP and gNB-CU-UP, respectively. The interface between CU-CP and CU-UP is called E1, which is purely a control plane interface. The gNB-CU-CP hosts the RRC and the control plane part of the PDCP protocol. It also terminates the E1 interface connected with the gNB-CU-UP and the F1-C interface connected with the gNB-DUs. The gNB-CU-UP hosts the user plane part of the PDCP protocol and the SDAP protocol of the gNB-CU for a gNB. The gNB-CU-UP similarly terminates the E1 interface connected with the gNB-CU-CP and the F1-U interface connected with the gNB-DU.

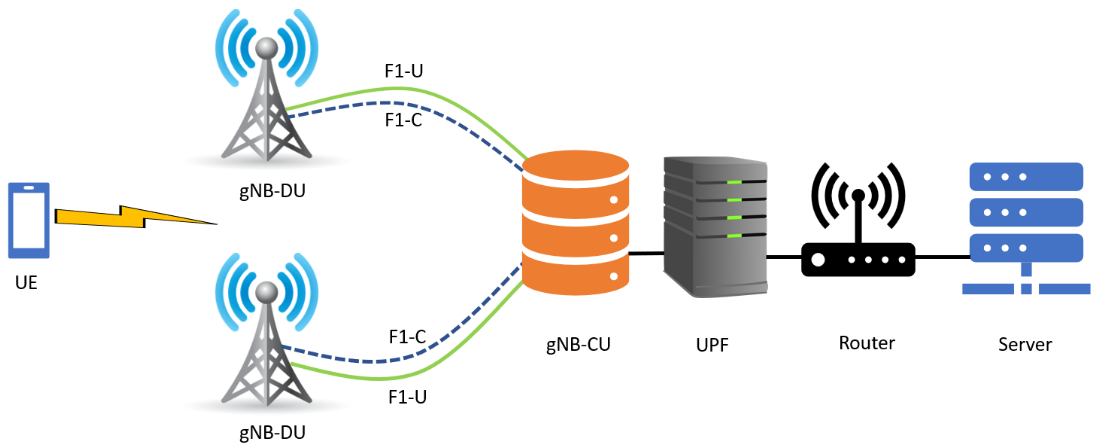


Figure 5.2: System Model for 5G CU-DU Split Architecture

5.2 Multiple Uplink Scheme and Simulation Methodology

Figure 5.2 shows our proposed architecture which consists of a gNB-CU that is connected to multiple mmWave-capable gNB-DUs. The RRC layer in the gNB-CU-CP is responsible for configuring the uplink grants for a given UE and will provide this UL grant information to all the gNB-DUs via the F1-C interface such that all the distributed units in the range of the UE's uplink communication are able to receive and accept the transmission. The UE uses this UL grant to perform a single transmission that can be received by multiple base stations. Each base station independently receives and processes these packets. The packet travels up through the PHY, MAC, and RLC layers on the gNB-DU stack. Each receiving distributed unit's RLC layer would then send the PDUs to the common PDCP layer in the gNB-CU-UP, where sequence numbering-based duplicate removal is performed. Only the earliest arriving packet is processed by the PDCP layer and passed on to the upper layers, whereas the copies of the PDU from other distributed units bearing the same sequence number that arrive at a later time are discarded.

Figure 5.3 and Figure 5.4 show the multilink process flow at the sender and receiver sides, respectively. At the sender, when a packet from the application layer of the UE arrives at the cellular protocol stack, it first checks if the multilink transmission is enabled for the packet. This decision can be based on the type of application and must be taken by the network if it wants to improve network reliability on the uplink. If enabled, the corresponding MultiLinkGroup ID is added to the packet by the UE before being transmitted. The MultiLinkGroup ID must be communicated to the

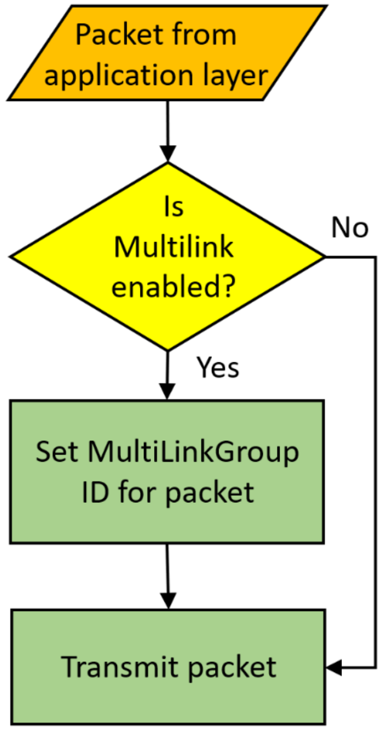


Figure 5.3: Multilink Process at the Sender

UE by the network as part of the PDU session establishment procedure and would indicate which gNB-DUs would be participating in the scheme. Further, the gNB-CU can decide to change the MultiLinkGroup ID through a PDU session update or could choose to dynamically add and remove gNB-DUs from the group based on traffic and congestion in the network.

At the receiver side, the gNB-DU checks the packet to see if the multilink scheme was enabled for the transmission by the UE. If multilink is not enabled, the packet is delivered straight to the higher layers. Else, the gNB checks if it is part of the MultiLinkGroup and deletes the packet if it is not. Otherwise, it checks if the packet sequence number is already present in the list of delivered packets. If the packet sequence number is present, meaning the packet was already delivered, it deletes the packet. Else, it adds the packet sequence number to the list of delivered packets and delivers the packet to the higher layers. The PDCP layer in the gNB-CU is responsible for in-order delivery and forwards the first successfully received PDU to the higher layers while discarding any duplicates that arrive later from the DUs.

In the deployment scenario, we consider a dense deployment of gNB-DUs such that at any

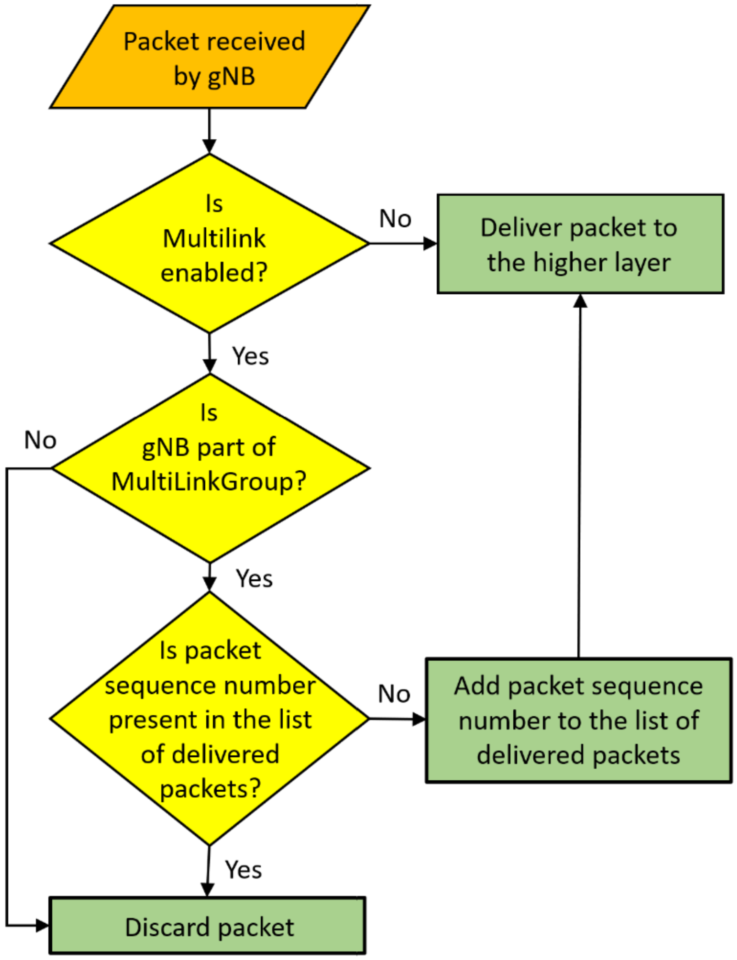


Figure 5.4: Multilink Process at the Receiver

given time, at least one gNB-DU is expected to be in the range of the UE. As the UE moves, it continues transmission using the same UL grant that was provided by the gNB-CU-CP RRC layer, and thus all the distributed units connected to this central unit are aware of the transmission and can receive it. Therefore, the protocol increases the probability of successful transmission and hence, increases the reliability and throughput of the network. Additionally, it also improves the latency in the network because of the lesser number of handovers between densely deployed mmWave base stations and lesser HARQ retransmissions. Furthermore, the introduction of this form of transmission redundancy does not adversely affect spectrum resource utilization. The proposed architecture was implemented and evaluated using Simu5G and is described in the next section.

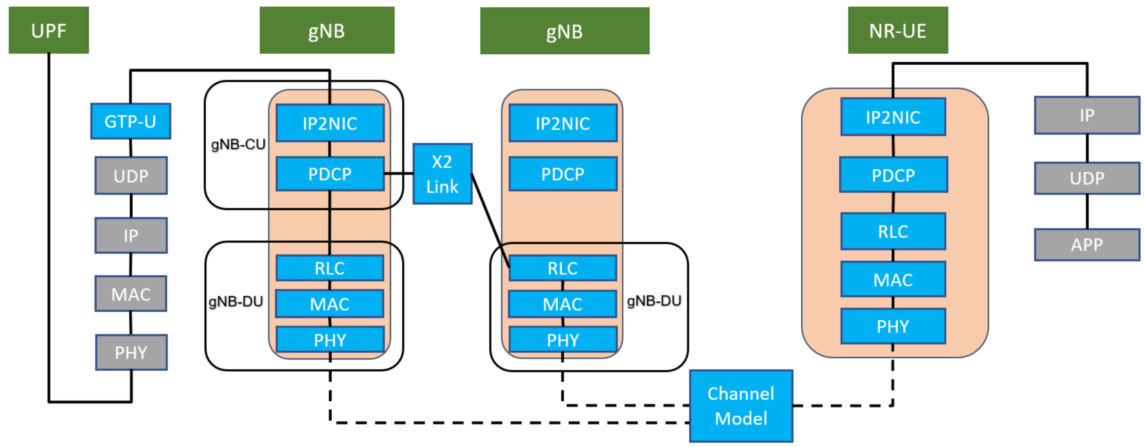


Figure 5.5: Simu5G Multilink Deployment Architecture

5.3 Evaluation

The performance of the proposed system was systematically evaluated for the single link (SL) and multilink (ML) architecture at 28 GHz under various mobility scenarios using the Simu5G simulator. A server is connected to the UPF via a router and receives data packets from the UE in our system. In the single-link architecture, packets from a UE are received by a single base station, and in the multilink system, the packets are received by multiple base stations. UE in the single-link system undergoes handovers during the simulation depending on the observed RSSI from the serving and target cells.

We first analyze the multilink uplink architecture by simulating a single UE in the system. We then run experiments to measure the performance of our proposed system for multiple UEs. We run experiments with 2, 4, 8, 16, and 32 UEs. Throughput, latency, and packet loss were used as metrics to evaluate the system and compare the performance of the different architectures. We also run another experiment to measure the number of handovers occurring in the single-link system to see how much resources and time the multilink system saves by not undergoing handovers.

Based on the existing architecture and protocol the latency in the uplink direction are usually higher than in the downlink direction. However, 5G architecture offers the flexibility to use higher numerology to change the sub-carrier spacing thereby reducing the observed latency. In this chapter evaluation, therefore, we run experiments to measure UL performance at numerology 3. We study the performance of these systems by simulating the data plane of an end-to-end 5G network

Table 5.1: Simulation Parameters for Multilink Deployment

Parameter	**Value**
mmWave Carrier Frequency	28 GHz
mmWave Bandwidth	20 MHz
UDP Packet Size	32 B
UDP Packet Inter-arrival Time	10 ms

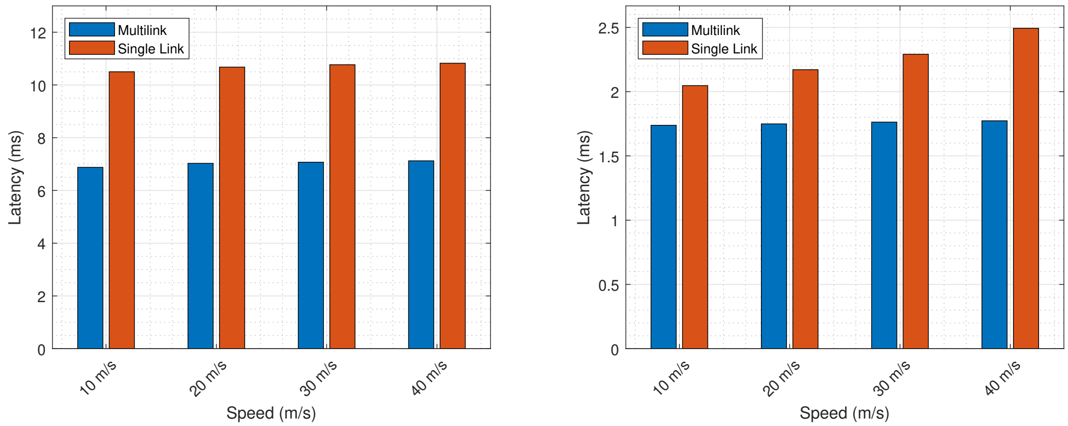
that transmits packets from a UE to a remote server. Figure 5.5 shows the multilink deployment using Simu5G.

5.3.1 Experimental Setup

The performance of our proposed multilink architecture was compared to the traditional single-link architecture for URLLC-type applications. The packet size was taken as 32 B, and the packet inter-arrival time is 10 ms. The channel bandwidth is taken as 20 MHz. The gNBs operate at a mmWave carrier frequency of 28 GHz and the path loss scenario considered for our experiments is based on the 3GPP recommended model of Urban Microcell. Three gNBs were considered in the simulation at coordinates of (100 m, 300 m), (250 m, 300 m), and (400 m, 300 m), and the UEs move along the x-axis to a distance of 500 m at varying speeds. The y-coordinates of the UE's starting position are changed for each UE in the multi-UE experiments but kept close together. Three user mobility patterns and four different speeds are considered for the experiments. For very high mobility cases, the speed of the users is set to 40 m/s, and for high mobility cases, it is set to 20 m/s and 30 m/s. The speed of the users is set to 10 m/s for the slow mobility scenarios. In the case of the experiment to count the number of handovers, we use the *RandomWayPoint* mobility model at the four different speeds to compare the savings in the multilink system. Additional simulation parameters used in our experiments are shown in Table 5.1.

5.3.2 Results for Single UE

We first present the results of our simulation for a single UE in the proposed multilink system architecture. We compare these with results from our experiments using the single-link



(a) Numerology 0 (b) Numerology 3

Figure 5.6: Latency vs UE speed result comparison in the uplink for the multilink and the single-link systems for different numerology

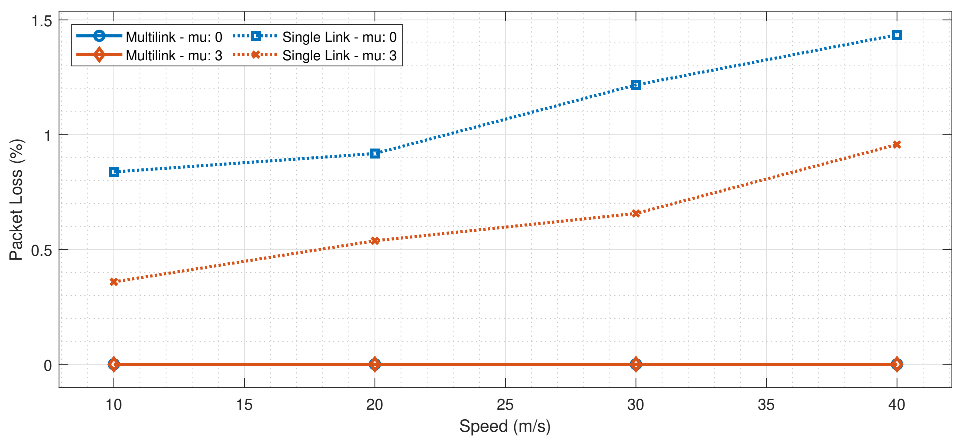
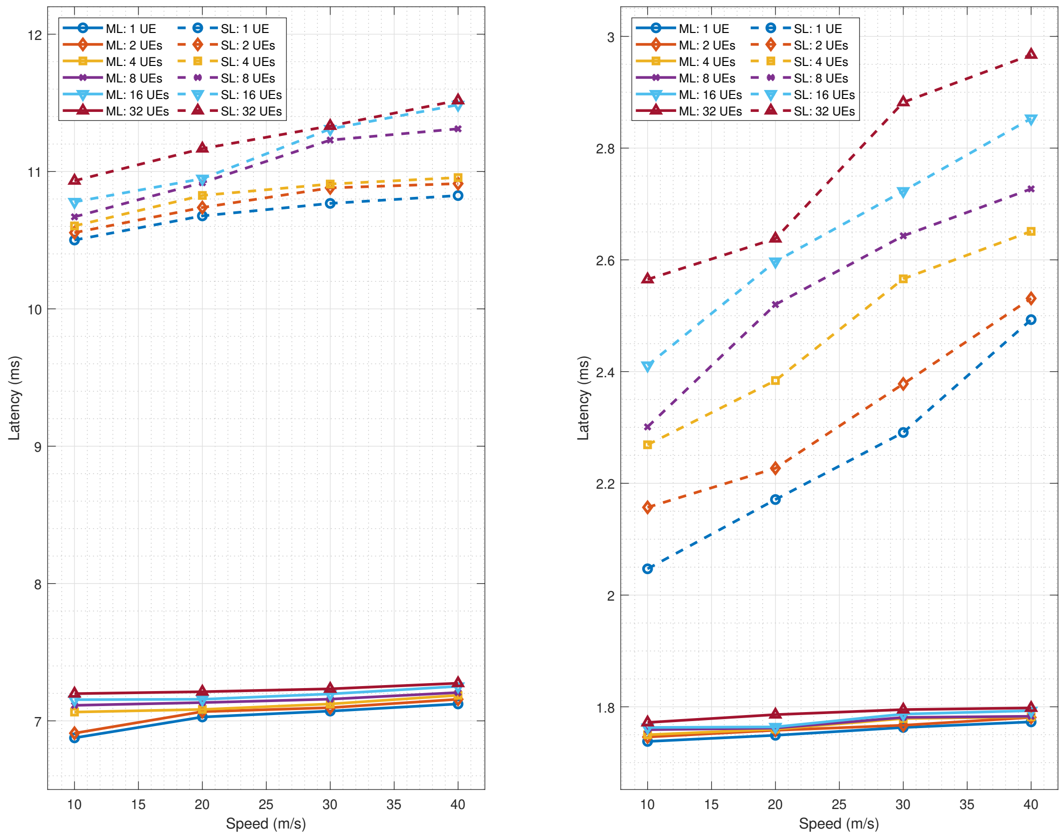


Figure 5.7: Packet Loss vs UE speed result comparison in the uplink for the multilink and the single-link systems for different numerology

system. The latency results are compared for numerology 0 and 3 in Figure 5.6. We present a bar graph comparison of latency values in the multilink and single-link systems at different speeds. We also plot the packet loss results versus the speed of the UEs from these experiments in Figure 5.7.

We observe that the multilink system results in lower latency of 6.876 ms for numerology 0 at a UE speed of 10 m/s compared to 10.501 ms in the case of the single-link system at 10 m/s. When the system uses numerology 3 then we see a reduction in latency to 1.738 ms in the case of multilink and 2.047 ms in the case of single link at 10 m/s. The latency increases as the speed of the UE is increased in both the numerology cases. For example, in the case of a multilink system,

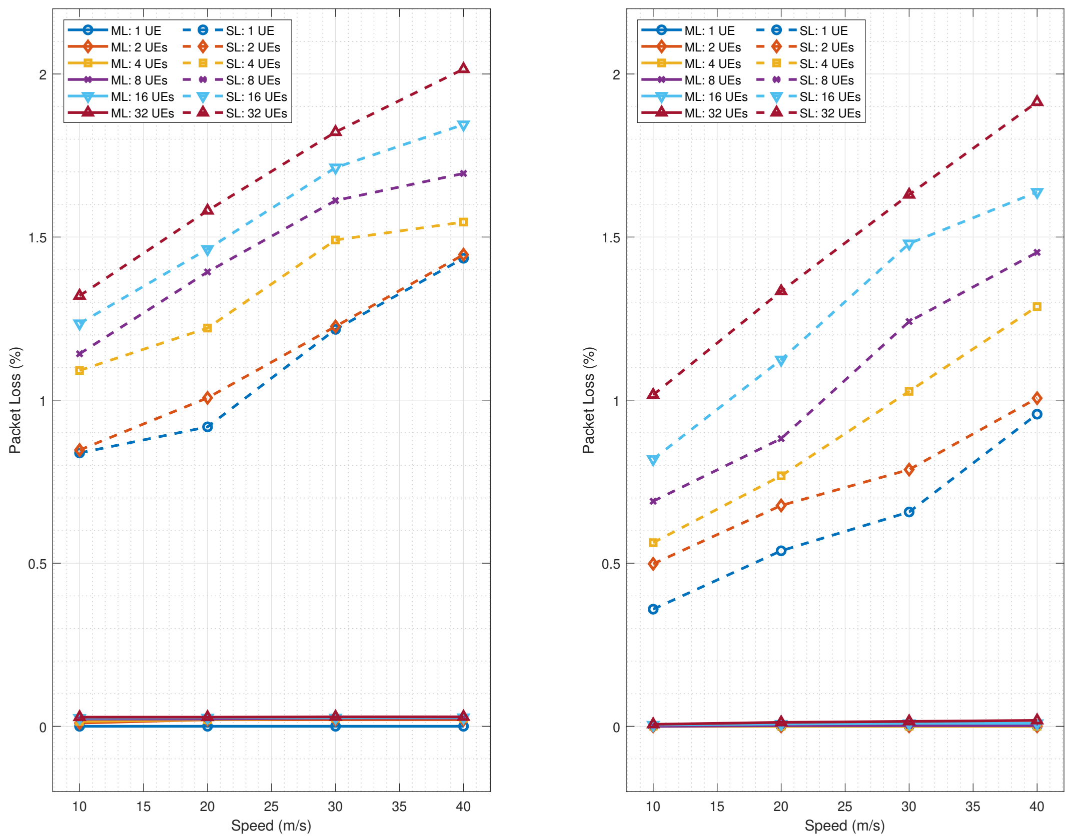


(a) Numerology 0

(b) Numerology 3

Figure 5.8: Latency vs UE speed result comparison in the uplink for the multilink and the single-link systems for different numerology

the latency rises by almost 0.25 ms when the speed of the UE increases to 40 m/s for numerology 0. In comparison to the single link system, this increase is 0.324 ms. We observe similar results for the packet loss as well. In the case of the multilink system for a single UE, we see that packet loss is zero for all levels of mobility. Whereas in the single-link system, the packet loss observed is 0.838% for a speed of 10 m/s and increases up to 1.435% when the speed increases to 40 m/s. For numerology 3, as well, we see the packet loss is zero for the multilink system and for the single-link system, the packet loss at 10 m/s is observed to be 0.359% and rises up to 0.957% for 40 m/s.



(a) Numerology 0

(b) Numerology 3

Figure 5.9: Packet loss vs UE speed result comparison in the uplink for the multilink and the single-link systems for different numerology

5.3.3 Results for Multiple UEs

Simulations are conducted using the multilink and single-link systems for varying numbers of UEs at different speeds. The latency results in the multilink and single link system for a numerology value of 0 and 3 are shown in Figures 5.8a and 5.8b, respectively. Similarly, the packet loss results in the multilink and single link systems are shown in Figures 5.9a and 5.9b for the numerology value of 0 and 3, respectively.

It is observed that as the number of UEs in the system increases the latency and packet loss also start to increase. Although we see that the rise is more in the case of single link as compared to multilink. We also see that the performance deteriorates with speed as both latency and packet loss increases with it. We again observe that for numerology 3 the latency and packet loss are significantly lower. The packet loss for numerology 3 remains at zero for up to 4 UEs in the system for all UE speeds in the multilink system. As the UEs in the system increase we see some packets lost as the load in the system starts to increase.

5.3.3.1 Multilink system outperforms the single link system

From the above results, we can clearly observe that the multilink system outperforms the single-link system in all the metrics. The system provides a low latency below 1.8 ms and low packet loss below 0.018% if numerology is taken to be 3. Also for a URLLC type of application, we can see that the system scales very well as the number of UEs is increased and is able to achieve high reliability with low packet loss and low latencies. Additionally, this increase in the system performance does not affect the spectrum efficiency as there is only a single uplink transmission. Moreover, compared to the single-link system the multilink system shows lower latency as there are no handovers and fewer retransmissions happening.

Our proposed system adds to the overall processing time at the gNBs on the network level, as every packet is processed independently by all receiving gNBs and passed onto the common PDCP layer. Hence, our system should perform well until all participating gNBs have a limited number of UEs to serve with applications that transit small packets. As these numbers increase the performance may start to deteriorate. From our simulation, we show here that for URLLC-type applications and for 20 MHz bandwidth our system can effectively serve at least 32 UEs in the system.

5.3.3.2 Performance improvement in terms of handovers

Figure 5.10 shows the number of handovers that were executed during the simulation of the single link system as the number of users in the system increases. In comparison, the proposed multilink system would not execute any handovers, which improves the latency and reliability of the system. We observe that the number of handovers in the system increases with speed as the simulation time is kept the same for all four cases. As the number of UE increases this statistic plays an important role as every handover adds to the latency and generates additional control signaling in the network that adds to the traffic. By using the multilink system we are able to reduce all the handover-related traffic and remove the latency.

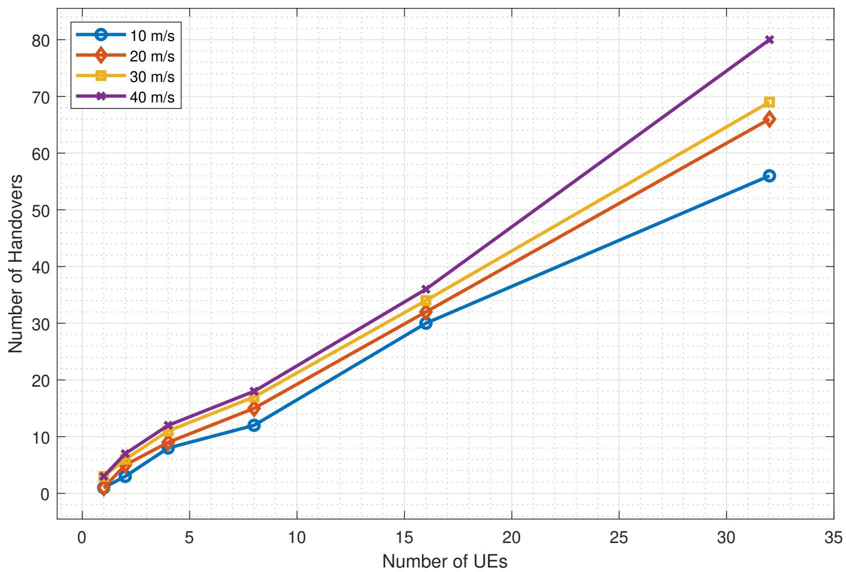


Figure 5.10: Number of handovers occurring at different speeds in single link system vs. number of users for a packet size of 32 B at packet inter-arrival time of 10 ms and numerology 0

Thus, multilink enables 5G networks to meet the requirements of high throughput and reliable low latency applications for different levels of UE mobility. The system is able to achieve high reliability with low packet loss and low latencies. We observe that our system was able to scale well for small 32 B packets for up to 32 UEs. Additionally, this increase in the system performance does not affect the spectrum efficiency as there is only a single uplink transmission. The proposed multilink system performs well for all the scenarios and shows great potential for next-generation applications and services.

Chapter 6

Conclusion

Cellular communication needs to handle real-time communication and data transmission for an enormous number of users and large amounts of data. 5G networks and beyond have the potential to support and enable new applications in areas like artificial intelligence (AI), the Internet of Things (IoT), etc. 5G and beyond 5G networks must also account for UE's mobility as they are expected to maintain their level of performance under different mobility scenarios and perform better than traditional networks. Although 5G technology has developed significantly in recent years, in this work, we showed that there still exists a critical gap in understanding how all these technologies would perform under mobility. There is also a need to analyze and identify issues that arise with scaling these networks to serve multiple UEs simultaneously. Hence, in this work, we come up with solutions to overcome these hurdles without compromising the performance of these networks. To address the network issues that come with mobility and fill that gap, this dissertation investigated next-generation networks from three distinct perspectives. We first analyzed the issue of mobility in 5G networks from the perspective of network architecture and proposed a multi-connectivity-based architecture. Next, we proposed an enhancement using packet duplication to improve the performance of the multi-connectivity architecture to support enhanced mobility. Finally, we explored a novel architecture focusing on a single grant-based multiple uplink transmission and its performance under mobility. We also analyzed each of these architectures and strategies by increasing the network load in terms of performance under multiple UEs in the system.

6.1 Mobility Enhancement using Multi-Connectivity in 5G Networks

Traditional cellular networks that cover large geographical areas struggle to deliver the dense coverage, low latency, and high bandwidth required by some 5G applications. User mobility directly impacts the design and performance of cellular wireless networks in terms of handovers, resource management aspects like channel allocation, traffic load estimation, etc., and radio propagation aspects like signal strength variation, interference level, etc. The mobility protocols ensure the delivery of useful data and control signals via a radio interface link to the desired user when commuting between coverage areas. However, we show that providing seamless mobility to all devices in a diverse, dense, multi-tier network like 5G with traditional protocols has repercussions on network performance. The 5G and next-generation cellular networks require flexibility in the network topology to ensure the level of performance that is expected of such services. Balancing the trade-offs among handoff rate, service delay, and achievable coverage/data rate in heterogeneous, dense, and diverse 5G cellular networks is, therefore, an open challenge that has been analyzed in this dissertation. Thus, in this work, we studied how multi-connectivity enhances mobility in 5G Networks.

Multi-connectivity (MC) is shown to be a key enabler in improving the performance of these next-generation networks. Simultaneous connectivity to multiple radio access technologies or multiple bands must be an important solution for ongoing and future 5G deployments. It not only enhances the system performance by providing multiple simultaneous links between the user equipment (UE) and the base stations (BS) for data transfer but also eliminates the time needed to deal with frequent handoffs, link establishment, etc. Hence, in this dissertation, we show how mobility impacts the performance of current UDN architecture in a city environment and discuss improvements for reducing the impact of mobility to meet 5G specifications using multi-connectivity. We further analyzed how UDN and NR-DC deployments scale with the number of users in the system. We also looked into how this new approach impacted handovers in a multi-connectivity-based NR-DC deployment and a mmWave-based UDN deployment.

6.2 Improving Reliability in Mobility Scenarios using Multi-Connectivity with Packet Duplication

The next-generation cellular networks also require ultra-reliability to ensure the reliability of modern mission-critical services is not affected. Multi-connectivity could potentially enhance the system performance by providing multiple simultaneous links between the user equipment (UE) and the base stations. We showed in our work that in such systems, packet duplication can be used to meet the stringent reliability and latency requirements of modern cellular networks. This is possible as data packets were duplicated and transmitted concurrently over two independent links enhancing their reception probability. Thus, in this part of the dissertation, we analyzed the system and evaluated the performance with packet duplication under various mobility scenarios for reliable low-latency communication. We also discussed the choice of using the PDCP sub-layer in the creation and removal of duplicate, redundant packets over other proposed mechanisms.

We then proposed an enhancement to the packet duplication feature based on three different UE metrics that vary in accuracy and collection complexity. We analyzed optimization techniques based on UE-gNB distance, UE velocity, and the power level reported by the UE for the attached gNB to improve radio resource utilization. In this work, we showed that packet duplication can be used to meet the stringent reliability and latency requirements of modern cellular networks, and optimization can be used to balance the performance with spectrum usage based on network requirements. This dissertation also looked in detail into the effects of scaling the packet duplication and its optimization to multiple UEs. We showed that this scheme performed well for a large number of UEs, but the performance would be compromised at some point depending on the resource availability in the network.

6.3 Single Grant Multiple Uplink Scheme for Performance Improvement in Mobility Scenarios

The 5G and beyond 5G networks serve diverse use cases and larger uplink traffic, including applications such as autonomous driving, IoT applications, live video, etc., which are much more uplink intensive as compared to traditional applications. We discussed how the uplink performance falls behind the downlink performance, and there is a need in 5G to improve the uplink performance.

Thus, in this work, we proposed a novel uplink scheme where the UE performs only a single transmission on a common channel, and every base station that can receive this signal would accept and process it. This scheme challenges the existing architecture and norms of cellular networks, but we show with results that such out-of-the-box thinking is critical to achieving the expected performance.

In this work, our proposed architecture simulated a UE that was connected to multiple mmWave capable distributed units (DUs), which were connected to a single gNB-central unit. In an ultra-dense deployment with multiple mmWave base stations around the UE, this mitigated the need to perform frequent handovers and allowed high mobility with reduced latency. Our work in this dissertation showed that the proposed system improved the performance in the uplink direction that could meet future requirements.

This dissertation work studied the impact of mobility and scalability on next-generation networks. This study develops and systematically analyzes the overall mobility and scalability problem with distinct requirements. We developed solutions that involve a deep understanding of the current architecture, deployment strategy, and behavior patterns under mobility and increased load. These studies are intended to understand better the network requirements for handling mobility and propose solutions for handling scalable real-world networks.

Appendices

Appendix A

Additional Results for Packet duplication with Multiple UEs

We present and compare the additional results of our experiment conducted on the NR-DC deployment with Packet Duplication and different optimization techniques. We present the network performance metrics for three forms of packet duplication optimizations with Multiple UEs i.e. Distance based, Velocity-based, and RSSI-based. The results are presented for different duplication thresholds with two different simulation areas of 1000 m and 1500 m. Each deployment scenario has two pairs of master plus secondary gNBs with a packet inter-arrival time of 10 ms similar to what was discussed in section 4.3.4. Thus, each table shows the comparison of performance under different gNB densities with the rows showing different thresholds and different numbers of UEs in the system. We present the results for the packet duplication ratio, minimum latency, maximum latency, average latency, and packet loss.

A.1 Distance-based Optimization

We present the detailed results for distance-based Packet duplication optimization in Table A.1 and A.2. For the 1000 m scenario, we present results for full duplication and distance thresholds of 150 m, 250 m, and 350 m. Whereas for the 1500 m scenario, we present results for full duplication and distance thresholds of 150 m, 250 m, 350 m, 450 m, 550 m, and 650 m.

Table A.1: Performance comparison for distance-based optimization of packet duplication in a 1000 m scenario in an NR-DC deployment with multiple UEs in the system

Number of UEs	**Distance Threshold**	**PDR (%)**	**Latency**			**Packet Loss (%)**
			Min (ms)	**Max (ms)**	**Avg (ms)**	
2	**FD**	100.00	4.73	5.05	4.89	0.00
	150	91.67	5.06	5.28	5.17	0.00
	250	70.99	5.26	5.47	5.36	0.01
	350	35.25	6.48	7.05	6.77	0.07
4	**FD**	100.00	4.90	4.95	4.92	0.00
	150	91.10	5.13	5.38	5.25	0.01
	250	72.01	5.53	5.89	5.71	0.03
	350	35.11	6.63	6.98	6.91	0.26
8	**FD**	100.00	4.99	7.92	5.45	0.08
	150	88.81	5.14	10.16	6.14	0.51
	250	66.91	5.72	11.79	7.26	1.26
	350	34.26	6.73	14.38	8.25	1.97
16	**FD**	100.00	5.47	17.53	6.85	1.08
	150	88.53	6.02	25.96	7.49	2.23
	250	64.15	6.17	24.49	8.83	2.92
	350	40.54	7.78	18.09	9.48	3.56
32	**FD**	100.00	6.74	59.43	9.09	4.86
	150	90.09	6.87	54.78	9.83	5.65
	250	74.65	7.17	42.80	10.29	5.89
	350	43.02	9.10	39.89	10.59	5.84
64	**FD**	100.00	10.14	922.63	75.89	10.69
	150	89.65	10.08	893.24	73.56	10.51
	250	73.67	10.22	726.99	63.61	9.16
	350	38.41	11.23	681.64	68.93	10.15

Table A.2: Performance comparison for distance-based optimization of packet duplication in a 1500 m scenario in an NR-DC deployment with multiple UEs in the system

Number of UEs	Distance Threshold	PDR (%)	Latency			Packet Loss (%)
			Min (ms)	Max (ms)	Avg (ms)	
2	**FD**	100.00	4.92	5.54	5.23	0.00
	150	96.15	5.01	5.92	5.47	0.00
	250	83.91	5.52	6.57	6.05	0.03
	350	75.53	6.26	6.62	6.44	0.18
	450	53.90	7.43	7.78	7.60	0.80
	550	34.51	8.82	9.10	8.96	2.72
	650	18.76	9.65	9.77	9.71	2.94
4	**FD**	100.00	5.11	5.94	5.32	0.00
	150	93.78	5.17	7.00	5.73	0.03
	250	83.41	5.56	8.53	6.49	0.08
	350	75.15	6.28	8.61	6.90	0.25
	450	55.64	7.49	8.73	7.83	1.30
	550	34.04	8.98	9.06	9.12	2.98
	650	24.45	9.99	9.27	9.93	3.28
8	**FD**	100.00	5.29	8.80	5.94	0.02
	150	97.69	5.25	9.04	6.14	0.03
	250	90.71	5.87	10.55	6.81	0.10
	350	70.24	6.38	10.85	7.31	0.31
	450	60.91	7.66	11.21	8.44	1.36
	550	43.25	9.33	11.89	9.76	3.97
	650	14.13	10.80	12.50	11.10	4.06
16	**FD**	100.00	5.39	18.68	12.54	1.74
	150	94.40	5.71	16.35	11.73	2.50
	250	88.13	6.40	17.86	10.88	3.01
	350	76.74	6.90	16.93	8.41	4.68
	450	64.66	8.47	17.52	9.59	5.65
	550	47.97	10.00	18.91	10.56	6.45
	650	21.93	11.46	21.12	12.19	7.40
32	**FD**	100.00	8.96	119.48	51.22	11.80
	150	97.15	9.79	107.76	43.78	10.85
	250	89.96	8.45	104.30	39.87	9.06
	350	81.02	8.41	98.65	36.03	8.09
	450	71.44	9.34	76.57	24.46	7.61
	550	47.31	10.98	56.33	15.26	7.99
	650	23.12	12.50	68.29	33.89	10.86
64	**FD**	100.00	14.09	1079.67	148.38	20.11
	150	94.88	13.69	1094.35	133.02	17.25
	250	89.85	13.30	984.29	123.30	16.31
	350	77.69	12.97	997.68	113.82	15.38
	450	68.16	10.74	610.63	90.69	13.65
	550	48.54	12.05	629.24	70.65	14.23
	650	27.64	18.56	895.12	88.34	16.61

A.2 Velocity-based Optimization

We present the detailed results for velocity-based Packet duplication optimization in Table A.3 and A.4. Results are shown for full duplication and velocity thresholds of 5, 15, 25, and 35 m/s for both 1000 m and 1500 m scenarios.

Table A.3: Performance comparison for velocity-based optimization of packet duplication in a 1000 m scenario in an NR-DC deployment with multiple UEs in the system

Number of UEs	Velocity Threshold	PDR (%)	Latency Min (ms)	Latency Max (ms)	Latency Avg (ms)	Packet Loss (%)
2	**FD**	100.00	4.78	5.00	4.89	0.00
	5	89.95	4.81	5.04	4.93	0.02
	15	56.53	5.34	6.07	5.70	0.49
	25	22.87	6.39	7.27	6.83	1.63
	35	10.88	6.82	7.29	7.05	2.38
4	**FD**	100.00	4.95	4.98	4.96	0.00
	5	91.40	4.98	5.09	5.01	0.20
	15	58.85	5.65	6.66	5.95	1.21
	25	26.32	6.65	8.05	7.00	2.13
	35	9.32	7.01	7.65	7.18	3.07
8	**FD**	100.00	5.02	6.11	5.21	0.01
	5	78.72	5.19	6.79	5.44	0.24
	15	51.64	5.92	9.02	6.37	1.83
	25	30.28	7.19	10.82	7.91	2.97
	35	5.67	7.67	12.11	8.39	4.37
16	**FD**	100.00	5.60	9.57	5.89	0.80
	5	80.08	5.84	9.90	6.12	0.84
	15	52.05	6.42	10.43	6.93	2.16
	25	24.16	7.68	13.94	8.61	3.76
	35	6.48	8.28	16.89	9.39	5.25
32	**FD**	100.00	6.25	77.51	8.98	3.61
	5	83.53	6.54	69.82	9.18	3.43
	15	56.30	6.97	45.63	9.65	3.15
	25	27.26	8.39	29.45	9.62	4.15
	35	10.40	8.97	40.28	10.48	8.80
64	**FD**	100.00	12.27	1090.49	83.38	12.32
	5	81.53	11.67	864.13	79.90	11.84
	15	59.62	10.37	681.11	73.74	5.07
	25	27.81	10.73	760.05	59.89	5.76
	35	10.21	10.99	869.24	64.27	9.64

Table A.4: Performance comparison for velocity-based optimization of packet duplication in a 1500 m scenario in an NR-DC deployment with multiple UEs in the system

Number of UEs	Velocity Threshold	PDR (%)	Latency Min (ms)	Latency Max (ms)	Latency Avg (ms)	Packet Loss (%)
2	**FD**	100.00	4.92	5.34	5.13	0.01
	5	89.95	4.97	5.46	5.51	0.06
	15	36.31	5.16	7.47	6.31	1.22
	25	19.10	6.84	7.76	7.30	3.12
	35	7.76	8.78	10.39	9.59	5.16
4	**FD**	100.00	5.11	5.67	5.39	0.08
	5	87.55	5.09	7.95	6.02	0.18
	15	38.32	5.61	8.33	6.97	2.09
	25	18.79	7.28	9.57	7.93	3.85
	35	6.75	9.02	12.84	10.33	5.82
8	**FD**	100.00	5.57	5.90	5.63	0.21
	5	88.55	5.73	8.77	6.75	0.54
	15	39.43	6.15	9.01	7.33	3.09
	25	17.98	7.87	11.11	8.35	4.25
	35	5.61	9.88	14.29	11.99	6.18
16	**FD**	100.00	6.14	18.68	11.91	2.56
	5	84.44	6.58	17.96	10.37	2.95
	15	40.61	7.00	14.76	9.38	4.35
	25	23.55	8.69	16.49	12.09	5.49
	35	4.98	10.89	17.65	15.27	7.52
32	**FD**	100.00	8.53	113.86	58.60	10.43
	5	76.32	9.23	108.01	49.12	9.79
	15	42.08	10.42	94.52	17.97	8.46
	25	22.85	14.24	101.95	18.59	9.37
	35	5.00	16.75	100.28	27.01	12.01
64	**FD**	100.00	12.92	1365.17	147.54	20.77
	5	78.82	12.27	1138.40	128.33	18.97
	15	44.66	11.76	845.00	89.58	15.69
	25	21.60	16.52	921.46	97.99	16.17
	35	6.80	18.47	1045.16	107.81	19.14

A.3 RSSI-based Optimization

The detailed results for RSSI-based Packet duplication optimization are presented in Table A.5 and A.6. Results are shown for full duplication and RSSI thresholds of 50, 40, 30, 20, and 10 for both 1000 m and 1500 m scenarios.

Table A.5: Performance comparison for RSSI-based optimization of packet duplication in a 1000 m scenario in an NR-DC deployment with multiple UEs in the system

Number of UEs	RSSI Threshold	PDR (%)	Latency Min (ms)	Latency Max (ms)	Latency Avg (ms)	Packet Loss (%)
2	**FD**	100.00	4.68	5.10	4.89	0.00
	50	59.27	5.30	6.35	5.83	0.21
	40	18.54	6.67	7.22	6.94	2.19
	30	2.51	7.74	8.21	7.98	2.78
	20	0.54	8.07	8.42	8.24	3.38
	10	0.00	8.60	9.15	8.87	4.31
4	**FD**	100.00	4.85	5.00	4.96	0.00
	50	59.35	5.47	6.97	6.23	0.42
	40	24.32	6.76	8.43	7.54	2.69
	30	1.62	7.85	9.14	8.19	3.53
	20	0.32	8.26	9.73	8.64	4.74
	10	0.00	8.61	10.36	9.06	5.07
8	**FD**	100.00	4.99	6.11	5.15	0.32
	50	64.48	5.77	8.29	6.48	1.30
	40	17.09	6.80	9.07	7.93	3.03
	30	2.50	7.98	10.87	8.69	4.55
	20	0.69	8.80	13.07	9.83	5.82
	10	0.00	9.42	15.69	10.53	7.25
16	**FD**	100.00	5.24	10.17	5.85	1.39
	50	58.44	6.27	11.31	6.93	1.80
	40	19.64	7.11	15.46	8.28	3.86
	30	2.07	8.23	22.03	9.18	5.29
	20	0.64	9.11	24.46	10.28	6.85
	10	0.00	9.99	29.24	11.61	9.08
32	**FD**	100.00	5.94	71.45	9.09	4.87
	50	64.80	6.88	67.67	9.18	4.39
	40	15.77	7.57	32.86	9.97	4.82
	30	1.78	8.92	33.95	12.44	5.84
	20	0.77	9.87	34.86	14.97	7.21
	10	0.00	10.64	42.73	16.89	10.85
64	**FD**	100.00	11.50	977.35	84.92	12.86
	50	69.40	11.34	879.56	73.45	9.16
	40	17.43	10.87	554.86	49.86	7.46
	30	1.39	10.81	626.55	69.68	8.72
	20	0.43	11.87	754.86	89.86	9.95
	10	0.00	11.87	837.96	95.85	13.93

Table A.6: Performance comparison for RSSI-based optimization of packet duplication in a 1500 m scenario in an NR-DC deployment with multiple UEs in the system

Number of UEs	**RSSI Threshold**	**PDR (%)**	**Latency**			**Packet Loss (%)**
			Min (ms)	**Max (ms)**	**Avg (ms)**	
2	**FD**	100.00	4.92	5.14	5.03	0.01
	50	75.26	5.40	5.63	5.51	0.61
	40	46.51	6.53	7.02	6.77	2.89
	30	11.76	8.74	10.08	9.41	3.90
	20	1.25	10.87	11.85	11.36	5.28
	10	0.00	13.00	14.49	13.74	5.75
4	**FD**	100.00	5.11	5.97	5.39	0.08
	50	76.12	5.58	6.51	5.85	0.97
	40	49.15	6.72	8.31	7.26	3.17
	30	10.99	8.94	10.77	9.61	4.55
	20	1.52	11.12	13.81	12.06	6.52
	10	0.00	13.18	19.48	15.13	7.33
8	**FD**	100.00	5.47	6.70	5.63	0.19
	50	79.81	5.89	8.42	6.25	1.33
	40	51.27	6.76	10.80	7.78	4.39
	30	8.15	9.35	11.84	9.74	5.87
	20	0.62	11.40	15.31	13.85	7.04
	10	0.00	14.03	20.21	18.62	8.02
16	**FD**	100.00	5.68	21.30	11.49	2.83
	50	86.01	6.42	19.10	10.26	3.31
	40	49.47	6.94	18.35	9.00	5.86
	30	14.12	9.79	16.87	10.83	7.67
	20	0.48	12.20	21.23	15.22	8.63
	10	0.01	14.92	31.98	21.45	9.78
32	**FD**	100.00	8.43	155.77	51.60	11.57
	50	88.90	9.20	125.86	43.08	10.60
	40	54.09	8.47	116.79	13.63	7.86
	30	14.63	10.18	83.97	12.78	8.77
	20	1.00	14.15	89.65	36.40	11.34
	10	0.02	16.15	95.62	43.89	12.53
64	**FD**	100.00	12.95	1330.99	146.97	20.51
	50	89.06	12.37	1131.00	124.18	16.12
	40	65.44	10.72	900.19	101.65	14.45
	30	15.96	10.85	841.13	93.49	12.19
	20	0.77	15.35	1001.56	112.95	17.28
	10	0.04	17.40	1026.75	132.58	18.95

Bibliography

[1] 3GPP. Technical Specification Group Radio Access Network; Study on 3D channel model for LTE. Technical Report (TR) 36.873, 3rd Generation Partnership Project (3GPP), 12 2017. Version 12.7.0.

[2] 3GPP. NG-RAN; Architecture description. Technical Specification (TS) 38.401, 3rd Generation Partnership Project (3GPP), 12 2021. Version 16.8.0.

[3] 3GPP TR 21.915. Release 15 Description; Summary of Rel-15 Work Items, Oct 2019. Version 15.0.0.

[4] 3GPP TR 21.916. Release 16 Description; Summary of Rel-16 Work Items, Jan 2022. Version 16.1.0.

[5] 3GPP TR 38.901. Study on channel model for frequencies from 0.5 to 100 GHz, Jan 2020. Version 16.1.0.

[6] 3GPP TR 38.912. Study on New Radio (NR) access technology, Jul 2020. Version 16.0.0.

[7] 3GPP TR 38.913. Study on scenarios and requirements for next generation access technologies, Jul 2020. Version 16.0.0.

[8] 3GPP TS 22.186. Service requirements for enhanced V2X scenarios, June 2019. Version 16.2.0.

[9] 3GPP TS 23.501. System architecture for the 5G System (5GS), Dec 2021. Version 16.11.0.

[10] 3GPP TS 36.300. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2, Dec 2021. Version 16.7.0.

[11] 3GPP TS 37.340. NR; Multi-connectivity; Overall description; Stage-2, Dec 2021. Version 16.8.0.

[12] 3GPP TS 38.101-1. NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone, Apr 2022. Version 16.11.0.

[13] 3GPP TS 38.214. NR; Physical layer procedures for data, Apr 2022. Version 16.9.0.

[14] 3GPP TS 38.300. NR; NR and NG-RAN Overall description; Stage-2, Dec 2021. Version 16.8.0.

[15] 3GPP TS 38.323. NR; Packet Data Convergence Protocol (PDCP) specification, Dec 2021. Version 16.6.0.

[16] Adnan Aijaz. Packet duplication in dual connectivity enabled 5G wireless networks: Overview and challenges. *IEEE Communications Standards Magazine*, 3(3):20–28, 2019.

[17] Mahmoud A Albreem, Alaa H Alhabbash, Shahriar Shahabuddin, and Markku Juntti. Deep learning for massive MIMO uplink detectors. *IEEE Communications Surveys & Tutorials*, 24(1):741–766, 2021.

[18] Rabe Arshad, Hesham ElSawy, Sameh Sorour, Tareq Y Al-Naffouri, and Mohamed-Slim Alouini. Handover management in 5G and beyond: A topology aware skipping approach. *IEEE Access*, 4:9073–9081, 2016.

[19] Nicola Baldo, Marco Miozzo, Manuel Requena-Esteso, and Jaume 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*, pages 293–298, 2011.

[20] Gilberto Berardinelli, Nurul Huda Mahmood, Renato Abreu, Thomas Jacobsen, Klaus Pedersen, István Z Kovács, and Preben Mogensen. Reliability analysis of uplink grant-free transmission over shared resources. *IEEE Access*, 6:23602–23611, 2018.

[21] Gabrial Brown. Service-Based Architecture for 5G Core Networks. *A Heavy Reading white paper produced for Huawei Technologies Co. Ltd. Online verfügbar unter: https://www. huawei. com/en/press-events/news/2017/11/HeavyReading-WhitePaper-5G-Core-Network, letzter Zugriff am*, 1:2018, 2017.

[22] Marco Centenaro, Daniela Laselva, Jens Steiner, Klaus Pedersen, and Preben Mogensen. System-level study of data duplication enhancements for 5G downlink URLLC. *IEEE Access*, 8:565–578, 2019.

[23] Patrik Cerwall, Anette Lundvall, Peter Jonsson, Stephen Carson, Richard Möller, Rachit Saksena, Dave Lu, Ilyas Celik, et al. Ericsson Mobility Report. Technical report, Ericsson, November 2021.

[24] R Fantini, W Zirwas, L Thiele, D Aziz, P Baracca, A Osseiran, J Monserrat, and P Marsch. Coordinated multi-point transmission in 5G. In *5G Mobile and Wireless Communications Technology*, pages 248–276. Cambridge University Press Cambridge, 2016.

[25] GSMA. 5G Implementation Guidelines v2.0. `https://www.gsma.com/futurenetworks/wp-content/uploads/2019/03/5G-Implementation-Guideline-v2.0-July-2019.pdf`, Jul 2019. Accessed: 2022-05-22.

[26] Ralf Irmer, Heinz Droste, Patrick Marsch, Michael Grieger, Gerhard Fettweis, Stefan Brueck, Hans-Peter Mayer, Lars Thiele, and Volker Jungnickel. Coordinated multipoint: Concepts, performance, and field trial results. *IEEE Communications Magazine*, 49(2):102–111, 2011.

[27] Akshay Jain, Elena Lopez-Aguilera, and Ilker Demirkol. Improved handover signaling for 5G networks. In *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 164–170. IEEE, 2018.

[28] Mahmoud Kamel, Walaa Hamouda, and Amr Youssef. Ultra-Dense Networks: A Survey. *IEEE Communications Surveys & Tutorials*, 18(4):2522–2545, 2016.

[29] Snigdhaswin Kar, Prabodh Mishra, and Kuang-Ching Wang. 5G-IoT Architecture for Next Generation Smart Systems. In *2021 IEEE 4th 5G World Forum (5GWF)*, pages 241–246. IEEE, 2021.

[30] Taehoon Kim and Bang Chul Jung. Performance analysis of grant-free multiple access for supporting sporadic traffic in massive IoT networks. *IEEE Access*, 7:166648–166656, 2019.

[31] Lawrence M Leemis. *Reliability: probabilistic models and statistical methods*. Prentice-Hall, Inc., 1995.

[32] Zexian Li, Mikko A Uusitalo, Hamidreza Shariatmadari, and Bikramjit Singh. 5G URLLC: Design challenges and system concepts. In *2018 15th international symposium on wireless communication systems (ISWCS)*, pages 1–6. IEEE, 2018.

[33] Yan Liu, Yansha Deng, Maged Elkashlan, Arumugam Nallanathan, and George K Karagiannidis. Analyzing grant-free access for URLLC service. *IEEE Journal on Selected Areas in Communications*, 39(3):741–755, 2020.

[34] M Carmen Lucas-Estañ, Javier Gozalvez, and Miguel Sepulcre. On the capacity of 5G NR grant-free scheduling with shared radio resources to support ultra-reliable and low-latency communications. *Sensors*, 19(16):3575, 2019.

[35] ITU-R Recommendation M.2083-0. IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond. `https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf`, Sep 2015. Accessed: 2022-05-22.

[36] ITU-R Report M.2320-0. Future technology trends of terrestrial IMT systems. `https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2320-2014-PDF-E.pdf`, Nov 2014. Accessed: 2022-05-22.

[37] Nurul Huda Mahmood, Melisa Lopez, Daniela Laselva, Klaus Pedersen, and Gilberto Berardinelli. Reliability oriented dual connectivity for URLLC services in 5G New Radio. In *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6. IEEE, 2018.

[38] Mediatek. 5G NR Uplink Enhancements. `https://newsletter.mediatek.com/hubfs/mwc/download/ul-enhancements.pdf`, 2018. Accessed: 2022-12-22.

[39] Prabodh Mishra, Snigdhaswin Kar, Vikas Bollapragada, and Kuang-Ching Wang. Multi-Connectivity using NR-DC for High Throughput and Ultra-reliable Low Latency Communication in 5G Networks. In *2021 IEEE 4th 5G World Forum (5GWF)*, pages 36–40. IEEE, 2021.

[40] Prabodh Mishra, Snigdhaswin Kar, Chun-Chih Lin, Kuang-Ching Wang, and Linke Guo. Enabling Robust Communication among Military Ground Vehicles using Multi-connectivity. In *2023 SAE WCX*, pages 1–10. SAE, 2023.

[41] Prabodh Mishra, Snigdhaswin Kar, and Kuang-Ching Wang. Performance Evaluation of 5G Multi-Connectivity with Packet Duplication for Reliable Low Latency Communication in Mobility Scenarios. In *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pages 1–6. IEEE, 2022.

[42] Zheng Naizheng, Malek Boussif, Claudio Rosa, Istvan Z Kovacs, Klaus I Pedersen, Jeroen Wigard, and Preben E Mogensen. Uplink coordinated multi-point for LTE-A in the form of macro-scopic combining. In *2010 IEEE 71st Vehicular Technology Conference*, pages 1–5. IEEE, 2010.

[43] Giovanni Nardini, Dario Sabella, Giovanni Stea, Purvi Thakkar, and Antonio Virdis. Simu5G–An OMNeT++ Library for End-to-End Performance Evaluation of 5G Networks. *IEEE Access*, 8:181176–181191, 2020.

[44] SB Nokia. UL grant-free transmission for URLLC. *3GPP DRAFT, Feb*, 2017.

[45] Mohamed W Nomeir, Yasser Gadallah, and Karim G Seddik. Uplink Scheduling for Mixed Grant-Based eMBB and Grant-Free URLLC Traffic in 5G Networks. In *2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 187–192. IEEE, 2021.

[46] David Ohmann, Ahmad Awada, Ingo Viering, Meryem Simsek, and Gerhard P Fettweis. Impact of mobility on the reliability performance of 5G multi-connectivity architectures. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2017.

[47] OMNeT++: Discrete Event Simulator. `https://omnetpp.org/`. Accessed: 2021-07-15.

[48] Hyun-Seo Park, Yuro Lee, Tae-Joong Kim, Byung-Chul Kim, and Jae-Yong Lee. Handover mechanism in NR for ultra-reliable low-latency communications. *IEEE Network*, 32(2):41–47, 2018.

[49] Natale Patriciello, Sandra Lagen, Biljana Bojovic, and Lorenza Giupponi. An E2E simulator for 5G NR networks. *Simulation Modelling Practice and Theory*, 96:101933, 2019.

[50] Jaya Rao and Sophie Vrzic. Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis. *IEEE Network*, 32(2):32–40, 2018.

[51] Vishal Sharma, Ilsun You, Fang-Yie Leu, and Mohammed Atiquzzaman. Secure and efficient protocol for fast handover in 5G mobile Xhaul networks. *Journal of Network and Computer Applications*, 102:38–57, 2018.

[52] Xuemin Shen, Jie Gao, Wen Wu, Kangjia Lyu, Mushu Li, Weihua Zhuang, Xu Li, and Jaya Rao. AI-assisted network-slicing based next-generation wireless networks. *IEEE Open Journal of Vehicular Technology*, 1:45–66, 2020.

[53] Marie-Theres Suer, Christoph Thein, Hugues Tchouankem, and Lars Wolf. Multi-connectivity as an enabler for reliable low latency communications—An overview. *IEEE Communications Surveys & Tutorials*, 22(1):156–169, 2019.

[54] Marie-Theres Suer, Christoph Thein, Hugues Tchouankem, and Lars Wolf. Comparison of Multi-Connectivity Schemes on Different Layers for Reliable Low Latency Communication. In *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1357–1362. IEEE, 2021.

[55] Fasil B Tesema, Ahmad Awada, Ingo Viering, Meryem Simsek, and Gerhard P Fettweis. Mobility modeling and performance evaluation of multi-connectivity in 5G intra-frequency networks. In *2015 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2015.

[56] Lorenzo Vangelista and Marco Centenaro. Performance evaluation of harq schemes for the internet of things. *Computers*, 7(4):48, 2018.

[57] Anna Maria Vegni, Gabriele Tamea, Tiziano Inzerilli, and Roberto Cusani. A combined vertical handover decision metric for QoS enhancement in next generation networks. In *2009 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, pages 233–238. IEEE, 2009.

[58] Xianzhong Xie, Xiaoyu Ou, Huabing Lu, and Qian Huang. Joint Uplink and Downlink Resource Allocation in NOMA for End-to-End URLLC Services. *IEEE Communications Letters*, 25(12):3942–3946, 2021.