Co-funded by the
Erasmus+ Programme
of the European Union

Erasmus Mundus Joint Master's Degree
"SMART Telecom and Sensing NETworks" (SMARTNET) (2019/2021 intake)
Aston University, Triangle, B4 7ET / Birmingham, UK
Email: aipt_smartnet@aston.ac.uk / Web-site: smartnet.astonphotonics.uk/

Acknowledgement

This Master Thesis has been accomplished in the framework of the European Funded Project: **SMART Telecom and Sensing Networks (SMARTNET)** - Erasmus+ Programme Key Action 1: Erasmus Mundus Joint Master Degrees – Ref. Number 2017 – 2734/001 – 001, Project number - 586686-EPP-1-2017-1-UK-EPPKA1-JMD-MOB, coordinated by **Aston University,** and with the participation of **Télécom SudParis, member of IP Paris** and **National and Kapodistrian University of Athens.**

Aston University
BIRMINGHAM UK

TELECOM SudParis
TELECOM ParisTech

National and Kapodistrian
UNIVERSITY OF ATHENS

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

MSc THESIS

Uplink Multi-user MIMO Scheduling in Next-gen Communication Systems

Areeb T. Tariq

Supervisor (or supervisors): **Dimitris Syvridis,** Professor
Pavan Koteshwar Srinath, Research Engineer
Nokia, Nozay, France

ATHENS

SEPTEMBER 2022

MSc THESIS

Uplink Multi-user MIMO Scheduling in Next-Gen Communication Systems

Areeb T .Tariq

S.N.: 7115192100008

SUPERVISOR: **Dimitris Syvridis,** Professor

Pavan Koteshwar Srinath, Research Engineer, Nokia, Nozay, France

ABSTRACT

With the sparse radio spectrum shared by extremely demanding advanced applications, notably Extended Reality (XR), optimization for resource allocation becomes considerably significant. To meet these ever-increasing requirements, Multi-User Multiple Input Multiple Output (MU-MIMO) has proven to be a key enabling technology for next-generation communication systems. MU-MIMO offers transmission reliability and throughput gains through spatial diversity and spatial multiplexing. Besides downlink transmission, several XR use cases consistently seek freshness in transmitted information in an uplink direction. This work introduces a metric called “Age of Transmission (AoT)” with tunable hyperparameters to quantify the freshness of information. A novel uplink MU-MIMO scheduling algorithm called “Weighted Proportional Fair (WPF)” has been proposed in this work to satisfy the AoT requirements of the users while maximizing the average user throughput. In addition to average AoT and throughput, scheduling fairness for users with poor channels is also included in KPIs to evaluate performance. The simulation results for MU-MIMO configuration demonstrate the improved overall QoS performance of the proposed approach against the standard uplink scheduling baseline algorithms.

SUBJECT AREA: Communication Systems

KEYWORDS: MU-MIMO communication, optimal scheduling, age of transmission, resource allocation, uplink

AKNOWLEDGMENTS

For the completion of this thesis, I would like to thank Dr. Alvaro Valcarce (Research Engineer at Nokia Bell Labs, Nozay, France) and Mr. Ravi Sharan (Intern at Nokia Bell Labs, Nozay, France) for their cooperation and valuable contribution to the completion of this work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This thesis is an original work by Areeb Tariq. No part of this thesis report has been published before. This thesis work also contributes to Nokia Bell Labs, Nozay, France, research for wireless systems optimization. This report has been internally reviewed and approved by Nokia Bell Labs, Nozay, France.

1. INTRODUCTION

Each year, the number of internet users and mobile devices using 5G and beyond technologies is increasing rapidly [1], as shown in the figure 1. This exponential growth of devices is not only limited to human users using cellular technologies. The spread of diverse IoT applications and extensive sensor networks has also tremendously contributed to this increased number of connected devices. Moreover, modern applications have put additional requirements on data and Quality of Service (QoS). To summarize, the immense increment in the number of user devices constituting diverse user applications has massively enhanced existing, and future data and QoS demands. These requirements call for implementing advanced wireless technologies and efficiently utilizing the available spectrum.

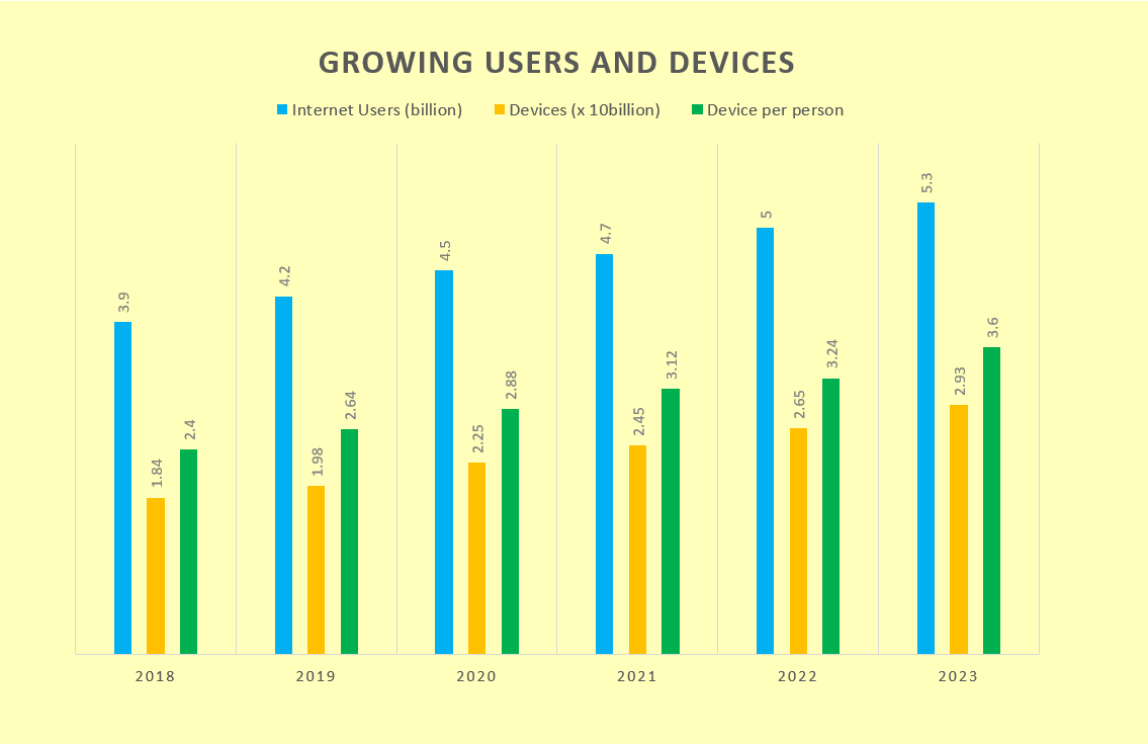

Figure 1: Growing Trend for the Mobile Users and Devices [1]

Current mobile operators have targeted cellular connectivity beyond mobile broadband with the rising IoT cellular applications [2]. IoT devices serve various applications, not limiting themselves to a particular key performance indicator (KPI). For example, some IoT applications require high data rates. Typical examples are unmanned aerial vehicles and drones. Similarly, some applications require lower latency and higher reliability in communication. The most common application for such use-cases is smart healthcare and modern intelligent transport networks involving machine-to-machine (M2M) communication. And numerous applications demand a part of both KPIs. Therefore, mobile networks need to be adaptive and intelligent to satisfy tradeoffs between higher data rates, lower delays, high reliability, and fairness.

Unlike typical mobile applications that emphasize higher data rates, with the presence of time-critical applications in different domains [3], the QoS requirements involve lower latencies. Therefore, in addition to data rate guarantees, for modern applications, QoS must also offer guarantees on the latencies. The general case of time-critical applications is real-time applications with an incredibly lower tolerance to delays and data corruption.

These real-time applications are spread over various sectors, from remote healthcare to real-time games and entertainment.

Mobile cloud gaming is getting quite popular with ever-increasing demands on data rates and, more importantly, on latencies. It offers an opportunity both to the consumers of cloud gaming and the service providers. The increasing market provides a window to service providers to meet the demands of consumers and grow their businesses. On the other hand, consumers with better service quality will enjoy greater Quality of Experience (QoE) [4].

In addition to mobile cloud gaming, another widespread real-time lower latency and higher data rate applications involve extended reality (XR). Extended reality and its overlap with augmented reality (AR) and virtual reality (VR) have caught the attention of a massive number of customers all over the world. However, XR and cloud gaming traffic isn't restricted to the downlink only. These applications require video transmission and streaming in uplink [5], which presents a challenge to satisfy QoS for these devices in the limited and shared spectrum. These QoS requirements are mapped to 5QI by 3GPP [6], as shown below in the Table 1.

Table 1: XR QoS mapping to 3GPP 5QI [6]

Service Name	5QI Value	Resource Type	Packet Delay Budget (PDB)	Packet Error Rate
Conversational Video	2	GBR	150 ms	10-3
Real-time gaming	3	GBR	50 ms	10-3
Video Interactive gaming	70	Non-GBR	100 ms	10-3
AR	80	Non-GBR	10 ms	10-6

5G and beyond must adapt and continuously evolve to meet these dynamic requirements [7]. Modern wireless cellular networks, including 5G, realize the needs of different technologies, including M2M communications [8], Internet of Things [9, 10], smart homes [11], low-latency networks [12], and various modern wireless networks [13, 14]. 5G has defined use-cases for different scenarios, as shown in the figure 2 below.

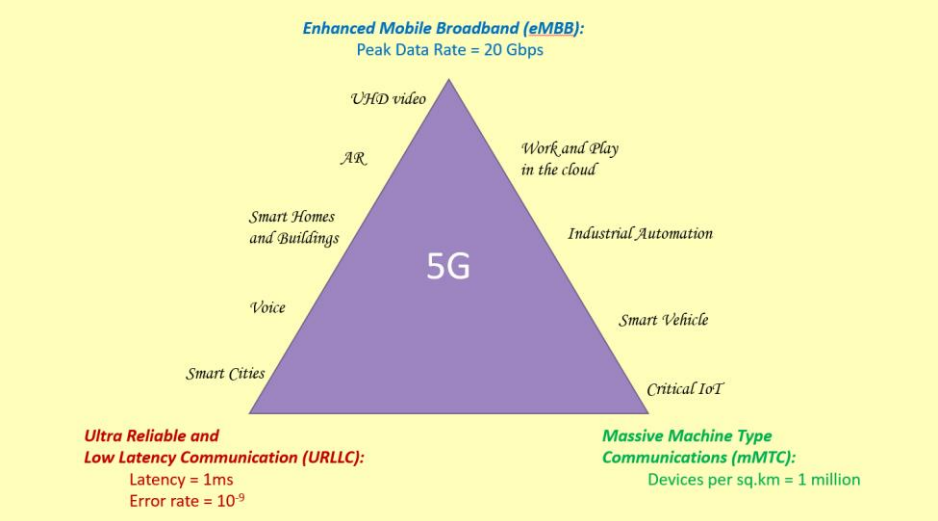

Figure 2: 3GPP 5G Use-Cases

5G and beyond envision achieving their targets by implementing key enabling technologies. Multiple input and multiple output (MIMO) has been a vital enabling technology for 5G and beyond [15]. MIMO technology enhances throughput and transmission reliability for the user [16]. MIMO employs spatial diversity and spatial multiplexing to optimize performance for wireless communication. In MIMO configuration, the base station and user equipment (UE) have more than one antenna [17]. The base station typically contains 10s or 100s of antennas to serve 10s of users. On the other hand, the user possesses antennas in the range of 2-16. MIMO has two main configurations, i.e., Single-user MIMO (SU-MIMO) and Multi-user MIMO (MU-MIMO).

SU-MIMO increases the reliability and throughput for the selected user, while MU-MIMO impacts the performance of the whole network more broadly by targeting multiple users. With the limited available spectrum and an exponential increase in the number of users and demands for radio resources, an intelligent spectrum-sharing strategy is the core of the research. Therefore, MU-MIMO is a critical focus in research to optimize the network performance and QoE for the users. Therefore, this thesis also focuses on the area of MU-MIMO specifically.

In modern wireless technologies, including cellular and WiFi, Orthogonal Frequency Division Multiplexing (OFDM) has been used to allocate time and frequency resources to different users [18]. This allocation process is described as the base station's scheduling decision. User scheduling identifies which user to utilize partial or complete bandwidth at a given time. However, with the intersection of OFDM with MU-MIMO, scheduling decisions must select the best possible combinations of users sharing the time and frequency resources. This signifies the importance of this decision since different users operate simultaneously, and the frequency of resources can negatively impact each other's performance. This will lead to system-level degradation.

User scheduling in MIMO systems, therefore, is crucial and arguably the most critical factor for network performance [19]. It must meet the specific needs of the particular user while at the same time ensuring the efficient and optimum utilization of the resources. Different approaches and criteria have been explored for user selection in the downlink transmission for MU-MIMO. These methods offer different tradeoffs between complexity and performance.

As discussed above, with the inception of real-time cloud gaming and XR applications, the uplink traffic has become as crucial as the downlink. Furthermore, with current phones supporting MIMO and requiring higher uplink QoS demands, the network must respond to such requirements accordingly. Moreover, with the ever-increase in device and application penetration, the demand for uplink MU-MIMO traffic has challenged the current network solutions. Therefore, it's equally significant for 5G and beyond networks to employ advanced uplink MU-MIMO user scheduling algorithms. Furthermore, it's vital to mention that these requirements include higher data rates and timely scheduling. Therefore, identifying the correct QoS parameters and implementing an effective user scheduling algorithm lead to better network performance and improved QoE for the respective users.

This work introduces a metric named age of transmission (AoT) in QoS. AoT quantifies the packet delay budget, inter-service delivery time, and user scheduling fairness. The XR use-case has been selected to define the data packet model, which will be explained in the next chapters. A novel uplink scheduling algorithm for MU-MIMO named weighted proportional fair has been proposed using the newly introduced metric and data model. The performance of this algorithm has been compared against the typical round-robin and proportional fair user scheduling algorithms.

This document is organized as follows. In chapter 2, the theoretical background for the MIMO and previous work on this topic have been mentioned in detail. Chapter 3 discusses the XR traffic use-case and data packet management related to the problem under consideration. Chapter 4 describes the AoT metric, system model, and problem formulation. The uplink MU-MIMO scheduling baselines and proposed algorithm are explained in chapter 5. Chapter 6 includes the simulation configurations and results. Finally, this work is concluded in Chapter 7. The technical terms and abbreviations used in this document are summarized in Table Abbreviations - Acronyms.

2. PRELIMINARIES

This chapter provides the theoretical background and setup for concepts and technologies relevant to the problem described in chapter 4.

2.1 MIMO

With the introduction of IoT and modern technologies like extended reality (XR), wireless traffic has increased exponentially. The bandwidth can be increased to a certain degree to meet these requirements. However, increasing bandwidth to meet data requirements isn't a sustainable solution to this problem. Therefore, different research areas are explored in 5G and beyond to satisfy user experience. Among numerous critical enabling technologies for 4G, 5G, and beyond, one of the most explored domains is Multiple-Input Multiple-Output (MIMO) [15].

MIMO is an antenna technology that contains several antennas in the range of tens or hundreds at the base station (BS) serving a selected group of UEs [20]. MIMO configuration provides numerous advantages for the overall system while requiring design and hardware updates [21]. Firstly, this large array of grouped antennas enhances spectral efficiency and higher throughput for the UEs. As the number of antennas in the MIMO increases, the radiated beam becomes narrower and more directive towards a specific group of users, consequently increasing the spectral efficiency (bits/s/Hz) [22]. Secondly, besides the spectral efficiency, MIMO also offers energy efficiency since the radiated beam is directed to a fixed direction rather than the spread in an omnidirectional manner in the azimuthal plane. Thirdly, the array gain of MIMO significantly increases the data rate for the user group through spatial multiplexing. Fourthly, MIMO also offers spatial diversity to improve the reliability of data transmission. Spatial diversity allows identical copies of data streams to be transmitted or received on multiple transmitting or receiving antennas, respectively. Furthermore, depending on the configuration, transmitter and receiver diversity compensates for multipath effects in the wireless channel. Last but not least, this directive beam in MIMO reduces the interference with the other non-targeted users.

As explained in various advantages of MIMO configuration, the number of correct bits in the same block of bandwidth (spectral efficiency, in other words) is remarkably increased. More importantly, this gain comes without increasing the transmit power or bandwidth. The most common positives with this configuration in signal processing are spatial diversity and spatial multiplexing.

2.1.1 Spatial Diversity and MIMO

In mobile communications, diversity is employed to counter fading effects. Diversity increases reliability in communication at the cost of utilization of additional resources. There are three types of diversity: time, frequency, and space diversity. In time diversity, identical signal copies are sent at different times to reduce communication errors. Similarly, in frequency diversity, exact signal copies are transmitted on various frequency resources simultaneously. Lastly, in space diversity (also called spatial diversity), multiple antennas send identical signal copies on the same time and frequency resources.

MIMO systems utilize spatial diversity to handle this multi-path effect and ISI. Spatial diversity can be employed at the transmitter, receiver, or both ends. MIMO uses both transmitter and receiver diversity with multiple antennas at both ends. For instance, in downlink communication, different antennas at the base station located in separate physical spaces transmit the same signal to the receiver using the same time and frequency resources. While the antennas at the UE, located in different spaces, receive different copies of the same signal coming from multiple paths. These redundant copies make it less likely for all the directions in the same time slot and frequency band to be degraded, resulting in enhanced transmission reliability.

The receiver can use different techniques to detect the original signal. The most common receiver diversity techniques are maximum ratio combining (MRC) and selection combining.

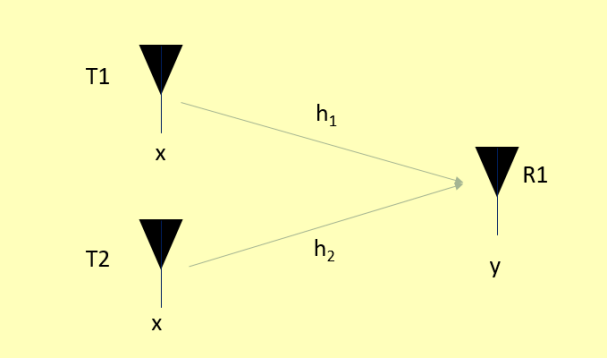

Figure 3: MIMO Transmit Diversity

$$y = x\mathrm{h}_1 + x\mathrm{h}_2$$

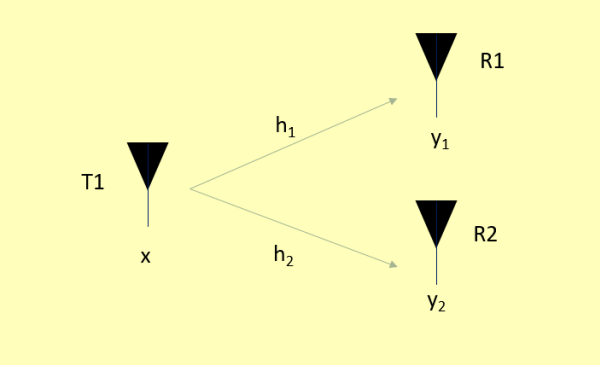

Figure 4: MIMO Receiver Diversity

$$y_1 = x\mathrm{h}_1$$

$$y_2 = x\mathrm{h}_2$$

2.1.2 Spatial Multiplexing and MIMO

As mentioned before, besides reliability, MIMO configuration also offers increased throughput for the user via spatial multiplexing. Spatial multiplexing incorporates several antennas separated in the space to transmit multiple data streams in parallel. Each antenna transmits different data content using the shared time and frequency resources, significantly increasing users' data rates. The target for this communication can be a single user or a group of users. Upon receiving parallel data streams, the receiver combines them and processes them with the received data.

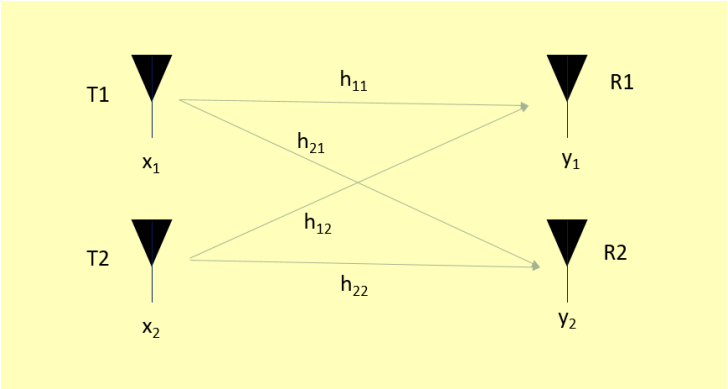

Figure 5: MIMO Spatial Multiplexing

$$\boldsymbol{Y = HX}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Spatial multiplexing converts the incoming higher data rate signal at the transmitter into multiple data streams to be transmitted on common time and frequency resources. In other words, the same channel bandwidth is reused by all possible transmission data streams. Because of the spatial multiplexing, the throughput for the targeted users is increased, and the interference with unintended users is decreased.

The minimum number of antennas at the transmitter and the receiver limit the maximum number of data streams to be transmitted. This work focuses on the spatial multiplexing implementation of the MU-MIMO systems.

2.1.3 Types of MIMO

There are following possible MIMO configurations:

2.1.3.1 Single-user MIMO (SU-MIMO)

In SU-MIMO, at least one end point of the transmission link (transmitter or receiver) is singular. Multiple data streams are transmitted on shared radio resources towards the single user, resulting in the performance improvement of the particular user. In this case, the channel information between the transmitter and receiver is not required.

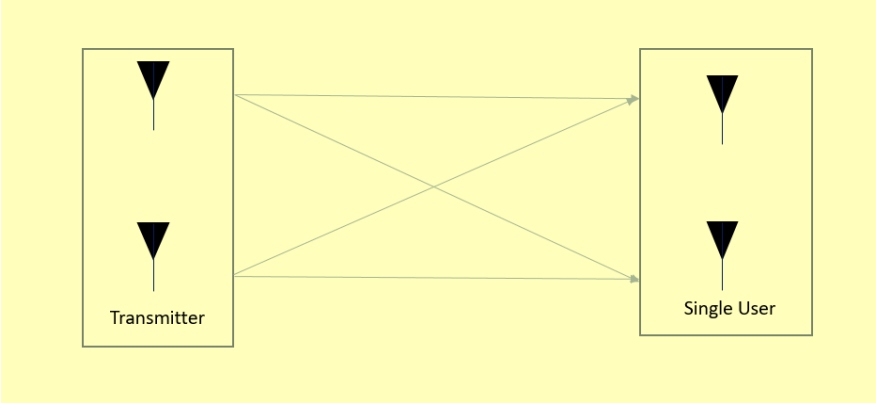

Figure 6: SU-MIMO Visualization

2.1.3.2 Multi-user MIMO (MU-MIMO)

MU-MIMO allows multiple users to simultaneously use the same time and frequency resources with varied spatial multiplexing [23]. However, it is essential to know the channel state information (CSI) between the transmitter and receiver to reduce interference between overlapping users. With correct channel information, precoding, and coding techniques, MU-MIMO can increase the overall system capacity and throughput.

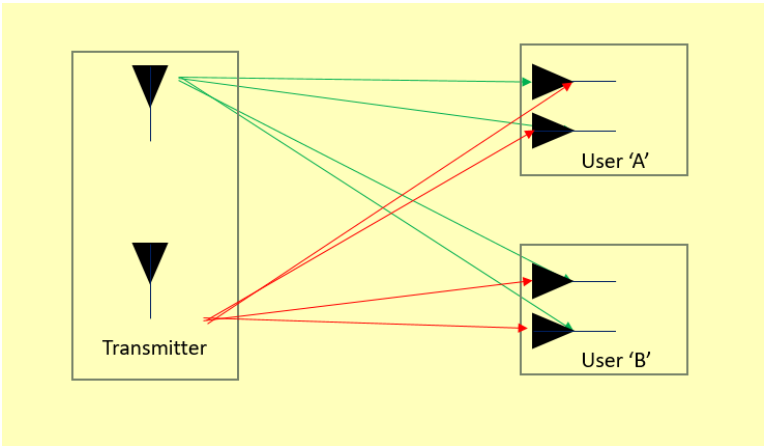

Figure 7: MU-MIMO Visualization

As shown in the figure 7 above, the MU-MIMO system has one transmitter sending signals simultaneously to multiple users. This optimizes the performance in terms of improved throughput, diversity, and multiplexing gains.

2.1.4 MIMO Transmission Modes

In wireless communication employing MIMO, there are following two transmission modes:

2.1.4.1 MIMO Downlink Transmission

As this work focuses on MU-MIMO, the MU-MIMO downlink transmission involves signaling and user data communication from the base station and multiple users in its

coverage range. A typical downlink MU-MIMO has a base station equipped with ND antennas transmitting signals to K UEs, each with NU antennas.

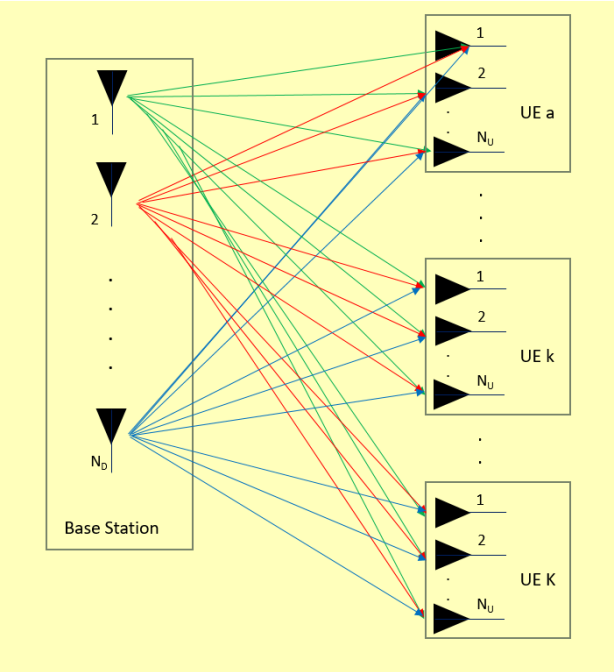

Figure 8: MU-MIMO Downlink Visualization

The signal received at the UE 'k' is following with the dimension of N_U x 1:

$$\boldsymbol{Y_k} = \boldsymbol{X_k}\,\boldsymbol{H_k}$$

The '$\mathbf{X_k}$' is the signal transmitted by the base station with N_D x1 dimension. $\mathbf{H_k}$ is the channel matrix between base station and user 'k' with dimension of N_D x N_U. Each channel element is independent and identically distributed.

2.1.4.2 MIMO Uplink Transmission

The MU-MIMO uplink transmission constitutes signaling and user data information from each UE to the camped base station.

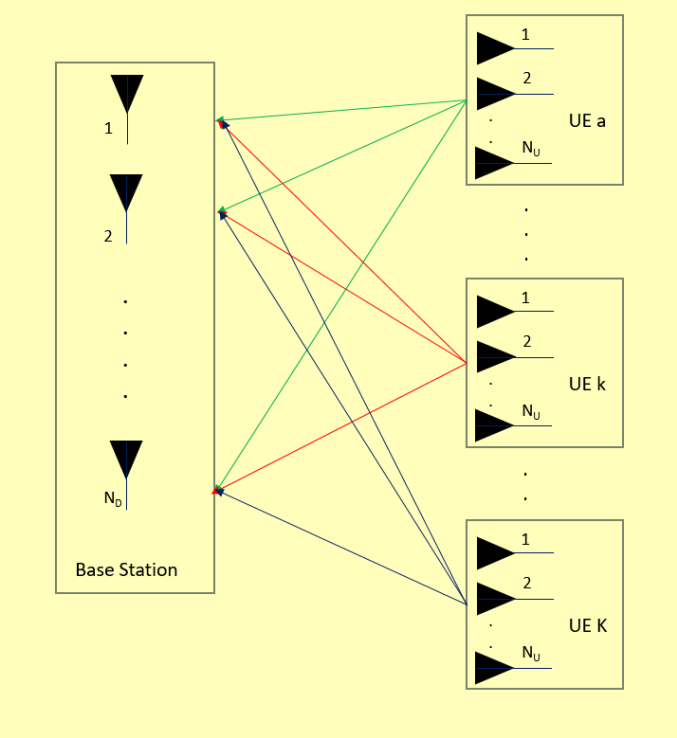

Figure 9: MU-MIMO Uplink Visualization

With the same configuration explained in the downlink section, the uplink signal received at the base station from the UE 'k' is as follows:

$$\boldsymbol{Y_k} = \boldsymbol{X_k}\,\boldsymbol{H_k}$$

The 'kth' signal received at the base station is the uplink signal from the 'kth' UE. The channel matrix $\mathbf{H_K}$ between UE k and the base station has a dimension of N_D x N_U. The elements of the channel are independent and identically distributed. The overall signal received at the base station is accumulation of the signals received from all 'K' UEs having K x ND x NU dimension and is given as:

$\mathbf{Y} = \sum_{k=1}^{K} \boldsymbol{Y_k}$

The overall uplink signal dimension is K x N_D x N_U.

2.2 Resource Allocation and OFDM

With the growing number of users and diverse applications, uplink transmission requires higher QoS and data requirements. Because of the limited resources in the network, resource allocation has become a significant research area for uplink multi-user systems. Therefore, the MU-MIMO has further emphasized resource utilization for uplink communication.

In wireless communication, radio resources include physical resources in different domains. Therefore, resource allocation comprises power allocation, spectrum allocation, time allocation, modulation and coding (MCS) allocation, precoding matrix allocation, antenna allocation, beam allocation, layer allocation, and so on. However, spectrum allocation, time allocation, layer allocation, and MCS allocation are focused on in this work.

Orthogonal Frequency Division Multiplexing (OFDM) is an underlying technology for all the current wireless services, i.e., 802.11, LTE, 5G NR, etc. OFDM is an extension of

classical Frequency Division Multiplexing (FDM). However, unlike FDM, OFDM has multiple orthogonal subcarriers rather than a single carrier. This orthogonality of subcarriers allows subcarriers to be squeezed together to increase available bandwidth. Mathematically, the two consecutive subcarriers are orthogonal if the integral of their product over time is zero. Because of this orthogonality, OFDM doesn't require guard bands between subcarriers to avoid interference. This technique also conserves and ensures efficient use of the transmission bandwidth. Each subcarrier is modulated independently in parallel with a digital modulation scheme (N-QAM, N-PSK).

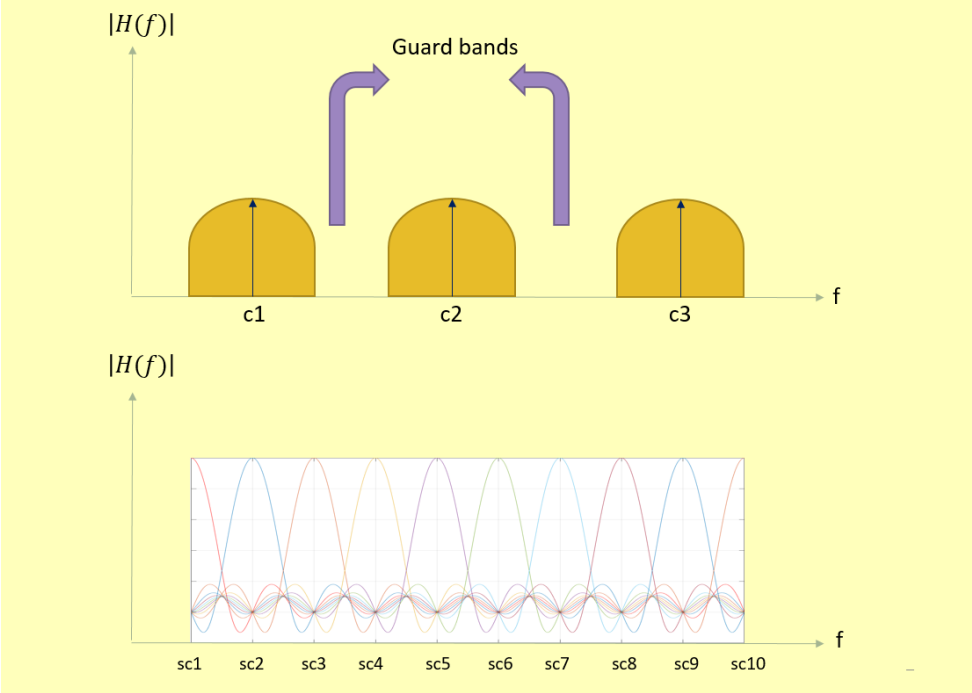

Figure 10: FDM (top) vs OFDM (bottom) Spectrum Utilization

As shown in the figure 10 above, the FDM requires a guard band between consecutive carriers, i.e., c1 and c2. However, in OFDM, the orthogonality between subcarriers, i.e., sc1 and sc2, allows multiple subcarriers to be squeezed, allowing higher bandwidth utilization and spectral efficiency. In addition to this spectrum efficiency, because of the higher number of carriers in comparison to the classical FDM, OFDM is far more resilient to frequency selective fading.

Each subcarrier independently gets modulated, all combined afterward to make an OFDM symbol. The period of the OFDM symbol is reciprocal to the subcarrier spacing; hence orthogonality is ensured.

In 5G and beyond, the OFDM resource grid typically contains time and frequency axes. Each time unit has multiple subcarriers placed over the entire channel bandwidth. This OFDM configuration allows different users to be scheduled at a given time with specific subcarriers. This decision of allocation of time and frequency resources is user scheduling. The time and frequency structures have different units, which are explained in this section.

In 5G and beyond, communication is carried out by the uplink, downlink, and special radio frames and sub-frames in the time domain. The uplink and downlink configurations determine the location of these frames. There are a total of 7 such configurations. In addition, special sub-frames are used to switch from downlink to uplink transmission.

Generally, these sub-frames contain uplink and downlink pilot time signals comprising synchronization and reference signals.

As per 3GPP specification 38.211 [24], the NR (5G) radio frame is of the size of 10ms. The half-sub-frame and sub-frame are 5ms and 1ms, respectively. However, unlike the previous generations, 5G NR offers flexible numerology for the slot duration. Based on the given numerology, slot size varies. Since the slot size varies with numerology, it results in variation in the number of slots per sub-frame. However, the number of OFDM symbols in a slot remains the same across all the numerologies. Typically per slot, there are 14 (normal cyclic prefix) or 12 (extended cyclic prefix) OFDM symbols. The Table 2 below shows the distribution of symbols, slots over different numerology.

Table 2: 5G NR Flexible Numerology

Numerology	Subcarrier Spacing (kHz)	Slot Duration (ms)	Slots in Sub-frame	Symbols in Sub-frame
0	15	1	1	14
1	30	0.5	2	28
2	60	0.25	4	56
3	120	0.125	8	112
4	240	0.0625	16	224

On a frequency domain, different units are available. The most common is called physical resource block (PRB), which contains the 12 consecutive subcarriers and 1 slot duration.

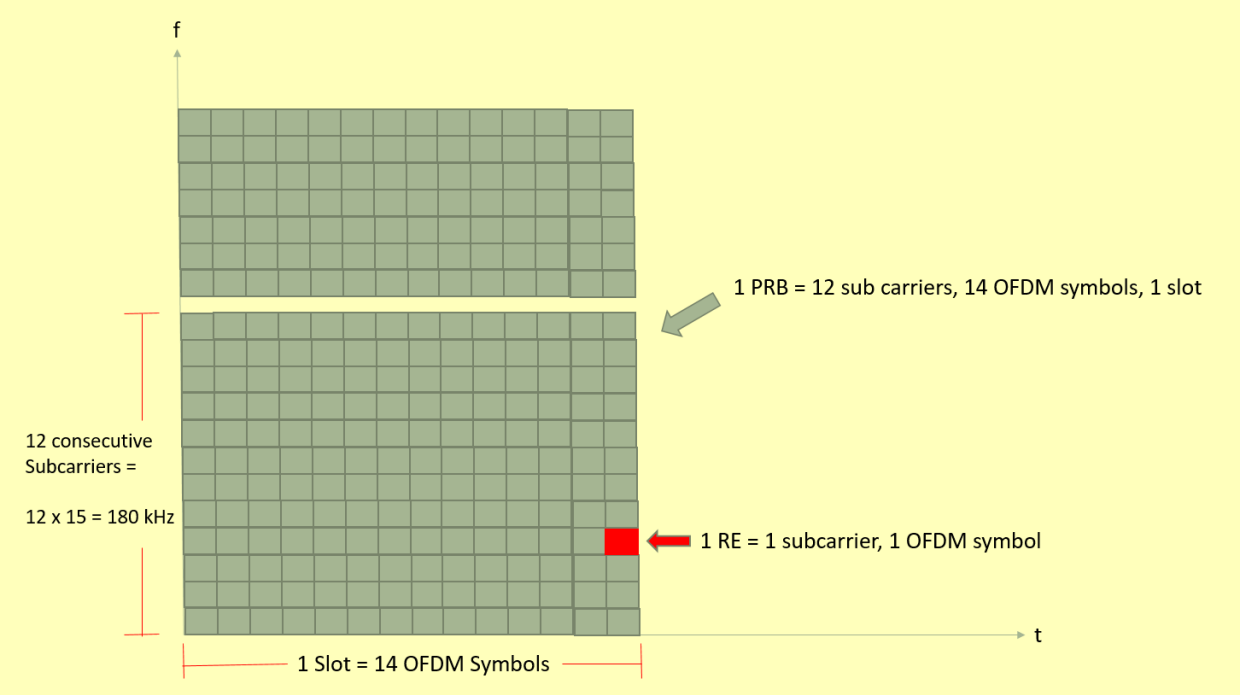

Figure 11: OFDM Resource Grid Visualization

As shown above, in the figure 11, the composition of 12 consecutive subcarriers in the frequency domain and one slot duration in the time domain, which constitutes 14 OFDM symbols, form 1 physical resource block (PRB). Resource Element (RE) is the smallest unit in the radio resource grid which comprises one subcarrier in the frequency domain and 1 OFDM symbol in the time domain. It's important to mention the figure 11 above is

for the case of numerology = 0. In the given case, 1 PRB has 180 kHz bandwidth in the frequency domain and 1 ms duration in the time domain. Moreover, 1 PRB contains 14 x12 = 168 REs, where each RE has 15 kHz bandwidth and 71.4 ms duration. In any other numerology, the number for PRB bandwidth, OFDM symbols, and slot time will vary accordingly.

With all the advantages of OFDM, the major disadvantage is a higher peak-to-average power ratio (PAPR). These variations in amplitude can make it difficult for linear amplifiers to adjust to higher amplitudes leading to lower amplifier efficiency. The second downside of OFDM is sensitivity to noise because of multiple carriers. The carrier frequency offset (CFO) negatively impacts OFDM systems far more than the single-carrier modes.

2.3 OFDM and MU-MIMO

With the prior explanation of spatial multiplexing in MU-MIMO and resource grid in OFDM, in this subsection, we will try to focus on how radio resources (time and frequency) are shared in the MIMO setup. In other words, this section focuses on the intersection of MIMO and OFDM concepts which are the integral building blocks of current and future wireless communication.

Spatial multiplexing (with the availability of multiple antennas) allows various data streams to be transmitted to the receiver. Radio resources (time and frequency) are allocated for each data stream. However, the resources are shared on each data stream to increase the throughput using spatial multiplexing. Intuitively is easy to understand how this impacts the system's performance in terms of throughput. However, the con side of this configuration is the co-channel interference between data streams since they operate on the same resources. This results in a higher likelihood of interference and, subsequently, higher errors and lower throughput. In such cases, the whole idea of spatial multiplexing is underachieved or unachieved in the worst case. Therefore, channel information is critical when transmitting multiple data streams on the same radio resources. If the channel conditions are suitable, spatial multiplexing will produce positive results. In the opposite case, the system performance can take a severe hit. As one can imagine, the higher the data streams or layers, the higher the likelihood of interference. In other words, channel conditions limit the number of layers for optimum performance.

Starting from the simple case of the single-user MIMO (SU-MIMO), employing spatial multiplexing will enable multiple parallel data streams to be transmitted to the single user using the same time and frequency resources. This logically must increase the user's throughput, given the low co-channel interference ideally, as we increase the number of layers throughput for the given user multiplied with the same factor.

The number of layers can be increased to a point where it doesn't exceed transmitting and receiving antennas. Typically in downlink communication, the number of layers is limited by the UE side since UE has fewer receiving antennas. Similarly, in the uplink, the sum of layers from different UEs must be less than or equal to the layers supported by the base station.

In MU-MIMO, the main subject of this work, multiple users are targeted using spatial multiplexing and shared radio resources, both in uplink and downlink transmissions. In this case, on the same time instant and frequency block, different users can be operated. Unlike SU-MIMO, which employed spatial multiplexing to increase throughput for the given user, MU-MIMO uses spatial multiplexing to serve more users to improve the overall system performance and network capacity. However, in MU-MIMO, in addition to interference between different layers, the interference between other users sharing the

same time and frequency response also limits the performance. Again, the better channel condition will allow the luxury to serve more users on the same resources.

The availability of channel information becomes a crucial determining factor for network performance. The accurate channel information will allow the base station to identify the users with lower interference, which can be served simultaneously on shared resources. This decision-making process is called user scheduling for both uplink and downlink transmissions. Traditionally, different approaches and algorithms have been proposed and discussed for downlink scheduling or MU-MIMO downlink scheduling. However, with diverse and demanding applications for uplink transmissions, uplink MU-MIMO scheduling demands a similar focus and research interest.

Specifically, with XR traffic which has massive uplink video traffic data and a large of simultaneously active users accessing the same network, the user scheduling for uplink MU-MIMO systems has become quite significant to satisfy respective users' uplink QoS requirements. This is the primary motivation behind this research work to explore different strategies to improve the uplink transmission experience for the users present in MU-MIMO configurations. User scheduling becomes quite significant when multiple users demand a share from the limited resource space. The decision to select how many users and which of the users to be served at the given time with a given spectrum share represents user scheduling. The better the scheduling method for such users, the better will be the radio resource management and the user satisfaction.

2.4 Related Work

User scheduling and resource allocation has always been significant for mobile network operators to meet user demands and increase revenues. MU-MIMO has just enhanced the significance of these processes. Multiple research works have been done for different components of resource allocations. However, most of the work has been targeted for the downlink MU-MIMO user scheduling.

The joint antenna selection and user scheduling (JASUS) in downlink MU-MIMO systems have recently been exceptionally researched topics. For example, the authors of [25] have proposed a Matrix Gauss Elimination Method (MGEM) for user scheduling in downlink transmission aimed at increasing the system's throughput. In [26], targeting a similar problem over time-varying fading channels, the authors tried to maximize the overall sum rate and satisfy the individual user's data rate requirements. In the research mentioned above, the simulation results increase the sum rate, but there are no definite guarantees of delays for particular users.

Machine learning and reinforcement learning (RL) have also been researched for MU-MIMO systems for user scheduling. The authors in [27] mapped the scheduling problem as Markov Decision Process (MDP) and utilized multi-agent Q-learning to perform resource allocation. The rewards in this work are again designed to maximize the sum rate, and there are no guarantees on the timely delivery of data. In addition to Q-learning, the policy gradient has also been explored to solve the same problem in the downlink direction. The authors in [28] has utilized deep deterministic policy gradient (DDPG) for SU-MIMO resource allocation. For MU-MIMO, the authors in [29] have divided a typical DDPG into offline training and real-time decision-making. The authors demonstrate 56% faster speed with the proposed DDPG incorporating asynchronous updates for actor and critic networks. In addition to the throughput, the authors have also considered fairness a reward parameter.

All the above works mentioned are targeted for the downlink MU-MIMO. In this document, it's been repeatedly emphasized how important is the uplink MU-MIMO user scheduling.

Unfortunately, there has been very little work on uplink MU-MIMO user scheduling. However, in [30], the authors have proposed joint user scheduling and transmit precoding matrix selection approach for uplink MU-MIMO using DDPG. Unlike the previous work focusing on downlink transmission, this work considers the antenna correlation at the UE side alongside the channel correlation between users. This problem has been modeled as combinatorial with user selection and respective precoding matrices. In this work, the authors again considered the system sum rate as the KPI and reward in the DDPG approach.

While the work in [30] improves the system throughput for uplink MU-MIMO systems, it is also essential to improve the other QoS requirements for the UE applications in the uplink MU-MIMO system. In addition, the RL approaches require higher data samples for both the online and offline training phase. For example, in the work carried out in [30], the number of user per base station are picked to be 4. However, for the more extensive and practical scenario, a BS has far more users in its coverage area, which induces more user interference scenarios. This eventually leads to the difficult scheduling problem with scalability since RL requires a large number of data samples for sub-optimal or optimal performance.

Therefore, this work puts significance on the KPIs other than the data rate for QoS while evaluating a scheduling algorithm. The decision not to pursue the machine learning paradigm in this work was to reduce the computational complexity and performance time. Furthermore, this work performs better with other scheduling baselines: round-robin and proportional fair. The simulations have been carried out on the full-scale uplink MU-MIMO scheduler with a reasonably bigger network size and footprint. The motivation behind this work is to contribute to the relatively less explored uplink MU-MIMO user selection domain for the XR use-cases.

3. TRAFFIC MODEL AND DATA PACKET MANAGEMENT

This chapter discusses the reference traffic model for the problem described in Chapter 4. The importance of extended reality (XR) applications and their QoS mapping is discussed first, followed by the explanation of data packet management and transmission.

3.1 Extended Reality

Immersive technologies and applications are attaining the interest of customers and small businesses. One such rapidly growing technology is Extended Reality (XR). XR has always been an abstract concept under discussion and research for a long time. However, the significant development in XR resulted from the modern smart devices and entertainment market. In addition to the widely popular entertainment and gaming industry, XR has various applications in communication, healthcare, transportation, education, and industry.

Extended Reality (XR), Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) are widely interchanged in everyday discourse. However, there lie differences per definition between these terms. VR environment offers the user a complete virtual 3D view of alienation from the real world. A typical example of VR is virtual games with headsets. In AR, the user experiences a virtual overlay in the real world. In other words, the virtual objects are immersed in the real world surrounding the user. A classic example of AR would be Pokémon Go. MR is also similar to AR and includes real objects immersed in the virtual view provided to the user. MR is everything in between the absolute real world and absolute VR. The figure 12 below summarizes the distinction between these terms.

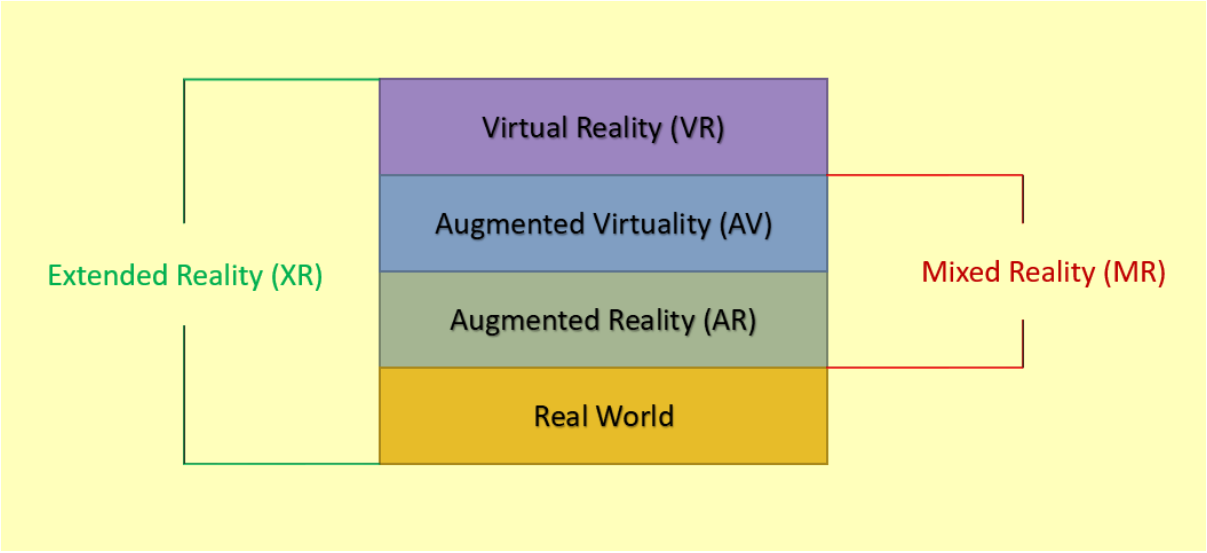

Figure 12: XR and its variants

XR traffic model equally incorporates low latency requirements as much as on higher data rates. In this work, the XR traffic model has been picked as a reference for evaluating the uplink MU-MIMO scheduling algorithm. To evaluate, the XR use-case and traffic model have to be quantified. 3GPP has introduced a 5G Quality Indicator (5QI) for services involving XR [31]. Moreover, 3GPP TR 38.838 Release 17 [32] documented the XR traffic models and respective KPIs.

Table 3: 3GPP TR 38.838 QoS parameters for XR and CG [32]

Traffic Model	Data Rate (Mbps)	PDB (ms)
VR DL	30, 45	10
VR UL	0.2	10
CG DL	30, 8	15
CG UL	0.2	10
AR DL	30, 45	10
AR UL	10	30

Significantly AR headsets demand has been continuously growing in the past five years and is expected to follow this trend more vigorously. From 2019 to 2025, the AR headsets market share is expected to grow by 73.8% [33]. AR headsets provide an opportunity to interact with individuals and enterprises in a customized virtual world. Moreover, the major global technologies companies have shown keen interest in investing in AR businesses. For example, Microsoft for HoloLens 2019 collaborated with AR service providers to enhance communication for HoloLens users. Other notable players investing in this AR are Google LLC, Sony Corporation, RealWear Inc., etc. This increasing AR popularity encourages defining problems suitable to these particular use cases and designing solutions accordingly.

These QoS parameters described above in the Table 3 help identify the correct evaluation metrics for this work. As explained earlier, the motivation behind this research is to satisfy data rate and timely scheduling of the users for the uplink MU-MIMO systems.

3.2 Data Packet Management

The Data packet is the fundamental unit in the communication system. Therefore, it's crucial to define its structure and flow for the selected problem. Different wireless systems, namely WiFi, Bluetooth, LTE, and NR, describe packets according to relevant standards. In a protocol stack, different layers have their functionalities and names for information blocks.

In 3GPP, in layer 2 for LTE and NR, Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), and Medium Access Control (MAC) have their packet sizes and methodologies for transmission. Typically, as moving downwards for uplink transmission data, packet size gets shorter. 3GPP also defines multiple timers for the retransmission lost or corrupted data packets. The packet size and retransmission method are at the core of data packet management.

Moreover, buffer status reporting (BSR) is implemented in LTE, NR, and other wireless technologies. BSR enables communicating the current status of transmittable data from the UE to the BS if any. This information is quite valuable for the BS in prioritizing UEs while scheduling.

The problem defined in this work operates on a "generate-at-will" model for the data packets. This means as soon as the current application packet is removed from the buffer, it is assumed that the new packet will arrive immediately without any slot delay. This ensures the continuous availability of data packets for transmission on the UE side. The real-time XR traffic, especially video streaming, aligns with this model, where the requirement of uplink transmission is always active. Every packet from the application layer has a defined Packet Data Budget (PDB) to meet QoS. However, on each slot, UE is allocated a particular transport block size which may or may not equal the size of the

entire packet. Generally, it takes multiple scheduling opportunities for the UE to transmit the whole application packet for three obvious reasons. First, UE won't get a scheduling opportunity for every slot because of the competition with other UEs for the same resources. Secondly, the transport block size is typically smaller than the large application packet size. Thirdly, because of connection errors or data corruption, the data packet has to be retransmitted, which will cost additional time.

UE keeps reporting its buffer status to the BS at each slot to facilitate BS making better scheduling decisions. With this communication flow, a UE may or may not successfully transmit the entire application packet in the defined PDB for the given packet. If the entire application packet is not being sent successfully during this defined PDB, the UE will discard all the remaining content of the current application packet. Upon discarding, a new application packet with its defined PDB will arrive at the UE buffer.

The data packet management for this work is well demonstrated in the figure 13 below.

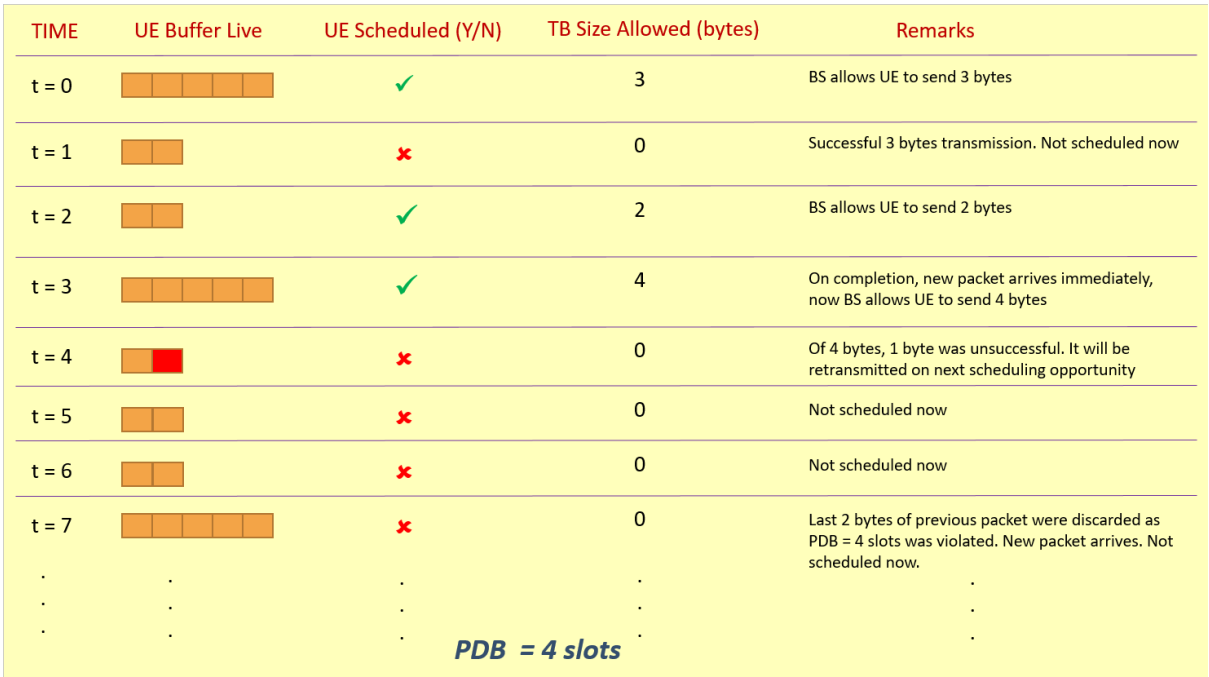

Figure 13: Demonstration of Data Packet Management for the problem

In the example demonstrated in the figure 13 above, the UE has applications packets with packet size = 5 bytes and PDB = 4 slots. The UE buffer initially has a complete application packet with 5 bytes. Then, UE gets an opportunity to be scheduled with transport block (TB) size = 3 bytes. Between slot t = 0 and t = 1, the signaling between BS and UE confirms the successful reception of three bytes; hence UE has discarded the three successful bytes from the buffer.

At t = 2, UE again gets a scheduling opportunity with TB = 2 bytes. Once again, this transmission is successful. As described before, upon a complete successful application packet transmission, UE immediately generates a fresh application packet with packet size = 5 bytes and PDB = 4 slots, as shown on t = 3. At this instant, UE gets an opportunity to transmit 4 bytes, and it does so. However, during the t = 3 and t = 4, signaling between UE and BS indicates that there has been an unsuccessful delivery of the 1 byte. Therefore, from the transmitted 4 bytes, UE discards the three successful from the buffer while keeping the one unsuccessful. UE now awaits the scheduling opportunity. At t = 7, the current application packet (with 2 bytes) exceeds the PDB requirement; hence UE discards the remaining 2 bytes of the current application packet, and a fresh new packet arrives once again with packet size = 5 bytes and PDB = 4 slots. This deletion of the old

application packet and arrival of a new packet ensures the continuity of transmission and freshness of transmitting information and avoids bottlenecking.

As shown in the figure 13 above, application packet and transport block sizes differ. The transport block size is the allowed transmitting size by the base station. It's decided by the base station when it's scheduling the particular UE. The calculation of the transport block size depends on the number of PRBs allocated to the given user and the Modulation and Coding Scheme (MCS) assigned to the same user.

MCS depends on the channel quality and defines how many bits are transmitted per resource element (RE). The coding rate is a ratio of useful bits to total transmitted bits. The coding rate is based on Channel Quality Indicator (CQI) sent from the UE in Cell Reference Signals (CRS). The modulation scheme defines the selection between different possible modulation options, i.e., 16-QAM, 4-QAM, QPSK, etc.

The transport block size calculation for uplink transmission is defined by 3GPP Release 36.213 [34]. The MCS level assigned to UE, TBS index is calculated from Table 8.6.1-1 in 3GPP TS 36.213. This table is quoted in this document below as Table 4. For example, MCS = 11 gives TBS Index = 10. Given the TBS index and allocated number of PRBs by the BS, Table 7.1.7.2.1-1 in 3GPP TS 36.213 maps the number of bits in a transport block. In the selected example, say for the number of PRBs = 4 and calculated TBS Index = 10, the number of bits per TB = 680.

Table 4: 3GPP TS 36.213, table 8.6.1-1: Modulation, TBS index and redundancy version table for PUSCH [34]

MCS Index I_{MCS}	Modulation Order Q'_m	TBS Index I_{TBS}	Redundancy Version rv_{idx}
0	2	0	0
1	2	1	0
2	2	2	0
3	2	3	0
4	2	4	0
5	2	5	0
6	2	6	0
7	2	7	0
8	2	8	0
9	2	9	0
10	2	10	0
11	4	10	0
12	4	11	0
13	4	12	0
14	4	13	0
15	4	14	0
16	4	15	0
17	4	16	0
18	4	17	0

19	4	18	0
20	4	19	0
21	6	19	0
22	6	20	0
23	6	21	0
24	6	22	0
25	6	23	0
26	6	24	0
27	6	25	0
28	6	26	0
29	reserved		1
30			2
31			3

4. SYSTEM MODEL AND PROBLEM FORMULATION

In this chapter, the core part of this work is explained, which explains formulating the problem and describing the system model for the given problem. In the beginning, the Age of Transmission (AoT) concept is described as a stepping stone

4.1 Age of Transmission

Age of Transmission (AoT) is a variant of a metric Age of Information (AoI). AoI quantifies the freshness of the information. In comparison to other time metric latency or delay, AoI captures the performance per data packet independently. It describes how fresh the transmitting packet is. A transmitting packet can be fast with higher speeds and lower latency but can be outdated information. The AoI quantifies this aspect of the data.

This is an essential metric in time-critical applications. The instructions passed remotely have varying importance depending on how fresh that instruction is. For example, a robot receiving signals remotely in a dynamic maze on a high-speed and low-latency channel can still perform poorly if the received signal contains outdated information. The same can be argued about telemedicine, space, vehicular networks, and military fields, where the freshness of the information determines its value of the information.

In diverse applications, AoI has been utilized in problem formulation [35] for communication and information systems. For example, AoI has been defined for problems concerning vehicle communications [36], radio networks [37], and IoT monitoring systems [38]. Modern applications rely on wireless communication and radio networks for most communication. As it's the case with many critical scenarios described before, the freshness of the information is as significant for the user experience.

XR, AR, or cloud gaming are real-time interactive communication platforms that rely on timely reception and information delivery for better performance. Since, in this work, the XR/AR use case has been considered as a reference, it makes more sense to incorporate a metric quantifying freshness of the information in the problem formulation.

AoI for a data packet is defined as the difference between the time the packet has been successfully received at the receiver and the time the packet was generated. For example, if the packet was generated on the transmitting side at $t = 2$ and is received entirely successfully at the receiver at $t = 5$, the AoI for this packet is 5 -2 = 3. This explains the information is not fresh by the duration of 3 time slots.

AoT is a variant defined for the problem-focused in this work. As described in chapter 3, the data packet generation in this work is continuous. In other words, UE must always have something to be transmitted. Therefore, the buffer on the UE side during the execution will never be empty. Because of the ever-existing transmittable packet on the UE side, the idea of packet generation is redundant. In this case, the duration for the last successful application packet becomes relevant.

AoT is defined as the time elapsed since the last application packet was successfully received at the receiver. With AoT, the delay for the current application packet will be quantified directly. This is valuable and relevant knowledge for the base station to schedule UEs to optimize the overall performance in an average sense. In addition, AoT covers PDB minimization, which is one of the principal motivations for this work.

AoT can be evaluated as an average value over a time horizon and a peak value. In this problem, the average AoT is considered a metric in the problem formulation. AoT is calculated for each UE under the base station's coverage independently, and the goal is

to minimize the average AoT for the base station. It's important to remember that AoT is measured per the application packet on the UE side, not concerning the transport block size allocated by the BS. The multiple transmissions of TBs for a single application packet will contribute to increasing AoT for the given application packet until successful delivery or dropout.

As described in chapter 3, the buffer status information for all the UEs is available at the BS. The buffering process at the UE side in the data packet management leads to the conclusion that every new application packet generation will lead to a state where the current buffer size will always be greater than the one in the state immediately before. Let's say B_{n_k}(t) represents the buffer size for the UE n_k at t. The event B_{n_k}(t+1) > B_{n_k}(t) will indicate the arrival of a new packet. The arrival of a new packet is synchronized with an event of completion of the previous packet. The completion can either be through successful delivery or timeout.

When a new packet is generated, the AoT for the given UE is reset. Till the completion of the packet, the AoT will keep incrementing by a factor of *1*. The data packet management described in the previous chapter in the figure 13 combined with the AoT will represent the flow as illustrated in the figure 14 below.

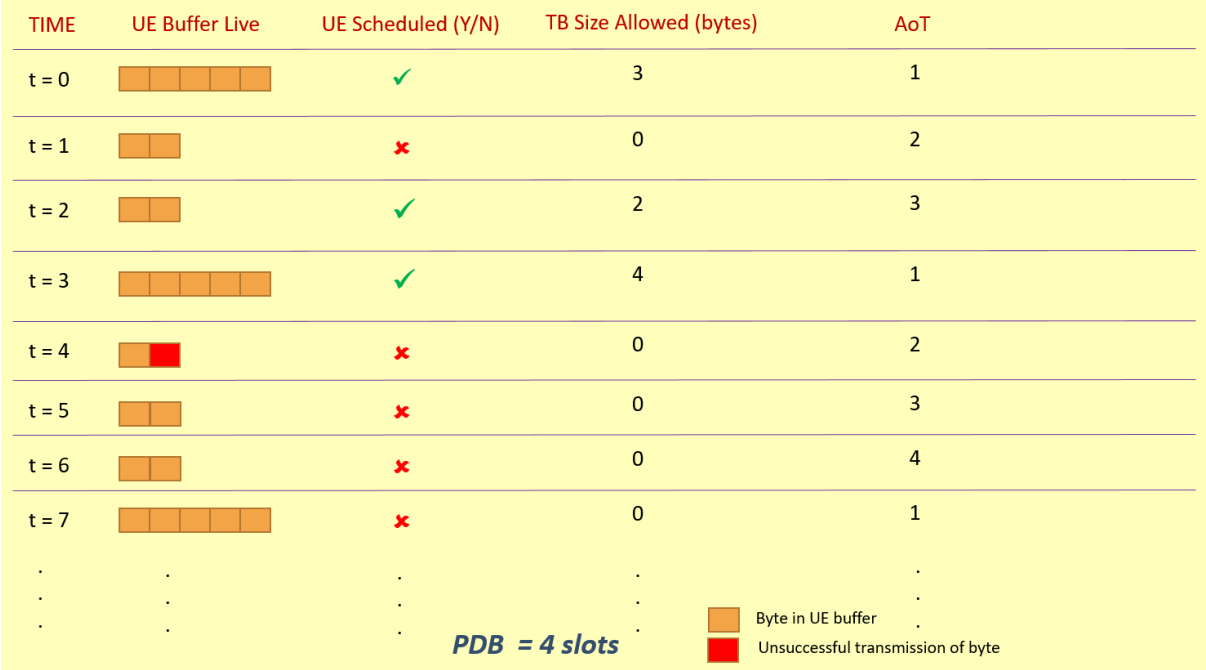

Figure 14: AoT flow demonstration

The AoT keeps on incrementing till the point a new packet arrives. The arrival of a new packet is consistent with the fact the new buffer size will always be greater than the last value. At this instant, the AoT for the given UE is reset to 1. Mathematically, this relationship is represented as shown below.

$$AT_{n_k}(t+1) = \begin{cases} 1, & B_{n_k}(t+1) > B_{n_k}(t) \\ AT_{n_k}(t) + 1, & else \end{cases} \quad (1)$$

For the UE n_k and time slot t + *1*, instantaneous AoT is represented by $AT_{n_k}(t+1)$. instantaneous AoT for UE n_k will keep on incrementing until the buffer status information indicates the arrival of new packet.

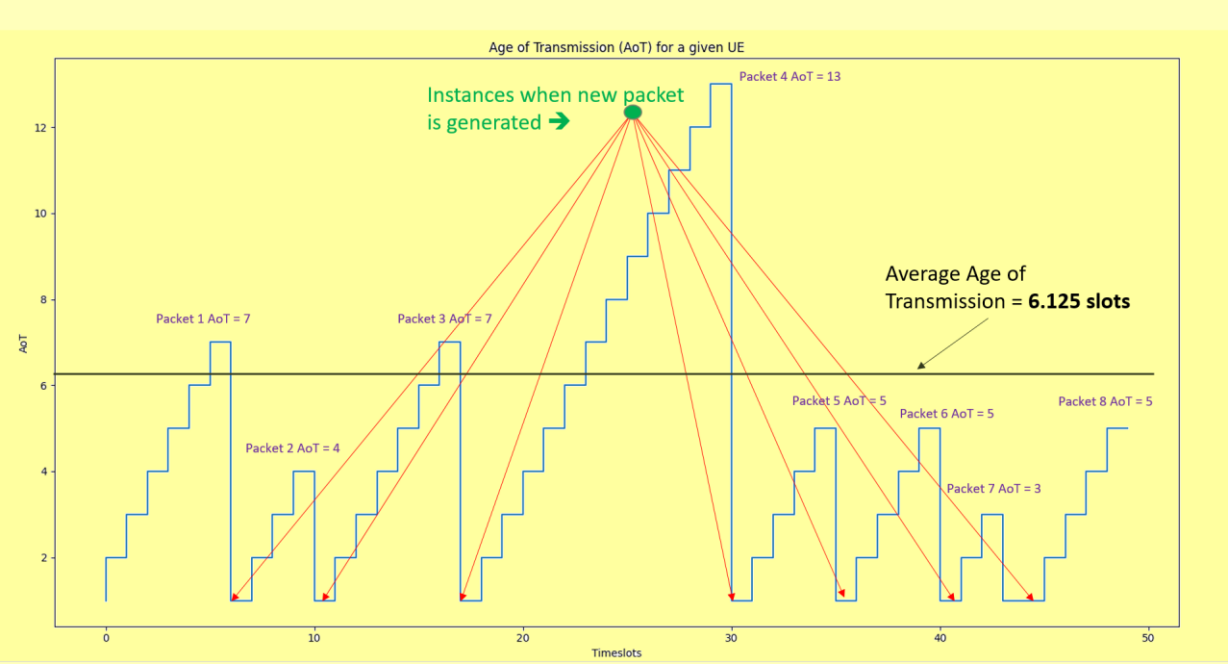

Figure 15: AoT plot over time

The figure 15 demonstrates the AoT flow during a transmission. For each UE, for each application packet arrived during the simulation, the AoT is calculated. The time duration between arrivals of the two consecutive application packets at the UE is termed as the AoT of the first of the two application packets. In other words, the AoT for the given packet is the time difference between the completion (successful delivery or timeout) of this packet and time the previous packet was completed (successful delivery or timeout). With all the AoT for all packets for a particular UE, the average AoT for given user is calculated.

For a UE n_k, if there are *P* application packets generated, each having respective AoT AT_p, then average AoT for a UE n_k is given in the equation (2):

$$avg\_AT_{n_k} = \frac{1}{P}\sum_{p=1}^{P} AT_p \quad (2)$$

For a BS *k* has total of N_k UEs in coverage, the average AoT for the BS *k* is given as:

$$avg\_AT_k = \frac{1}{N_k}\sum_{n=1}^{N_k} avg\_AT_{n_k} \quad (3)$$

In a network consisting of all *K* base stations, the average AoT per system:

$$avg\_AT = \frac{1}{K}\sum_{k=1}^{K} avg\_AT_k \quad (4)$$

4.2 System Model

An uplink multi-cell system is considered for this work as shown in figure 16. In this system model, there are *K* base stations (BS) denoted by the set $\mathcal{K}$. The BS *k* serves N_k UEs , $\forall k \in \mathcal{K}$. The set of served UEs by the BS *k* is represented by the set $\mathcal{N}_k, \forall k \in \mathcal{K}$.

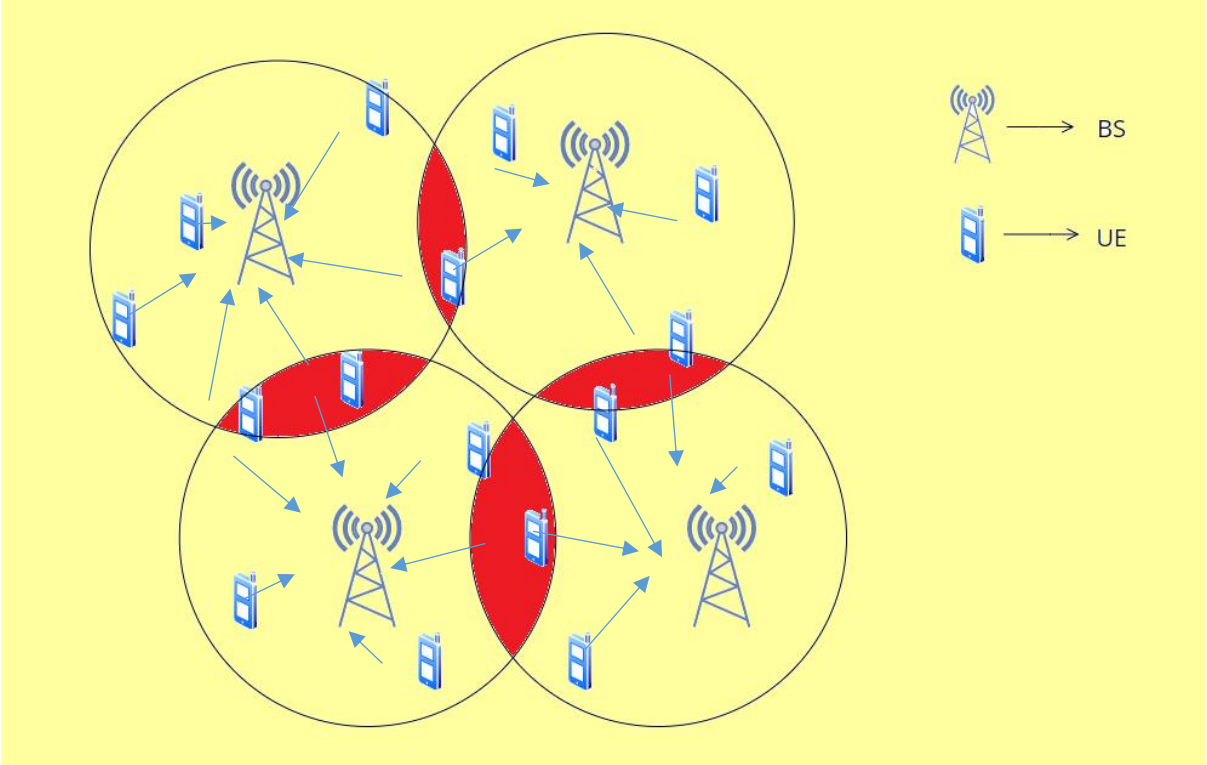

Figure 16: Network System Model

BS *k* is equipped with N_D antennas , $\forall k \in \mathcal{K}$. The maximum number of layers supported by the BSs is denoted by the set $\mathcal{L}$. $\overline{L_k}$ indicates the maximum number of layers supported by the BS *k*, $\forall k \in \mathcal{K}$. UE n_k is equipped with N_U antennas, such that $N_D \gg N_U, \forall n_k \in \mathcal{N}_k \; and \; k \in \mathcal{K}$. The maximum rank of the UE n_k is denoted by $\overline{L(n_k)}, \forall n_k \in \mathcal{N}_k \; and \; k \in \mathcal{K}$.

Set of base stations $\mathcal{K} = \{1, 2, \ldots . K\}$

Set of served UEs in BS *k* $\mathcal{N}_k = \{1, 2, \ldots . N_k\}$, $\forall k \in \mathcal{K}$

Set of maximum layers for BSs $\mathcal{L} = \{\overline{L_1}, \overline{L_2}, \ldots . \overline{L_K}\}$, $\forall k \in \mathcal{K}$

For physical transmission in this work, open-loop MIMO system is considered which performs transmission in time slots. The time slot is denoted by $t \in \mathcal{T} = \{0,1, \ldots . T\}$. The uplink transmission follow the OFDM resource allocation mechanism consisting of PRBs to be divided or shared between multiple UE in a given time slot. As described in chapter 3, it's assumed all the UEs to be served have data to be sent every time slot. At each slot, each BS selects a set of UEs from the list of available UEs for the given BS to be scheduled in current slot. The number of UEs selected by the BS *k* at time slot *t* for scheduling is represented by $\acute{N}_k$. These UEs belong to a selected UE set for the BS *k* at time slot *t* and are denoted by the set $\acute{\mathcal{N}_k}(t) \subseteq \mathcal{N}_k(t), \forall k \in \mathcal{K} \; and \; \forall t \in \mathcal{T}$.

Set of selected UEs by BS *k* at time slot *t*:

$$\acute{\mathcal{N}_k}(t) = \{1, 2, \ldots . \acute{N}_k\}, \forall k \in \mathcal{K}, \forall t \in \mathcal{T} \; and \; \acute{\mathcal{N}_k}(t) \subseteq \mathcal{N}_k(t)$$

In a scheduling decision, total layers assigned to BS k at time slot t is denoted by $L_k(t)$. The number of layers assigned to the UE n_k is denoted by $L(n_k; t) \forall n_k \in \mathcal{N}_k, k \in \mathcal{K} \; and \; t \in \mathcal{T}$.

The sum of assigned layers for all the UEs scheduled at a time slot *t* for BS *k* must be less than or equal to the layers assigned to the BS *k* at time slot *t*.

$$\sum_{k=1}^{N_k} L(n_k; t) \leq L_k(t), \quad \forall n_k \in \acute{\mathcal{N}_k}(t), \; \forall k \in \mathcal{K} \; and \; \forall t \in \mathcal{T}$$

When the BS makes a scheduling decision by selecting a set of UEs to be given an opportunity to transmit uplink data in a given time slot, it also assigns the respective MCS scheme for all the selected UEs. The MCS selection is based on channel condition and interference with co-scheduled UEs in the given time slot. MCS itself contains the modulation order and the coding rate for each scheduled UE.

For all the scheduled UEs for BS *k* at time slot *t*, denoted by $\mathcal{N}_k^{'}(t)$, set $\mathcal{M}_k(t)$ represents the modulation order and coding rate (MCS) values.

Set of MCS values for scheduled UEs: $\mathcal{M}_k(t) = \{M_1, M_2, \dots M_{\acute{N}_K}\}, \forall k \in \mathcal{K}$

In addition to MCS selection, BS also allocates the number of PRBs to each user. Based on these two information, transport block size for each user is calculated using the Table 4 and process explained in section 3.2

The BS has functionality of Hybrid Automatic Repeat Request (HARQ) which allows UE to receive ACK or NACK depending on reception status of the uplink transmission. UE has an opportunity to retransmit the unsuccessful packet for maximum of four times whenever given a scheduling opportunity next time.

The wireless uplink channel between UE and the BS is assumed to be flat and block fading. In other words, the channel for a given time slot remains same and changes between different time slots. It's assumed that the BS has availability of perfect channel information while making a scheduling decision. The BS has channel information through Sounding Reference Signals (SRS) or Demodulation Reference Signal (DMRS). The channel information is per PRB and is assumed to remain constant on all subcarriers within that PRB. For further work, the channel information can be further categorized on a level of resource elements (RE).

The received signal at the BS *k* at the time slot *t* is given as:

$$\boldsymbol{Y(t)} = \sum_{n_k \in \mathcal{N}_k^{'}(t)} \boldsymbol{H_{n_k}(t)\, X_{n_k}(t)} + \boldsymbol{I(t)}$$

Here, $\boldsymbol{H_{n_k}(t)}$ is the uplink MU-MIMO channel between the receiving antennas (RX) of the BS *k* and total layers for all the scheduled users in the BS *k* i.e. $n_k \in \mathcal{N}_k^{'}(t)$. If the base station *k* has $\mathbf{N_D}$ RX antennas and $\mathbf{N_L}$ are the total uplink layers for all the scheduled users, then $\boldsymbol{H_{n_k}(t)}$ channel has dimension of $\mathbf{N_D \times N_L}$. For instance, if there are total 14 layers for 6 scheduled users for time slot *t* and the base station *k* having 64 receiving antennas, the $\boldsymbol{H_{n_k}(t)}$ has dimension of 64 x 14. The elements of channel matrix are not identical and identity distributed (i.i.d) and has correlation among them because the antenna correlation both at the base station and user side is considered in channel characterization.

$\boldsymbol{X_{n_k}(t)}$ is the uplink signal vector from all the scheduled UEs' layers to the base station *k* in time slot *t*. The dimension of this vector is $\mathbf{N_L \times 1}$. In the example given in last para, the dimension will be 14 x 1. The $\boldsymbol{H_{n_k}(t)\, X_{n_k}(t)}$ overall has dimension of $\mathbf{N_D \times 1}$, which in current example is equivalent to 64 x 1.

The $\boldsymbol{I(t)}$ represents the interference at the BS *k* from all the users scheduled at the rest of the BSs. $\boldsymbol{I(t)}$ is expanded as given below:

$$\boldsymbol{I\,(t)} = \mathbb{E}\left[\sum_{l \neq k} \sum_{n_l \in \mathcal{N}_l^{'}(t)} \boldsymbol{H_{n_l}(t)\, X_{n_l}(t)}\right]$$

This is the sum of interference matrices from all the base stations $l \neq k$ to the base station *k*. Each base station $l, \forall l \in \mathcal{K}\ and\ l \neq k$ has set of respective scheduled users represented by $\mathcal{N}_l^{'}(t)$. The channel matrix between an interfering BS *l* and current base station *k* is denoted by $\boldsymbol{H_{n_l}}(\boldsymbol{t})$ and has a dimension of $\boldsymbol{N_D\ x\ \acute{N}_L}$. Here N_D are the BS *k* antennas and $\boldsymbol{\acute{N}_L}$ are the total number of layers for scheduled UEs in the interfering BS *l*. $\boldsymbol{X_{n_l}}(\boldsymbol{t})$ represents the vector containing uplink signals on all the transmitting layers for all the scheduled UEs at the interfering BS $l, \forall l \in \mathcal{K}\ and\ l \neq k$. $\boldsymbol{X_{n_l}}(\boldsymbol{t})$ has a dimension of $\boldsymbol{\acute{N}_L\ x\ 1}$.

Continuing the same example, if there are 5 interfering BSs each having 8 layers for 4 scheduled users in the same time slot *t*, $\boldsymbol{H_{n_l}}(\boldsymbol{t})$ for each of the five BSs has a dimension of 64 x 8 and $\boldsymbol{X_{n_l}}(\boldsymbol{t})$ has 8 x 1. The product $\boldsymbol{H_{n_l}}(\boldsymbol{t})\ \boldsymbol{X_{n_l}}(\boldsymbol{t})$ has dimension of 64 x 1. $\boldsymbol{I}(\boldsymbol{t})$ is the statistical expectation of sum of 5 vectors, each having 64 x 1 dimension.

Lastly, it's important to remind that the channel matrix and interference matrix have not i.i.d elements because of antenna correlations. In addition, the experiments are conducted in multiple drops represented by $\mathcal{D} = \{0, 1, \ldots D\}$, each having fixed number of time slots. During each drop, UEs are physically randomly placed in the entire coverage region consisting of all the BSs $\mathcal{K}$ with the initial channel configurations. The channel is time-varying channel over slot time within a single drop. When the drop changes, UEs are once again placed randomly with different initial and time-varying channel configurations.

4.3 Problem Formulation

In a system model explained in the previous sub-section, the problem in this work focuses on optimizing selected QoS parameters for uplink MU-MIMO XR traffic. In this problem, the following QoS parameters are considered:

a. Maximize the number of UEs within AoT threshold on a system level

b. Promote fairness in scheduled UEs with poor channel conditions

c. Maximize the system-wide average throughput given (a) and (b)

To meet these conditions, the problem can be demonstrated as shown in figure below:

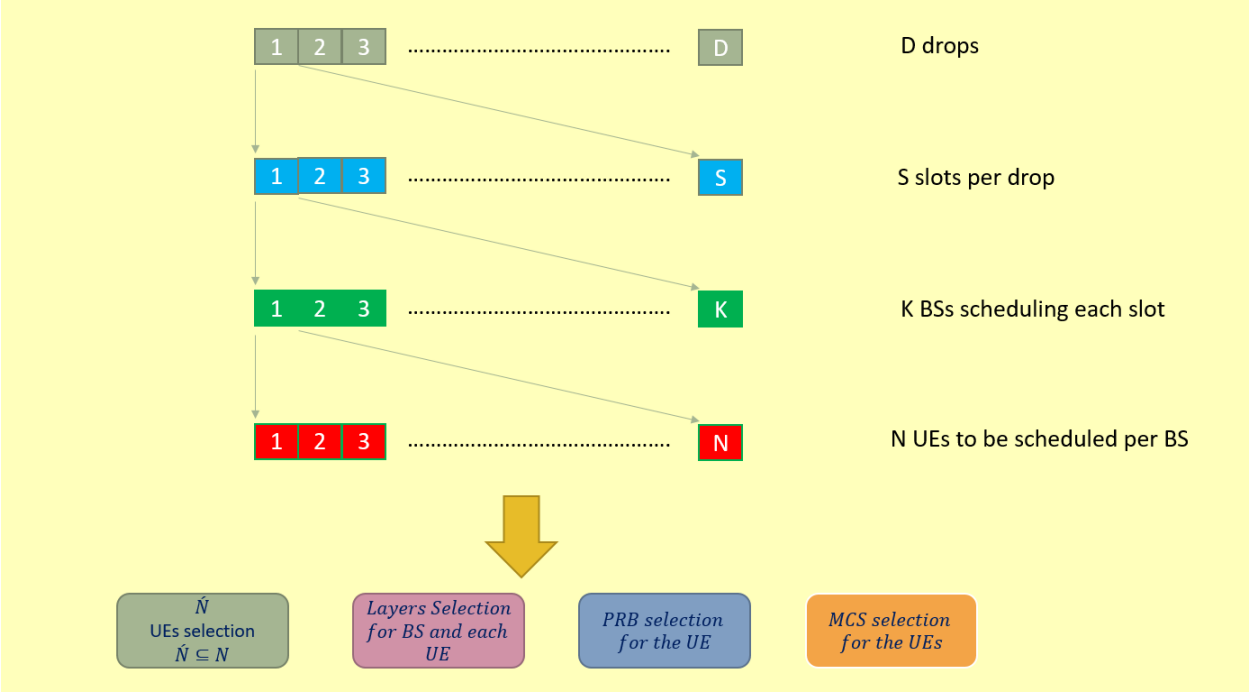

Figure 17: Problem Formulation

In descriptive sense, the optimization requires each base station in each time slot within each drop to schedule some UEs, allocate number of uplink scheduling layers, number of PRBs, and MCS levels to each of them. In this process, each BS ends up with optimum number of its layers for each time slot within each drop.

4.4 Scheduling Flow

The scheduling flow can be visualized as shown in figure 18 below.

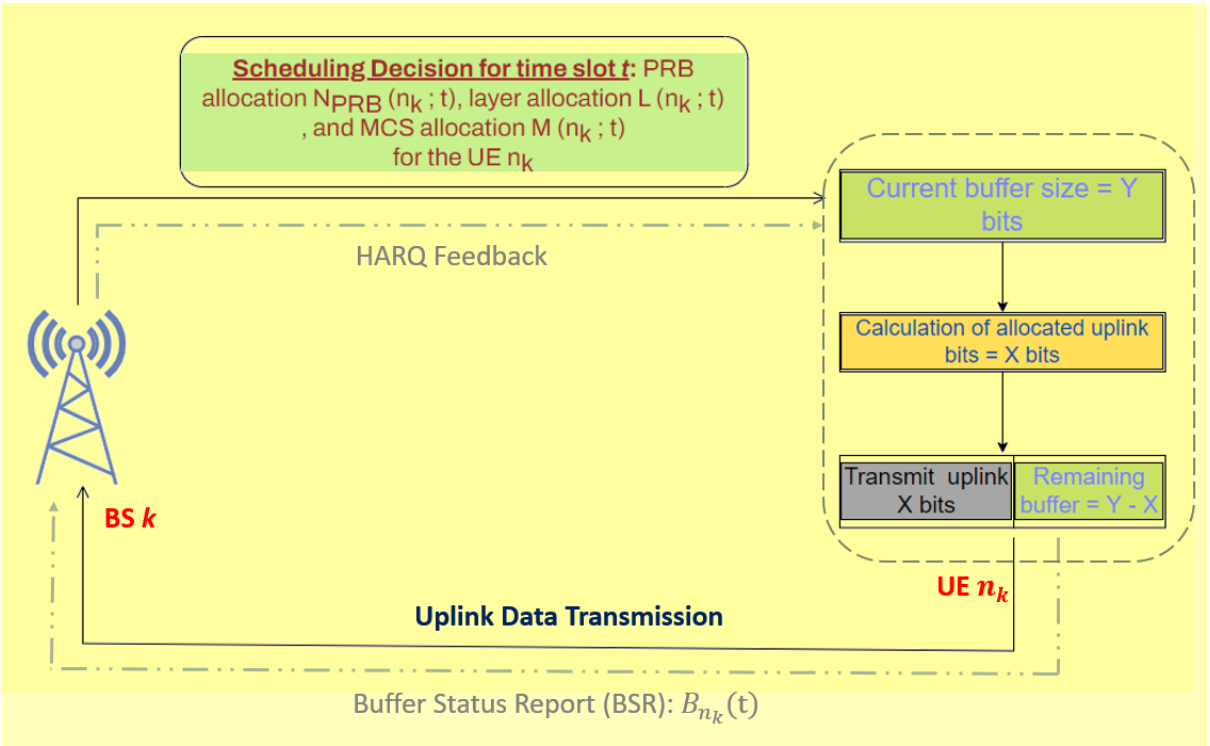

Figure 18: Scheduling Flow in the Problem

As it is evident the BS *k* makes a scheduling decision for a UE n_k which includes number of PRBs, number of uplink transmission layers, and MCS values. Based on these received information, UE assigns uplink transmitting bits and compute transport blocks. These TBs are forwarded for uplink data transmission from UE n_k to BS *k*. Also UE n_k keeps reporting its buffer information B_{n_k}(t) to the BS *k* for it to make optimum scheduling decision. Based on the uplink data received, BS evaluates the status of the reception and provides feedback in ACK/NACK through HARQ procedure.

4.5 KPI for the Problem

To evaluate and compare the performance of the proposed solution in the chapter 5 for the problem discussed in this chapter 4, the following key performance indicators (KPI) are considered:

1. The long term average Confidence Interval Throughput
2. Long-term average AoT

5. BASELINES AND PROPOSED APPROACHES

This chapter explains different baselines for uplink MU-MIMO scheduling algorithms for the problem described in chapter 4. After describing the baselines, the uplink MU-MIMO scheduling algorithm is proposed in this chapter.

All the scheduling algorithms are implemented on each of the BS at every time slot.

5.1 Baseline Algorithms

In order to compare the proposed approach in this work, it's important to define the uplink multi-user scheduling benchmarks.

5.1.1 Uplink Naive Round Robin Algorithm

Round robin (RR) is scheduling algorithm based on first in first out (FIFO) principle. The UEs to be scheduled in a sequential order as long as sum of layers of all the UEs is equal or less than max layers for the BS. This is named naive algorithm because it simply schedules UEs from the list of UEs and doesn't consider the any objective metric to select UEs. The algorithm pseudocode is provided below.

Algorithm: Uplink Naive Round Robin for BS *k* at time slot *t*

Initialize:

1 Co-scheduled UEs set = $\mathcal{N}_k^{'}(t)$ = [ ]

2 Served UE set = $\mathcal{N}_k(t) = \{u_1, u_2, \dots u_{N_k}\}$

3 $L_k(t) \leftarrow 0$

4 $i = 1$

5 while $\mathcal{N}_k^{'}(t) \neq \mathcal{N}_k(t)\ and\ L_k(t) < \overline{L_K}$ do

6 $L(u_i; t) \leftarrow \min\{\overline{L_K} - L_k(t),\ \overline{L(u_i)}\}$

7 $\mathcal{N}_k^{'}(t) \leftarrow \mathcal{N}_k^{'}(t) \cup \{u_i\}$

8 $L_k(t) \leftarrow L_k(t) + L(u_i; t)$

9 $\mathcal{N}_k(t) = \{u_{i+1}, \dots u_{N_k}, .. u_i\}$

10

11 end while

The notations used in the algorithm are in consistent with their description in sections 4.2 and 4.3. These notations are briefly described in Table 5.

Table 5: Description of notations used in Naive RR algorithm

Notation	Definition (for BS *k* in time slot *t*)

$\boldsymbol{\mathcal{N}_k^{'}(t)}$	Set of co-scheduled UEs
$\boldsymbol{\mathcal{N}_k(t)}$	Set of served UEs
$\boldsymbol{L_k(t)}$	Assigned number of layers for BS k
$\boldsymbol{\overline{L_K}}$	Maximum number of layers for BS k
$\boldsymbol{L(u_i; t)}$	Assigned number of layers to UE u_i served by the BS *k*
$\boldsymbol{\overline{L(u_i)}}$	Maximum rank (maximum number of layers) for UE u_i

This algorithm is used by the BS *k* at the time slot *t*. Initially, there are two lists for co-scheduled UEs and served UEs for the BS *k*. The co-scheduled UEs are empty at the beginning and the served UEs are the available UEs for the BS. The current layers for the BS *k* at time slot *t* is initialized to 0. The variable *i* represents the index for UEs in the algorithm.

A loop is run until the loop termination conditions are triggered. The two conditions are either the co-scheduled list of UEs becomes identical to served UEs list or the current layers for BS *k* exceed the maximum possible layers for the BS *k*. In each iteration of the loop, the current UE is selected from the front of served UEs list. This UE is assigned layers which is resultant of *min* operation on current available BS *k* layers and maximum rank for current UE as indicated in step 6. In step 7, the current UE is added to the co-scheduled UEs list. The step 8 increments the layers for BS k by the layers assigned to given user in step 6. Finally, in step-9, the current UE is removed from the front of the list and added back to the rear of the list.

As it can be clearly observed that this algorithm is quite primitive and basic in sense of selecting co-scheduled UEs which doesn't involve intelligent decision making factors and decision process. However, it ensures each user is served in a round robin manner. It doesn't particularly focuses on BS's throughput and UEs' QoS requirements in scheduling process. The utility function for this algorithm corresponds to the sum throughput for all the co-scheduled UEs for the current BS *k* at current time slot *t* within a current drop *d*.

5.1.2 Uplink Advanced Round Robin Algorithm

Uplink advanced round robin algorithm is partially similar to the previous algorithm. However, unlike the naive RR, advanced RR considers the channel and interference information while making a scheduling decision. The scheduling decision greedily focuses on expected sum rate of co-scheduled UEs for a BS in time slot. But since it's still a round robin algorithm, it selects UEs in a sequential order. The pseudocode for the uplink advanced round robin algorithm is given below.

Algorithm: Uplink Advanced Round Robin for BS *k* at time slot *t*

Initialize:

1 Co-scheduled UEs set = $\mathcal{N}_k^{'}(t)$ = [ ]

2 Served UE set = $\mathcal{N}_k(t) = \{u_1, u_2, \dots u_{N_k}\}$

3 $L_k(t) \leftarrow 0$

4 $\mathcal{R}\left(\mathcal{N}_k^{'}(t)\right) \leftarrow 0$

$i = 1$

$update\_flag \leftarrow True$

while $update\_flag$ is $\acute{T}rue$ and $\mathcal{N}_k\acute{}(t) \neq \mathcal{N}_k(t)$ and $L_k(t) < \overline{L_K}$ do

$L(u_i; t) \leftarrow \min\{\overline{L_K} - L_k(t),\ \overline{L(u_i)}\}$

if $\mathcal{R}\left(\mathcal{N}_k\acute{}(t) \cup \{u_i\}\right) > \mathcal{R}\left(\mathcal{N}_k\acute{}(t)\right)$ do

$\mathcal{N}_k\acute{}(t) \leftarrow \mathcal{N}_k\acute{}(t) \cup \{u_i\}$

$L_k(t) \leftarrow L_k(t) + L(u_i; t)$

$\mathcal{N}_k(t) = \{u_{i+1}, \ldots u_{N_k}, .. u_i\}$

$i \leftarrow i + 1$

else

$update\_flag \leftarrow False$

end if

end while

The notations used in the algorithm are in consistent with their description in sections 4.2 and 4.3. These notations are briefly described in Table 5 in previous section and Table 6 below.

Table 6: Description of notations used in Advanced RR algorithm

Notation	Definition (for BS *k* in time slot *t*)
$\boldsymbol{\mathcal{R}\left(\mathcal{N}_k\acute{}(t)\right)}$	Expected sum rate for co-scheduled UEs. It's a function of long term interference and channel matrices.
$\boldsymbol{\mathcal{R}\left(\mathcal{N}_k\acute{}(t) \cup \{u_i\}\right)}$	Expected sum rate for co-scheduled UEs and current UE u_i under consideration. It's a function of long term interference and channel matrices.
$\boldsymbol{\mathcal{R}_u(\mathcal{N}_k\acute{}(t))}$	Expected rate for the UE u when co-scheduled with other UEs within $\mathcal{N}_k\acute{}(t)$. It's a function of long term interference and channel matrices.
$\boldsymbol{update\_flag}$	Boolean variable to indicate to whether to continue or not the scheduling making process

As it's shown both in the algorithm code and Table 6, advanced RR considers the channel and interference information and their impact on the expected sum rate of the co-scheduled UEs. The term $\mathcal{R}\left(\mathcal{N}_k\acute{}(t)\right)$ is equivalent to sum rate of currently selected co-scheduled UEs. In advanced RR, the utility function described is the expected sum rate of co-scheduled UEs for the BS *k* in each time slot *t* within each drop *d*, which is equivalent to $\mathcal{R}\left(\mathcal{N}_k\acute{}(t)\right)$.

$$\mathcal{R}\left(\mathcal{N}_k\acute{}(t)\right) = \sum_{u \in \mathcal{N}_k\acute{}(t)} \mathcal{R}_u(\mathcal{N}_k\acute{}(t)) \quad (5)$$

The expected sum rate $\mathcal{R}\left(\mathcal{N}_k^{'}(t)\right)$ is initialized to zero in step 4 since no UE has been selected for scheduling. All the served UEs for BS k in time slot t are traversed iteratively in the loop. The loop has the same termination conditions as the one in naive RR with the addition of Boolean variable $update\_flag$ indicating the continuation of the process. As long as this flag is *True*, the process keeps on running and UEs are being traversed. However, once the algorithm identifies that adding the current UE to the currently selected co-scheduled UEs list will decrease the overall performance, the advanced RR algorithm halts the process and proceeds with the currently selected co-scheduled UEs.

It's important to remind that co-scheduling more UEs is a double-edged sword. Till a point, the more number of UEs will contribute positively to overall throughput since more users will be sending data hence system rate will increase. However, beyond a certain point the further addition of UEs actually degrades the overall BS performance. This is because the higher number of co-scheduled UEs will interfere with each other causing lower successful transmission bits (throughput) for each other. The goal of the uplink advanced RR algorithm is to select the optimum number of UEs for BS's overall performance.

Advanced RR unlike naive RR focuses on BS's throughput but doesn't consider other QoS parameter (AoT) for the served UEs. In addition, this algorithm is biased in scheduling the UEs with superior channel quality. Other UEs will struggle to get a chance to transmit uplink data.

5.1.3 Uplink Proportional Fair Algorithm

Uplink proportional fair (PF) algorithm focuses on both optimizing the overall long-term average throughput for the BS and the maintaining a sense of fairness in providing the UEs opportunity to transmit uplink data. Unlike the previous algorithms which either doesn't consider channel conditions or prefers only the UEs with better channel quality, PF does consider the UEs with poor quality while making a scheduling decision. Generally, it's not possible to mutually maximize system's throughput and fairness among users. However, the idea of PF is to find an optimum trade-off between two performance indicators. Uplink PF algorithm is described below.

Algorithm: Uplink Proportional Fair for BS *k* at time slot *t*

Initialize:

1 Co-scheduled UEs set = $\mathcal{N}_k^{'}(t) = [\,]$

2 Served UE set = $\mathcal{N}_k(t) = \{u_1, u_2, \dots u_{N_k}\}$

3 $L_k(t) \leftarrow 0$

4 $\mathcal{R}\left(\mathcal{N}_k^{'}(t)\right) \leftarrow 0$

5 $update\_flag \leftarrow True$

6 while $update\_flag^{'}\, is\, True\ and\ \mathcal{N}_k(t) \neq \emptyset\ and\ L_k(t) < \overline{L_K}$ do

7 $u^* \leftarrow \underset{u \in \mathcal{N}_k(t)}{\operatorname{argmax}} \mathcal{R}\left(\mathcal{N}_k^{'}(t) \cup \{u\}\right)$

8 $L(u^*; t) \leftarrow \min\{\overline{L_K} - L_k(t),\ \overline{L(u^*)}\}$

9 if $\mathcal{R}\left(\mathcal{N}_k^{'}(t) \cup \{u^*\}\right) > \mathcal{R}\left(\mathcal{N}_k^{'}(t)\right)\ do$

10 $\mathcal{N}_k^{'}(t) \leftarrow \mathcal{N}_k^{'}(t) \cup \{u^*\}$

11 $L_k(t) \leftarrow L_k(t) + L(u^*; t)$

12 $\mathcal{N}_k(t) = \mathcal{N}_k(t) \setminus \{u^*\}$

13 else

14 $update\_flag \leftarrow False$

15 end if

16 end while

The notations used in the algorithm are in consistent with their description in sections 4.2 and 4.3. These notations are briefly described in Table 6 and Table 7 in previous sections and Table 7 below.

Table 7: Description of notations used in PF algorithm

Notation	Definition (for BS *k* in time slot *t*)
$\boldsymbol{\mathcal{R}(\mathcal{N}_k^{'}(t))}$	Expected sum utility value for co-scheduled UEs. It's a function of long term interference and channel matrices as well as the PF exponential factor.
$\boldsymbol{\mathcal{R}(\mathcal{N}_k^{'}(t) \cup u^*)}$	Expected sum utility value for co-scheduled UEs and current UE u^*under consideration. It's a function of long term interference and channel matrices as well as the PF exponential factor.
$\boldsymbol{\mathcal{R}_u(\mathcal{N}_k^{'}(t))}$	Expected rate for the UE u when co-scheduled with other UEs within $\mathcal{N}_k^{'}(t)$. It's a function of long term interference and channel matrices as well as the PF exponential factor.
$\boldsymbol{Q_u^{\beta}}$(t)	Historical achieved goodput for the UE u with an initial value of $\epsilon > 0$
β	PF Exponential factor ∈ [0,1]

The major distinction in this algorithm is definition of expected sum utility value for the co-scheduled UEs rather than expected sum rate for the co-scheduled UEs. The utility function in PF algorithm considers both the throughput and fairness with expression given below:

$$\mathcal{R}(\mathcal{N}_k^{'}(t)) = \sum_{u \in \mathcal{N}_k^{'}(t)} \frac{\mathcal{R}_u(\mathcal{N}_k^{'}(t))}{Q_u^{\beta}} \quad (6)$$

As the equation (6) indicates the sum of utilities values' of all the co-scheduled UEs. The utility value for each UE is computed as a ratio between expected rate $\mathcal{R}_u(\mathcal{N}_k^{'}(t)$ of UE u when co-scheduled with other UEs within $\mathcal{N}_k^{'}(t)$ and past goodput $Q_u^{\beta}(t)$ for this UE u. As understandable, during the first slot, there is no past good put for any UE, which is why the initial value for $Q_u^{\beta}(t)$ is some small number ϵ.

The PF exponential factor β determines the weight of fairness in an algorithm and has a value between 0 and 1. The β = 1 indicates fully PF algorithm which weighs the past good put $Q_u^{\beta}(t)$ as much as it can while making a scheduling decision for the given UE u. On the other hand, β = 0 indicates algorithm doesn't weigh the past good put $Q_u^{\beta}(t)$ at all while making a scheduling decision for the given UE u. In this case, this algorithm is equivalent to uplink advanced RR described in section 5.1.2.

In the algorithm, after the initialization, a UE u^* from the list of the unselected UEs which maximizes the sum utility value (with already selected UEs $\mathcal{N}_k^{'}(t)$) is picked. If with this picked UE, the combined sum utility value of this UE u^* and already selected UEs $\mathcal{N}_k^{'}(t)$ is higher than the sum utility value of already selected UEs $\mathcal{N}_k^{'}(t)$, the UE $u^*$is added to the selected UEs list. This process will keep on repeating as long as the UE u^*under consideration is adding to the sum utility value. Otherwise, the process halts and schedules the so far selected UEs.

Uplink PF algorithm considers both system throughput, channel information, and UEs with poor channel condition for scheduling. Nonetheless, still this algorithm is limited to meet the UEs with QoS requirements considering AoT.

5.2 Proposed Algorithm

In this section, the proposed algorithm which is the core part of this research work is presented. As emphasized the importance of Age of Transmission (AoT) metric in chapter 3 for XR/AR and cloud gaming application, the proposed algorithm named ***Weighted Proportional Fair (WPF)*** considers optimizing the system's long-term throughput as well as the UEs AoT.

The traditional algorithms are either devoid of consideration for channel (uplink naive RR in section 5.1.1) or biased towards the UEs with better channel quality in order to maximizing the sum throughput (uplink advanced RR in section 5.1.2) or focusing on throughput and fairness metric without consideration for AoT requirements for the UEs (uplink PF in section 5.1.3). The motivation behind coming up with proposed weighted proportional fair (WPF) algorithm for uplink MU-MIMO scheduling is to consider optimization throughput, fairness, AoT requirements, and opportunity for poor channel UEs. As one can imagine, it is a challenging task to satisfy every user in terms of every aspect.

WPF after modeling requires the parameter optimization stage to find the optimum parameters for better results. The parameters optimization is discussed in next chapter 6. The WPF algorithm is described below.

Algorithm: Uplink Weighted Proportional Fair for BS *k* at time slot *t*

Initialize:

1 Co-scheduled UEs set = $\mathcal{N}_k^{'}(t)$ = [ ]

2 Served UE set = $\mathcal{N}_k(t) = \{u_1, u_2, \dots u_{N_k}\}$

3 $L_k(t) \leftarrow 0$

4 $\mathcal{R}\left(\mathcal{N}_k^{'}(t)\right) \leftarrow 0$

5 $update\_flag \leftarrow True$

6 while $update\_flag$ ́ $is\ True\ and\ \mathcal{N}_k(t)\ \neq \emptyset\ and\ L_k(t) <\ \overline{L_K}$ do

7 $u^* \leftarrow \underset{u \in \mathcal{N}_k(t)}{\operatorname{argmax}} \mathcal{R}\left(\mathcal{N}_k^{\acute{}}(t)\ \cup \{u\}\right)$

8 $L(u^*; t) \leftarrow \min\{\overline{L_K}\ - L_k(t),\ \overline{L(u^*)}\}$

9 if $\mathcal{R}\left(\mathcal{N}_k^{\acute{}}(t)\ \cup \{u^*\}\right) >\ \mathcal{R}\left(\mathcal{N}_k^{\acute{}}(t)\right)\ do$

10 $\mathcal{N}_k^{\acute{}}(t) \leftarrow \mathcal{N}_k^{\acute{}}(t)\ \cup \{u^*\}$

11 $L_k(t) \leftarrow L_k(t) + L(u^*; t)$

12 $\mathcal{N}_k(t) = \mathcal{N}_k(t) \setminus \{u^*\}$

13 else

14 $update\_flag \leftarrow False$

15 end if

16 end while

The notations used in the algorithm are in consistent with their description in sections 4.2 and 4.3. These notations are briefly described in Table 5, Table 6, and Table 7 in previous sections and Table 8 below.

Table 8: Description of notations used in WPF algorithm

Notation	Definition (for BS *k* in time slot *t*)
$\boldsymbol{\mathcal{R}\left(\mathcal{N}_k^{\acute{}}(t)\right)}$	Expected sum utility value for co-scheduled UEs. It's a function of long term interference and channel matrices as well as the UEs AoT requirements.
$\boldsymbol{\mathcal{R}\left(\mathcal{N}_k^{\acute{}}(t)\ \cup u^*\right)}$	Expected sum utility value for co-scheduled UEs and current UE u^*under consideration. It's a function of long term interference and channel matrices as well as the UEs AoT requirements.
$\boldsymbol{\mathcal{R}_u(\mathcal{N}_k^{\acute{}}(t)}$	Expected rate for the UE u when co-scheduled with other UEs within $\mathcal{N}_k^{\acute{}}(t)$. It's a function of long term interference and channel matrices as well as the UEs AoT requirements.
$\boldsymbol{ATF_u(t)}$	Age of Transmission (AoT) factor value for the UE u

The WPF algorithm evaluates the expected sum utility value differently by considering channel information, interference information and UEs AoT requirements. In other words, while estimating the utility throughput and AoT requirements are considered. The expected sum utility value for co-scheduled UEs is given by the equation (7).

$$\mathcal{R}\left(\mathcal{N}_k^{\acute{}}(t)\right) = \sum_{u \in \mathcal{N}_k^{\acute{}}(t)} \mathcal{R}_u\left(\mathcal{N}_k^{\acute{}}(t)\right)\ x\ ATF_u(t) \quad (7)$$

Here the sum utility value is the sum of individual utilities for all the co-scheduled UEs $\mathcal{N}_k^{\acute{}}(t)$. The individual utility value for each UE u is the product of expected rate $\mathcal{R}_u(\mathcal{N}_k^{\acute{}}(t)$ of UE u when co-scheduled with other UEs within $\mathcal{N}_k^{\acute{}}(t)$ and AoT factor value $ATF_u(t)$ for this UE u.

The calculation of $ATF_u(t)$ is explained below with the help of equations (8) and (9).

$$\omega_u(t) = \gamma F + (\overline{A_u} - AT_u(t)) - \frac{\eta F \; x \; B_u(t)}{Z_u(t)} \quad (8)$$

$$ATF_u(t) = sgn(\beta F - e^{-\omega_u(t)})e^{-\omega_u(t)} \quad (9)$$

The AoT factor value $ATF_u(t)$ for the UE u is computed in two steps as shown in above equations. The intermediate step in equation (8) computes $\omega_u(t)$ for UE u. These equations contain different variables which are explained below in Table 9.

Table 9: Description of variables used for AoT factor value in WPF algorithm

Variable Name	Description (time slot *t*)	Value Range	Date Type
$\overline{A_u}$	Average AoT threshold for the UE u.	≥ 1	Integer
$AT_u(t)$	Instantaneous AoT value for the UE u as modeled in equation (1) in sub-section 4.1	[1,T]	Integer
$B_u(t)$	Instantaneous buffer status report (number of current unsent bits/bytes) for the UE u	[1, 1e^5]	Integer
$Z_u(t)$	Current time-average goodput for the UE u	≥ 0	Float
γF	Hyperparameter: Weighing factor for the net instantaneous AoT difference ($\overline{A_u} - AT_u(t)$)	≥ 0	Float
ηF	Hyperparameter: Weighing factor for the required slots ratio ($\frac{B_u(t)}{Z_u(t)}$)	(0,1)	Float
βF	Hyperparameter: Weighing factor for consideration or not for UEs with exceptionally poor channel quality	≥ 1	Float

The algorithm in sequence is similar to the one described for uplink PF in section 5.1.3. However the metric to compute utility is updated with incorporation of AoT factor value.

The $\gamma F, \eta F, and\ \beta F$ are the tunable hyperparameters which influence the AoT factor value and consequently the utility value for Uplink WPF algorithm. In equation (8), $\left(\overline{A_u} - AT_u(t)\right)$ captures how far or close UE's instantaneous AoT from the threshold. If UE's instantaneous AoT is higher (UE has not been able to transmit application packet for long duration), the difference will be lower. Hence it will decrease the $\omega_u(t)$, impacting positively the $ATF_u(t)$. In other words UE u will be given higher weightage for scheduling. The γF increases the flexibility around AoT threshold.

The second term in equation (8) is a ratio $\frac{B_u(t)}{Z_u(t)}$ between current UE buffer size $B_u(t)$ (in bits/bytes) and time-average goodput $Z_u(t)$ (in bit per second/ bytes per second) for the UE. The ηF provides the weighting factor for this ratio. The calculation of $Z_u(t)$ considers the past good put and current expected rate. If UE has no history of goodput (scheduled for the first time), only the current expected rate is considered. Otherwise, both rates are weighed equally.

The WPF algorithm ensures that some exceptionally outliers UEs (if any) which have consistently extremely poor channel quality and close to zero success odds are not scheduled in order to avoid overall degradation of the system's performance. This is managed by the βF with the sgn operator. If UE is a long lost case, the $\omega_u(t)$ will be so that that the $e^{-\omega_u(t)}$ will be quite positive value. In case this very positive value exceeds the βF, a negative $ATF_u(t)$ will discourage the scheduling of such rare UE.

In addition to the co-scheduled UE selections and their respective transmitting layers, the PRB allocation $N_{PRB}(u;t)$ is done through open loop power control subjected to available channel bandwidth and UE available power.

Uplink WPF algorithm considers system throughput, channel information, and UEs with poor channel condition for scheduling. Unlike with all the baselines, the proposed algorithm focuses on meeting QoS requirements consisting of AoT for XR/AR and cloud gaming uplink traffic in MU-MIMO configuration.

5.3 Comparison of Algorithms

In the last part of this chapter, a brief comparative summary of baseline and proposed algorithms is provided in the Table 10 below.

Table 10: Comparative Summary for uplink scheduling algorithms

Uplink Algorithm	Optimizes Throughput	Optimizes AoT	Optimizes Fairness	Scheduling Metric
Naive Round Robin	NO	NO	NO	No metric
Advanced Round Robin	YES	NO	NO	Expected sum throughput of the co-scheduled UEs
Proportional Fair	YES	NO	YES	Expected fairness weighed sum throughput of the co-scheduled UEs
Weighted Proportional Fair [PROPOSED]	YES	YES	YES	Expected AoT weighted sum throughput of the co-scheduled UEs

6. SIMULATIONS AND RESULTS

This chapter describes different tools, simulation scenarios, simulation configurations, and obtained results for the proposed solution. The results are shown to compare and evaluate the performance of proposed approach against the baselines as described in chapter 5.

6.1 Tools

The Table 11 provides information about different tools and resources used for the experiments.

Table 11: Tools used for the simulations

Tool	Optimizes Throughput
Development Language	Python
Development Platform	Jupyter Notebook
Processing	GPUs and CPUs
Libraries/Packages	TensorFlow, NumPy, Docker, etc.
Link Level Simulations	Sionna
Channel Generation	Nokia Bell Labs internal package

6.2 Common Simulation Configurations

In this section, the shared simulation configurations related to communication systems are mentioned. The settings which distinguish different simulation experiments are described in the respective experiment sections later. The Table 12 provides the simulation common settings for all the experiments.

Table 12: Common Simulation Configurations

Index	Simulation Setting	Value
1	Channel Model	38.901-UMi [39]
2	Channel Bandwidth	100 MHz
3	Carrier Frequency	3.5 GHz
4	Number of Antenna subpanels	1
5	UE mobility speed	3 kmph
6	Subcarrier Spacing	30 kHz
7	Number of PRBs per channel bandwidth	273
8	Number of subcarriers per PRB	12
9	Number of OFDM symbols per slot	14
10	Maximum UE transmit power over full bandwidth	23 dBm
11	Target SNR	14 dB
12	Path loss compensation factor	0.85

13	Channel Estimation	Perfect
14	Inter-cell interference	Yes
15	Minimum PRBs for UE	4
16	Maximum PRBs for UE	273
17	HARQ Enabled	Yes
18	Maximum number of retransmissions	4
19	Number of drops	5
20	Number of time slots per drop	100
21	Application packet size	$1e^5$ bits

The settings defined in index 10 to 16 in Table 12 above are concerned with the open loop power control procedure to allocate the PRBs and maximum UE layers to the co-scheduled UEs. From a given channel, UEs are assigned to the primary BS with the available power for each UE.

The settings defined at index 19 and 20 are considered simulation duration. Each experiment is carried with 5 drops where each drop comprises of 100 time slots.

6.3 Experiments and Results

The first part of this section is focused on hyperparameter selection for the AoT factor calculation in the proposed algorithm as mentioned in equations (8) and (9). The purpose of this section is to demonstrate how impactful these hyperparameters are for the performance of the proposed approach.

Afterwards, the two main experiments are conducted to test the proposed solution for the problem described in chapter 5 and 4 respectively. The experiment 'A' considers the small scale network scenario with 3 BSs equipped with 16 RX antennas and 30 UEs within a network. On the other hand Experiment 'B' focuses on large scale network containing 21 BSs equipped with 64 RX antennas with 210 UEs within a network.

Finally, the comparison between two experiments is drawn to conclude which of the configurations enhances the performance of proposed approach over the baselines.

Before going directly to subsections, let's provide a brief summary of plots and their explanation which are going to be seen repeatedly for comparison in sections below.

Table 13: Summary of result plot parameters

Plot Parameter	Category	Explanation
Arithmetic Mean (AM) Throughput	Throughput	The average throughput for a UE across all drops, $\frac{1}{N}\sum_{n=1}^{N} TP_n$
Geometric Mean (GM) Throughput	Throughput	Geometric mean of the throughputs for all UEs across all drops, $(\prod_{n=1}^{N} TP_n)^{\frac{1}{N}}$
5th Percentile Average Throughput	Throughput	The average of throughputs of bottom 5%

		of UEs
10th Percentile Average Throughput	Throughput	The average of throughputs of bottom 10% of UEs
Average AoT	AoT	The average of AoT of all UEs across all the drops
Number of co-scheduled UEs	Co-scheduled UEs	Number of UEs transmitting simultaneously per time slot
Number of Spatial Layers per UE	MIMO layers	Layers for UE for transmitting uplink transmission
Number of Spatial Layers per BS	MIMO layers	Layers for BS for receiving uplink transmission

It's important to mention that for the AM and GM throughput values, the 90% confidence interval is considered. In other words, to calculate the mean and standard deviation, the 90% of the UEs are considered. Since these throughput values are estimations, that's why confidence interval is 90%. However, for each experiment in total of 500 time slots are executed per UE which is good enough size for a good estimation.

Moreover, for the calculation of AM throughput and GM throughput the bottom 5% of the UEs are not considered. The means are evaluated for the best 95% UEs in terms of throughput performance. While the 5th percentile and 10th percentile average throughput metrics give the average throughput values for the bottom 5% and 10% of the UEs respectively. The idea to segregate throughputs into best performing UEs and lower performing UEs it to compare how different algorithms in good conditions as well as in poor channel conditions.

6.3.1 AoT Hyperparameter Selection for Proposed Approach

The proposed approach in section 5.2 required hyperparameter values for the calculation of AoT factor. The two hyperparameters which impact the performance of the algorithm are γF and βF. The impact of these two hyperparameter on throughput and AoT performance for the proposed algorithm is described in subsections below. These experiments are conducted on large network containing 21 BSs and 210 UEs with 64RX antennas. The motivation is to demonstrate the effect these hyperparameters have on the performance.

6.3.1.1 Effect of γF on proposed algorithm

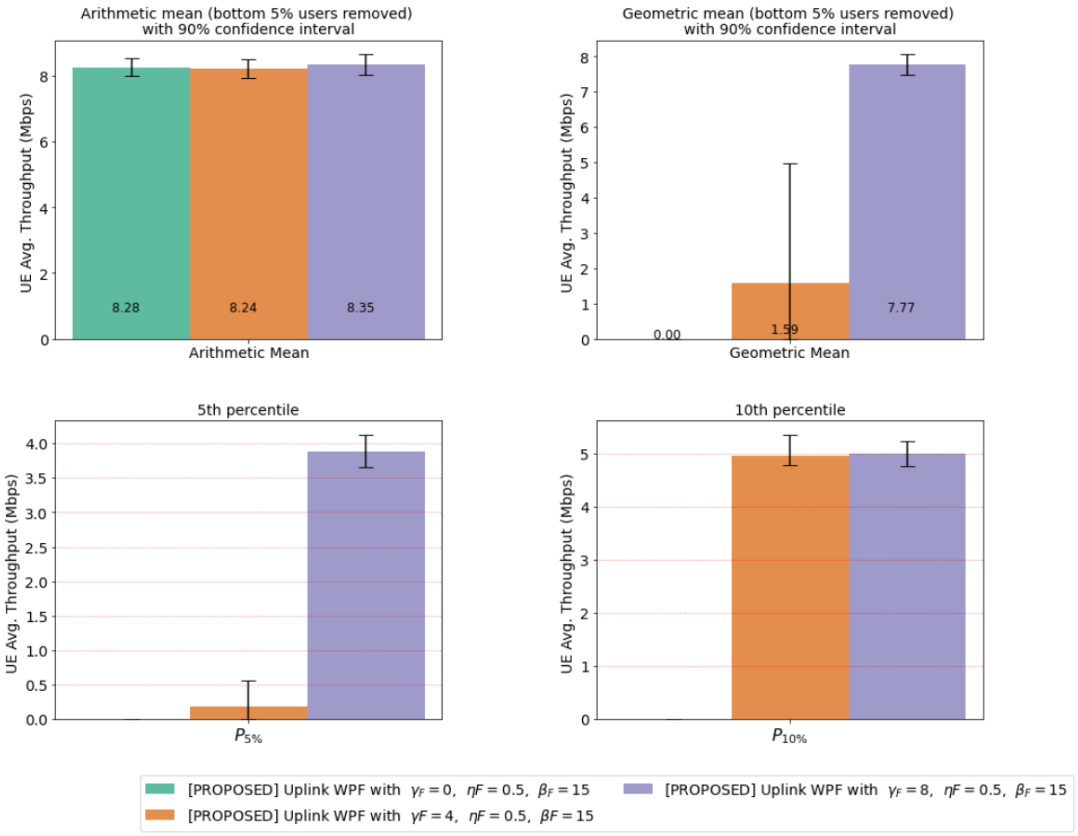

Figure 19: Effect of γF for throughput on proposed algorithm

The three different experiments are conducted with $\gamma F = [0, 4, 8]$ with $\eta F = 0.5$ and $\beta F = 15$. As it is obvious, the AM throughput has similar performance but for $\gamma F = 8$, the performance is massive for the UEs with poorer channel conditions. Similarly, the AoT comparison is shown below in figure 20.

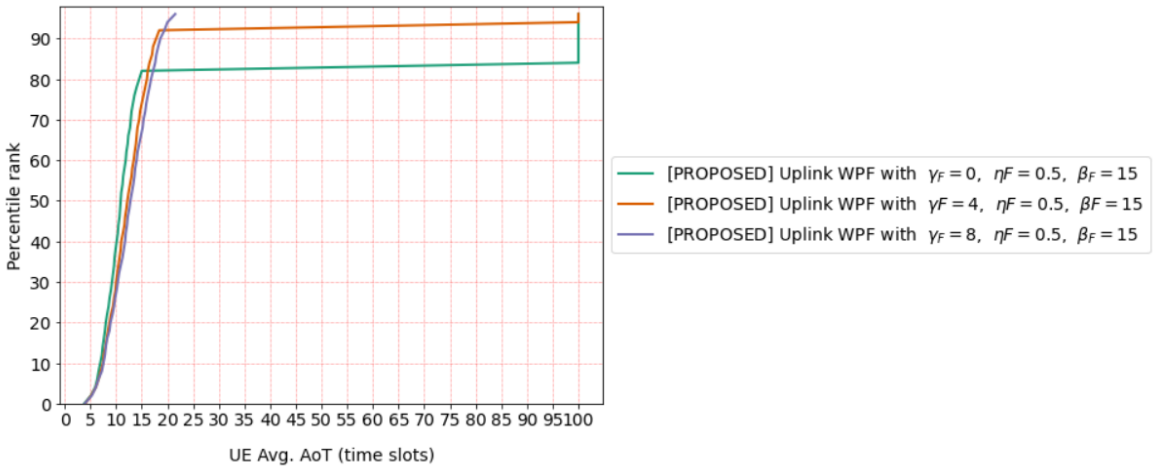

Figure 20: Effect of γF for AoT on proposed algorithm

Again $\gamma F = 8$ outperforms the other cases with able to satisfy the AoT requirements for 7% more UEs.

6.3.1.2 Effect of βF on proposed algorithm

Once the $\gamma F = 8$ is fixed let's see the impact of βF. The $\beta F = [15, 25]$ are experimented.

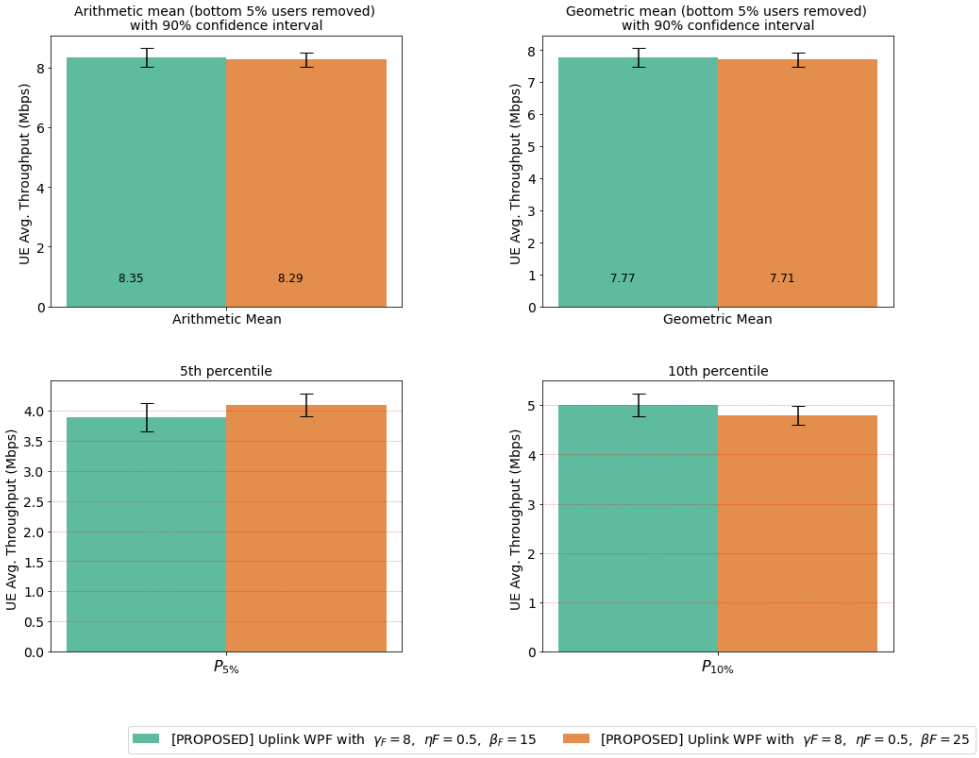

Figure 21: Effect of βF for throughput on proposed algorithm

There has been a marginal gain for $\beta F = 15$.

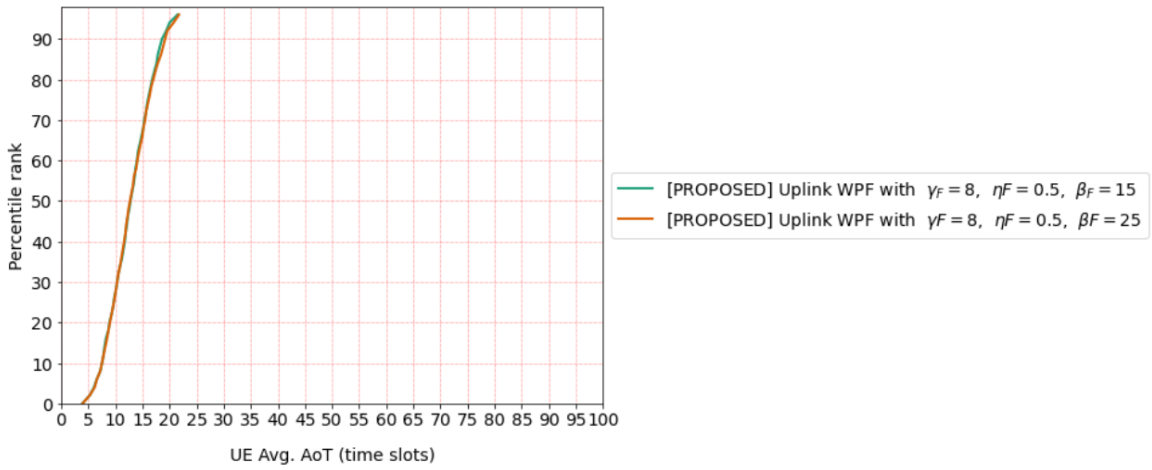

Figure 22: Effect of βF for throughput on proposed algorithm

There has not been much AoT difference between two configurations but overall $\gamma F = 25$ and $\beta F = 15$ the chosen configurations for the experiment with larger network as will be explained in section 6.3.3.

6.3.2 Experiment A: Small Network 3 BSs, 30 UEs, 16 RX

The simulation configurations concerned with Experiment 'A' are defined in Table 14 below.

Table 14: Simulation Configuration for Experiment A, Small Network 3 BSs, 30 UEs, 16RX

Index	Simulation Setting	Value
1	Number of BSs in network	3
2	Number of UEs in network	30
3	Average number of UEs per BS in network	10
4	Maximum BS layers	16
5	Number of BS RX antennas	16
6	PF exponential factor (β), Table 7	1
7	γF (hyperparameter for AoT calculation) , Table 9	15
8	ηF (hyperparameter for AoT calculation) , Table 9	0.5
9	βF (hyperparameter for AoT calculation) , Table 9	25
10	$\overline{A_u}$ (Average AoT threshold) , Table 9	25 time slots

In this experiment, a small scale network is simulated with configurations defined in Table 12 and Table 14. The performance of proposed algorithm against the baselines is observed and evaluated. The parameters concerning these algorithms are selected as given in index 6 to index 10 in Table 14.

6.3.2.1 Throughput comparison for Experiment A

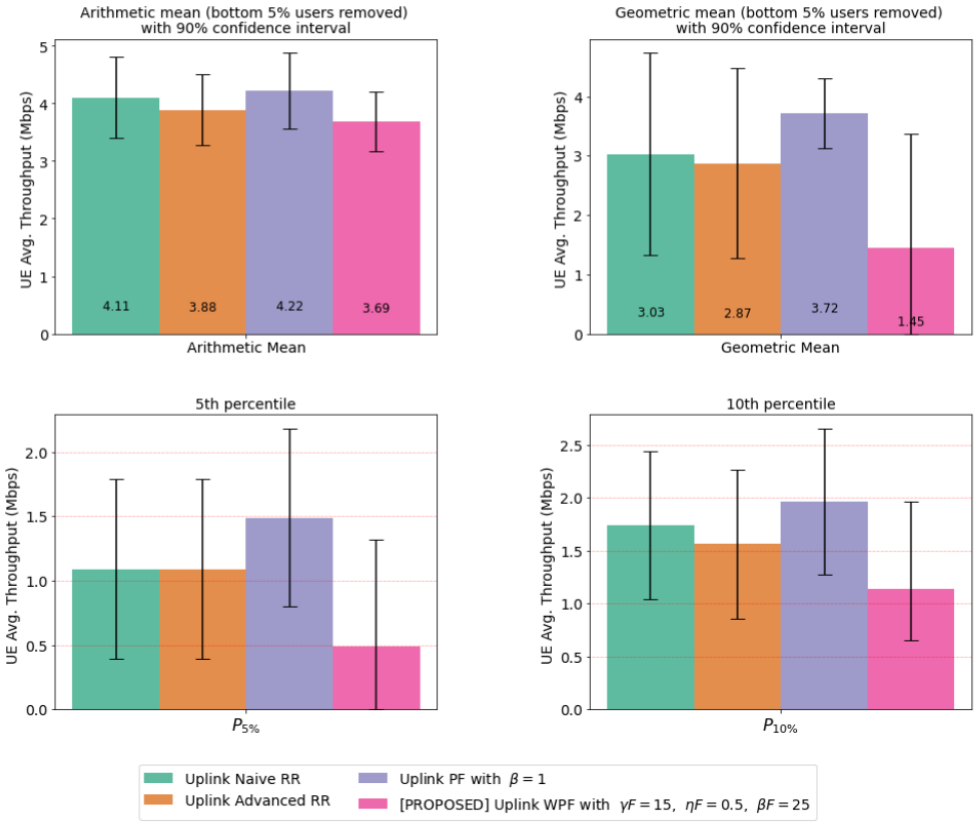

Figure 23: Throughput comparison for Experiment A

The throughput comparison between the three baselines algorithms and the proposed algorithms in figure 23 demonstrate that the proposed solution has actually marginally underperformed in terms of throughput.

The top left is the arithmetic mean (AM) of the throughput per UE in the network. The proposed uplink WPF has a small 0.53Mbps arithmetic mean throughput difference with the best performing baseline uplink PF. As expected uplink PF has been best performing baseline since it has the fairer and channel aware metric for scheduling among rest of the baselines. The top right is the geometric mean (GM) of the throughput. The uplink WPF has the lowest geometric mean throughput because it considers the UEs with poorer channel quality in scheduling henceforth comprising for throughputs for such UEs. Similarly 5th and 10th percentile UE average throughput for proposed algorithm is lesser than other algorithms. This is fundamentally because all these algorithms are heavily biased towards optimizing long-term average throughput for the system.

6.3.2.2 AoT comparison for Experiment A

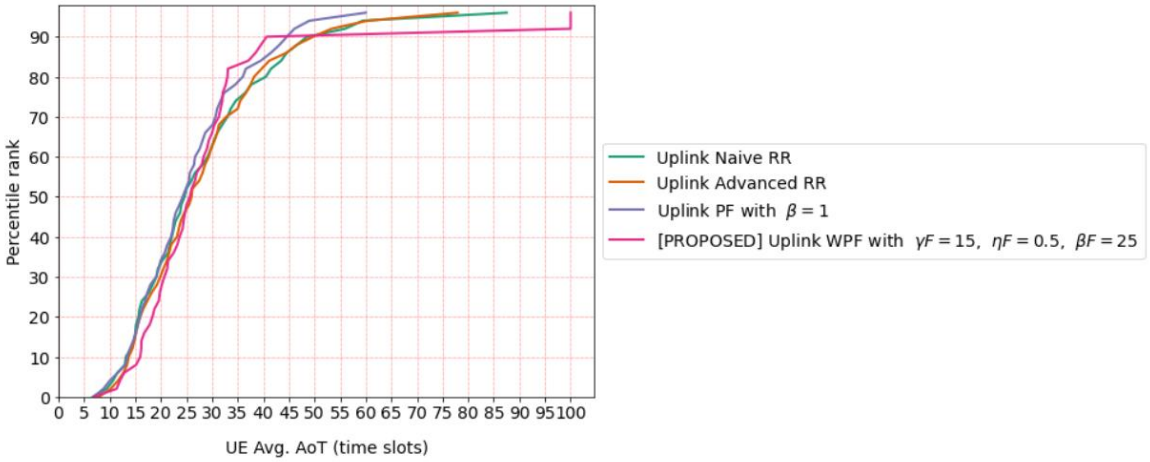

Figure 24: AoT comparison for Experiment A

The figure 24 shows the comparative performance of different algorithms from AoT perspective. The horizontal axis describes the time slots and vertical mentions the percentile value of respective time slot value. As we saw in the section 3.1, most XR/AR uplink traffic has PDB of around 30ms. There has not been massive difference in performance between different algorithms since the small scale network is under consideration in this experiment. Nonetheless, the proposed uplink WPF algorithm marginally outperforms the rest of the baselines around 35 time slots. Uplink WPF has 3% more UEs within 35 time slots than the second best uplink PF algorithm. It can be argued the marginally benefit PF had in throughput comparison has been slightly balanced by WPF in AoT aspect. This is intuitively understandable since the proposed WPF focuses on improving both the aspects which naturally has a trade-off.

6.3.2.3 Number of co-scheduled UEs comparison for Experiment A

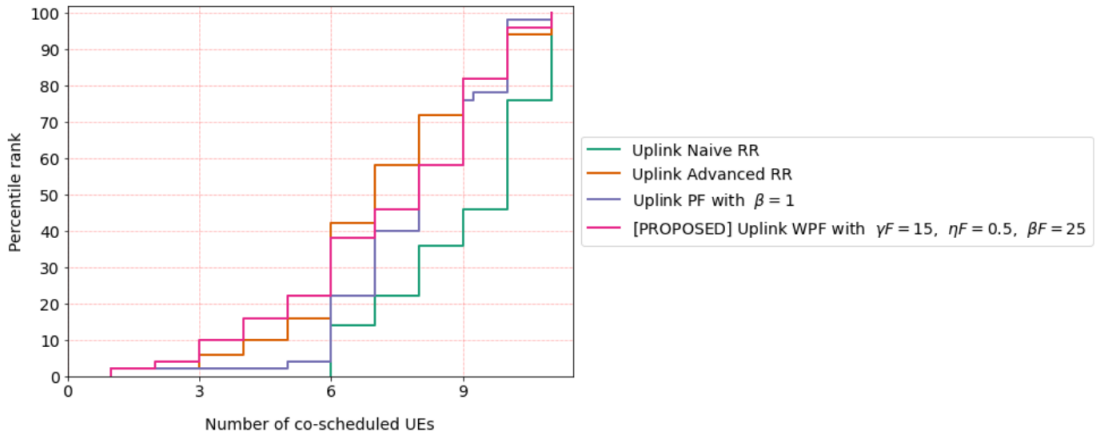

Figure 25: Number of co-scheduled UEs comparison for Experiment A

The figure 25 shows the percentile distribution for number of co-scheduled UEs per time slot. The uplink naive RR co-schedules the most number of UEs. This is primarily because the naive RR greedily focuses on filling as much UEs within the allowed layers at the BS. For half of the execution cycle, uplink advanced RR co-schedules the least 7 UEs followed by uplink PF and proposed uplink WPF co-scheduling 8 UEs. For the same frequency, uplink naive RR co-schedules 10 UEs per time slot.

6.3.2.4 Number of spatial layers comparison for Experiment A

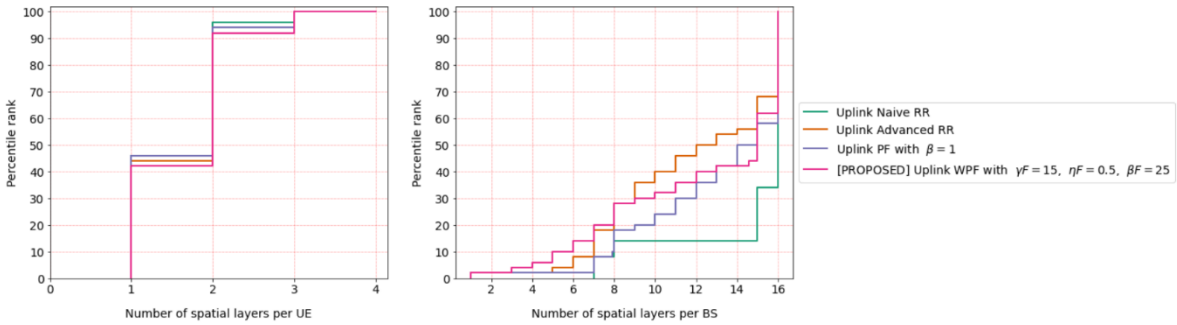

Figure 26: Number of spatial layers comparison for Experiment A

The two plots in figure 26 show the distribution of number of spatial layers per UE and number of spatial layers per BS. Generally, UE has 2 layers for 90% time for all the algorithms. Furthermore, BS operates on 16 layers throughout for all the algorithms.

6.3.3 Experiment B: Large Network 21 BSs, 210 UEs, 64 RX

The simulation configurations concerned with Experiment ‘B’ are defined in Table 15 below.

Table 15: Simulation Configuration for Experiment B, Large Network 21 BSs, 210 UEs, 64RX

Index	Simulation Setting	Value
1	Number of BSs in network	21
2	Number of UEs in network	210
3	Average number of UEs per BS in network	10

4	Maximum BS layers	16
5	Number of BS RX antennas	64
6	PF exponential factor (β), Table 7	1
7	γF (hyperparameter for AoT calculation) , Table 9	8
8	ηF (hyperparameter for AoT calculation) , Table 9	0.5
9	βF (hyperparameter for AoT calculation) , Table 9	15
10	$\overline{A_u}$ (Average AoT threshold) , Table 9	15 time slots

In this experiment, a large scale network is simulated with configurations defined in Table 12 and Table 15. The performance of proposed algorithm against the baselines is observed and evaluated. The parameters concerning these algorithms are selected as given in index 6 to index 10 in Table 15.

As in the large network, with large number of BSs and UEs, higher interference is expected. Moreover, with more RX antennas at BS (64 from 16), the performance of the algorithm is expected to improve.

6.3.3.1 Throughput comparison for Experiment B

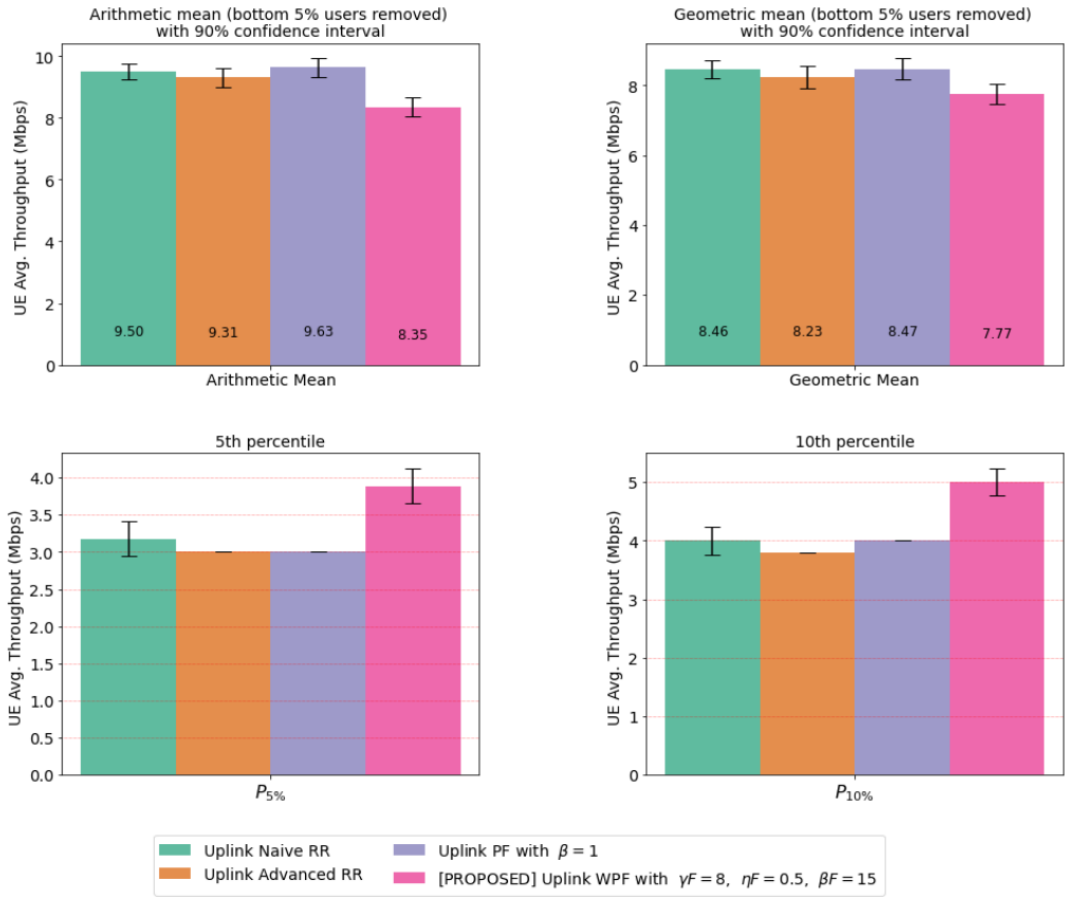

Figure 27: Throughput comparison for Experiment B

As expected with the higher RX and more practical model of 21 BSs and 210 UEs, the performance for the proposed approach has been better. In a broader sense, the proposed WPF algorithm actually performs the best among all candidate algorithms. The uplink WPF has 1.28Mbps lesser arithmetic mean throughput than the best performing uplink PF. With the large scale network, a remarkable improvement for geometric mean throughput for WPF with 64 RX than with 16 RX.

As it can been seen in figure 27, the AM throughput and GM throughput for all the algorithms are within similar range. This demonstrates that WPF AM and GM throughput performance is comparable to the baselines. In fact, the proposed algorithm performs best among all the algorithms for 5th and 10th percentile UEs in terms of throughput. The 5th percentile average throughput for proposed algorithm is 18.75% better than 2nd best. Similarly, the 10th percentile average throughput for proposed algorithm outperforms the 2nd best by 25%. With arguably comparable or better performance (for certain UEs) of proposed algorithm against the baselines in terms of throughput, the best is yet to come with second key QoS parameter AoT which is not focused in baseline algorithm. This validates the fact the proposed approach is heading towards right direction.

6.3.3.2 AoT comparison for Experiment B

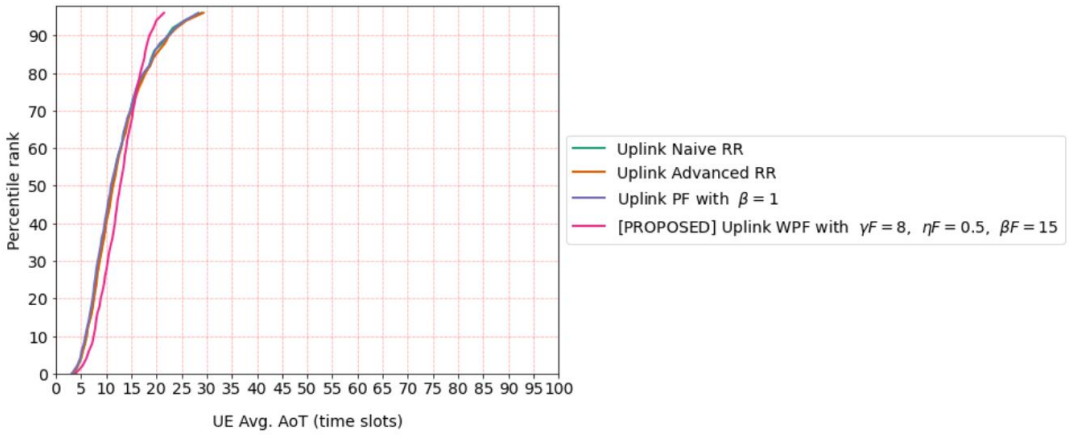

Figure 28: AoT comparison for Experiment B

Now with 64RX antennas at BS and larger network, it can be clearly seen how the proposed algorithm outperforms its competitors. The proposed algorithm on average schedules all users within 22 time slots. On the other hand, for the same number of UEs, baseline algorithms take around 29 time slots. This is a significant 7 time slots advantage for the XR/AR applications to transmit data with lesser delay and higher freshness.

As discussed in the section above, the throughput performance for the proposed algorithm was comparable to other algorithms. Now with superior AoT performance, it's reasonable to claim that the proposed algorithm is best performing uplink scheduling algorithm for the UEs running XR/AR or cloud gaming applications.

6.3.3.3 Number of co-scheduled UEs comparison for Experiment B

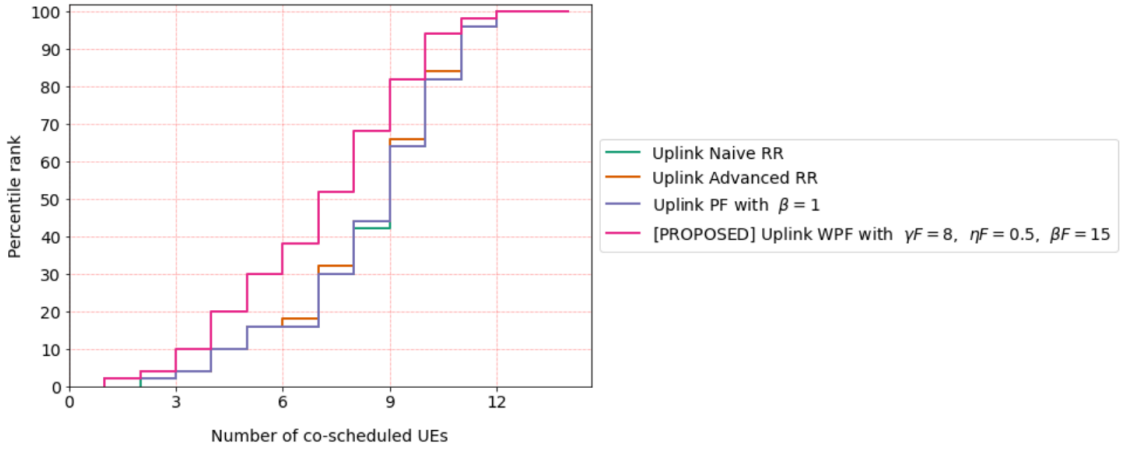

Figure 29: Number of co-scheduled UEs comparison for Experiment B

The figure 29 demonstrates not much difference in number of co-scheduled UEs for all the algorithms. Typically, the proposed algorithm on most occasions schedules 1 UE less than other algorithms. This can be explained with the communication theory concept of co-scheduling too much UEs leads to global negative behavior with higher interference.

6.3.3.4 Number of spatial layers comparison for Experiment B

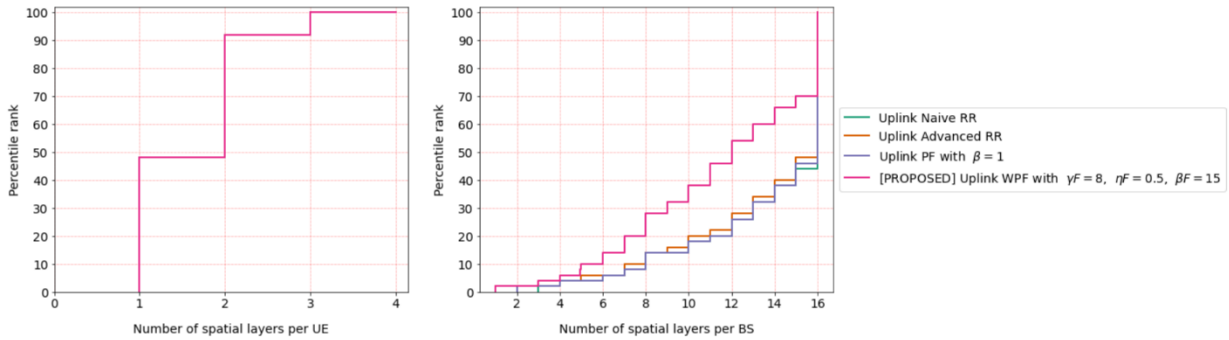

Figure 30: Number of spatial layers comparison for Experiment B

The layers per UE for all the algorithm is same. However, the layers for the BS are less for the proposed algorithm than the rest of the algorithms. This is an additional benefit since the performance is already better than other approaches and lesser number of layers for the BS can lead to power conservation at the BS.

6.3.4 Experiment A VS Experiment B

In this section, the performance of the proposed uplink WPF algorithm for experiment ‘A’ and experiment ‘B’ is compared. With the same configuration described in Table 14 and Table 15 for both experiments respectively, let’s find out the difference in terms of the same KPIs as we transcend from smaller network, lower RX antennas to bigger network, higher RX antennas.

6.3.4.1 Experiment A vs Experiment B Throughput

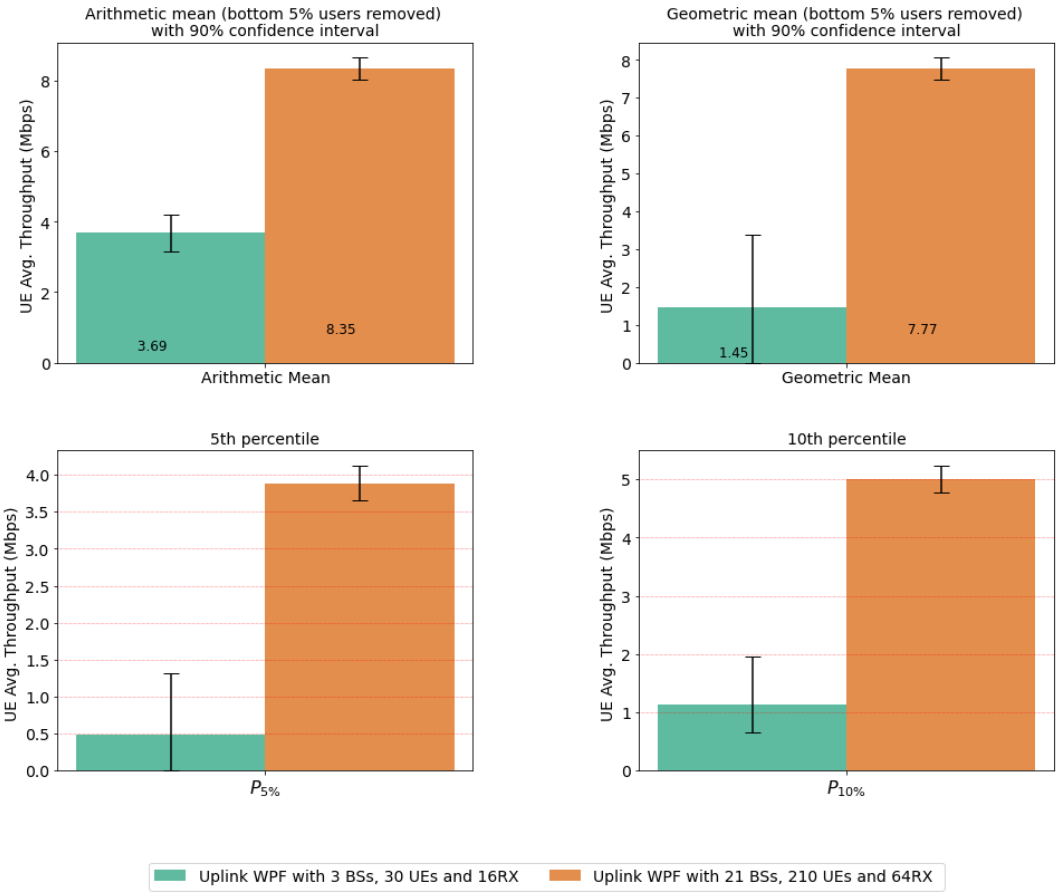

Figure 31: Experiment A vs Experiment B Throughput

The figure 31 speaks for itself how bigger MU-MIMO network improves the throughput massively from all aspects. The WPF algorithm has AM throughput with 64 RX 126% higher than 16RX. Similarly, the GM throughput difference is 436% in favor of the 21 BSs, 210 UEs, 64RX configuration. The 5th percentile throughput and 10th percentile throughput are 660% and 355% more in larger network scenario.

6.3.4.2 Experiment A vs Experiment B AoT

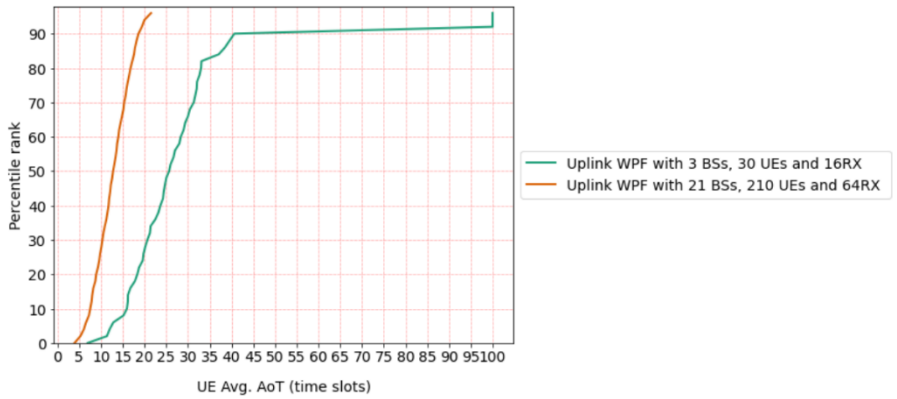

Figure 32: Experiment A vs Experiment B AoT

Same is the case with AoT performance as it is with throughput. With the higher RX BS, 57% more UEs are satisfied with AoT = 15 time slots. The higher RX scenario satisfies

AoT for all the users within 22 time slots. On the other hand, lower RX BS offers 90% of the UEs within 40 time slots (18 time slots later).

As with the both KPIs of throughput and AoT, it is evident how higher number RX antennas overwhelmingly boosts the performance and ensures meeting the UEs' QoS requirements. It's important to remind that both the experiments have same number of UEs per BS indicating the same amount of challenges and requirements to meet in both cases. If nothing, the bigger network has higher likelihood of the interference since more BSs and interfering UEs exist. Therefore, it makes sense how in experiment 'B', the proposed uplink WPF algorithm outperforms the rest of the algorithms.

6.3.4.3 Experiment A vs Experiment B number of co-scheduled UEs

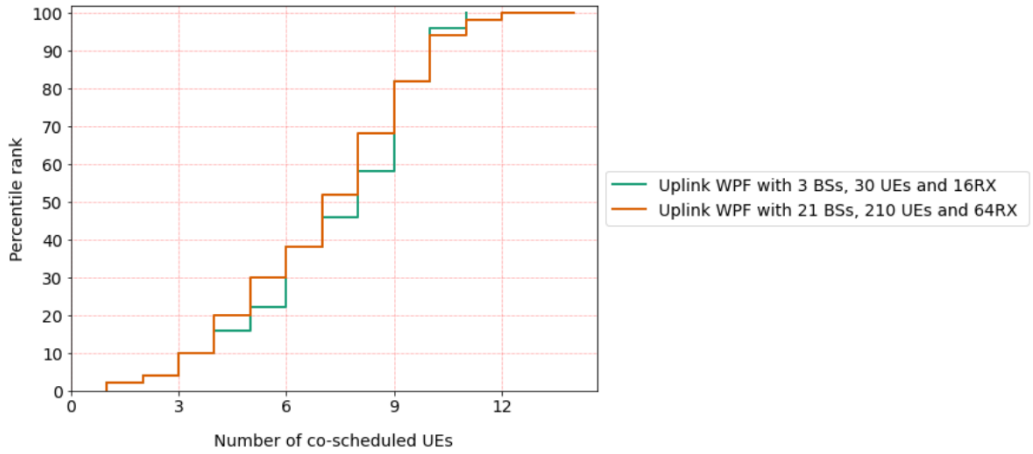

Figure 33: Experiment A vs Experiment B number of co-scheduled UEs

The number of co-scheduled UEs per time slot is same for both the configurations. It is understandable since number of competing UEs for each BS remains same (ten) for both configurations.

6.3.4.4 Experiment A vs Experiment B number of spatial layers

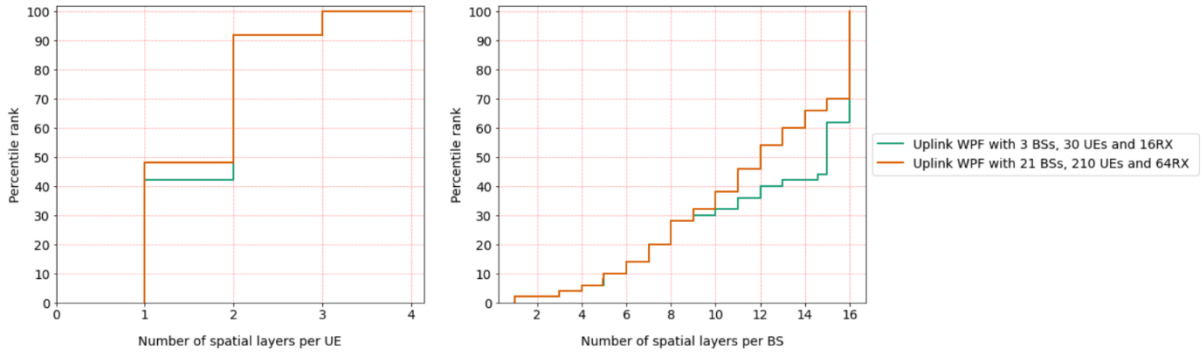

Figure 34: Experiment A vs Experiment B number of spatial layers

As was the case with co-scheduled UEs, the number of layers for UE and BS is also similar for both the cases as shown in figure 34.

6.4 Discussion and Future Directions

This work is the first of its kind. To the best of my knowledge, research has not been done focusing on uplink MU-MIMO user scheduling algorithms incorporating throughput and Age of Transmission (AoT) as QoS parameters. The motivation behind this work is to lay a foundation for this direction. As the results have demonstrated, MU-MIMO configuration with BS equipped with 64 RX antennas (Experiment B) outperforms all the baseline scheduling algorithms to meet UEs' data rate and delay requirements.

XR/AR and cloud gaming have explicitly been chosen as traffic use cases to emphasize that MU-MIMO uplink scheduling for modern applications is the need of the hour. In addition, the introduction of AoT as a QoS metric can provide valuable insight for future work to quantify and evaluate the freshness of the transmission.

Implementing machine learning and, in particular, reinforcement learning for the same problem can optimize the results of the proposed algorithm. The AoT hyperparameters can be mapped to machine learning models or policy network weights. The selection of algorithm parameters through artificial intelligence can optimize performance, computational complexity, and processing speed.

7. CONCLUSIONS

This work has explored the uplink scheduling algorithm for MU-MIMO configuration. Firstly, the XR/AR and cloud gaming use case has been considered a traffic model. The QoS requirements for XR applications consisting of throughput, PDB, and accuracy are defined per 3GPP standards. After finalizing the traffic model, the application packet model and management have been explained for the problem focused on in this work. Secondly, to quantify and evaluate the freshness of the transmitted information in the uplink direction, the AoT metric has been introduced qualitatively and quantitatively as one of this work's core contributions. Thirdly, the system model and problem formulation have been discussed, involving multiple BSs scheduling several UEs in each time slot within a drop. Fourthly, as a significant contribution to this work, a novel uplink scheduling algorithm has been proposed intended to satisfy the AoT requirements of the user while maximizing the average throughput. Finally, the performance of the proposed algorithm on the metrics of average throughput, AoT, and scheduling fairness has been evaluated against the standard baseline scheduling algorithms. The simulation results demonstrate how the proposed approach outperforms other candidate solutions in a broader QoS scale consisting of the abovementioned metrics.

ABBREVIATIONS - ACRONYMS

5QI	5G Quality Indicator
AoI	Age of Information
AoT	Age of Transmission
AR	Augmented Reality
BSR	Buffer Status Reporting
DMRS	Demodulation Reference Signal
HARQ	Hybrid Automatic Repeat Request
IoT	Internet of Things
M2M	Machine to machine
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
MR	Mixed Reality
MU-MIMO	Multi-user MIMO
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
PAPR	Peak to average power ratio
PDB	Packet Data Budget
PDCP	Packet Data Convergence Protocol
PF	Proportional Fair
PRB	Physical Resource Block
SRS	Sounding Reference Signal
QoE	Quality of Experience
QoS	Quality of Service
QAM	Quadrature Amplitude Modulation
RLC	Radio Link Control
RR	Round Robin
RX	Receiving Antennas
SU-MIMO	Single-user MIMO
TB	Transport Block
TX	Transmitting Antennas
UE	User Equipment

VR	Virtual Reality
XR	Extended Reality

REFERENCES

[1] Cisco Annual Internet Report: Trends, 2018-2023, White Paper C11-741490 [accessed on 18 September 2022]; Available Online: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[2] Ericsson Cellular IoT Evolution for Industry Digitalization: White Paper [accessed on 18 September 2022]; Available Online: https://www.ericsson.com/en/reports-and-papers/white-papers/cellular-iot-evolution-for-industry-digitalization.

[3] Ericsson Enabling time-critical applications over 5G with rate adaptation: White Paper [accessed on 18 September 2022]; Available Online: https://www.ericsson.com/en/reports-and-papers/white-papers/enabling-time-critical-applications-over-5g-with-rate-adaptation.

[4] Ericsson Mobility Report 2020 [accessed on 18 September 2022]; Available Online: https://www.ericsson.com/4adc87/assets/local/reports-papers/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf.

[5] Ericsson Technology Review: XR and 5G, Extended reality at scale with time-critical communication [accessed on 18 September 2022]; Available Online: https://www.ericsson.com/4a492d/assets/local/reports-papers/ericsson-technology-review/docs/2021/xr-and-5g-extended-reality-at-scale-with-time-critical-communication.pdf.

[6] Extended Reality (XR) in 5G (Release 16) TR 26.928, V16.1.0, 3GPP Technical Specification Group Services and System Aspects, December 23, 2020, Available Online: https://www.3gpp.org/ftp/Specs/archive/26_series/26.928/26928-g10.zip.

[7] W. Xiang, K. Zheng, and X.S. Shen, 5G Mobile Communications. Cham: Springer International Publishing, 2018.

[8] O. A. Amodu and M. Othman, “Machine-to-machine communication: An overview of opportunities,” Computer Networks, vol. 145, pp. 255–276, 2018.

[9] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," in IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2347-2376, Fourthquarter 2015, doi: 10.1109/COMST.2015.2444095.

[10] G. Marques, R. Pitarma, N. M. Garcia, and N. Pombo, “Internet of things architectures, technologies, applications, challenges, and future directions for enhanced living environments and healthcare systems: A Review,” Electronics, vol. 8, no. 10, p. 1081, 2019.

[11] F. Al-Turjman, H. Zahmatkesh, and R. Shahroze, “An overview of security and privacy in Smart Cities' IOT Communications,” Transactions on Emerging Telecommunications Technologies, vol. 33, no. 3, 2019.

[12] S. Zhang, X. Xu, Y. Wu and L. Lu, "5G: Towards energy-efficient, low-latency and high-reliable communications networks," 2014 IEEE International Conference on Communication Systems, 2014, pp. 197-201, doi: 10.1109/ICCS.2014.7024793.

[13] E. Borgia, “The internet of things vision: Key features, applications and open issues,” Computer Communications, vol. 54, pp. 1–31, 2014.

[14] M. Agiwal, A. Roy and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1617-1655, thirdquarter 2016, doi: 10.1109/COMST.2016.2532458.

[15] R. Chataut and R. Akl, “Massive MIMO systems for 5G and beyond networks—overview, recent trends, challenges, and future research direction,” Sensors, vol. 20, no. 10, p. 2753, 2020.

[16] X. Wu, N. C. Beaulieu and D. Liu, "On Favorable Propagation in Massive MIMO Systems and Different Antenna Configurations," in IEEE Access, vol. 5, pp. 5578-5593, 2017, doi: 10.1109/ACCESS.2017.2695007.

[17] F. Rusek et al., "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," in IEEE Signal Processing Magazine, vol. 30, no. 1, pp. 40-60, Jan. 2013, doi: 10.1109/MSP.2011.2178495.

[18] S. Dixit and H. Katiyar, "Performance of OFDM in Time Selective Multipath Fading Channel in 4G Systems," 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015, pp. 421-424, doi: 10.1109/CSNT.2015.107.

[19] W. Ajib and D. Haccoun, "An overview of scheduling algorithms in MIMO-based fourth-generation wireless systems," in IEEE Network, vol. 19, no. 5, pp. 43-48, Sept.-Oct. 2005, doi: 10.1109/MNET.2005.1509951.

[20] E. Björnson, E. G. Larsson and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," in IEEE Communications Magazine, vol. 54, no. 2, pp. 114-123, February 2016, doi: 10.1109/MCOM.2016.7402270.

[21] N. H. M. Adnan, I. M. Rafiqul and A. H. M. Z. Alam, "Massive MIMO for Fifth Generation (5G): Opportunities and Challenges," 2016 International Conference on Computer and Communication Engineering (ICCCE), 2016, pp. 47-52, doi: 10.1109/ICCCE.2016.23.

[22] "5G/NR - Massive MIMO", ShareTechnote. Available online: https://www.sharetechnote.com/html/5G/5G_MassiveMIMO_Motivation.html [Accessed: 28-Sep-2022].

[23] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst and M. Haardt, "An introduction to the multi-user MIMO downlink," in IEEE Communications Magazine, vol. 42, no. 10, pp. 60-67, Oct. 2004, doi: 10.1109/MCOM.2004.1341262.

[24] Physical channels and modulation (Release 16) TS 38.211, V16.10.0, 3GPP Technical Specification Technical Specification Group Radio Access Network, June 23, 2022, Available Online: https://www.3gpp.org/ftp/Specs/archive/38_series/38.211/38211-ga0.zip.

[25] H. Li, H. Zhang, D. Li, Y. Liu and A. Nallanathan, "Joint Antenna Selection and User Scheduling in Downlink Multi-User MIMO Systems," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 2018, pp. 1072-1076, doi: 10.1109/CompComm.2018.8780593.

[26] Y. -X. Zhu, D. -Y. Kim and J. -W. Lee, "Joint Antenna and User Scheduling in the Massive MIMO System Over Time-Varying Fading Channels," in IEEE Access, vol. 9, pp. 92431-92445, 2021, doi: 10.1109/ACCESS.2021.3092754.

[27] G. Bu and J. Jiang, "Reinforcement Learning-Based User Scheduling and Resource Allocation for Massive MU-MIMO System," 2019 IEEE/CIC International Conference on Communications in China (ICCC), 2019, pp. 641-646, doi: 10.1109/ICCChina.2019.8855949.

[28] S. -C. Tseng, Z. -W. Liu, Y. -C. Chou and C. -W. Huang, "Radio Resource Scheduling for 5G NR via Deep Deterministic Policy Gradient," 2019 IEEE International Conference on Communications Workshops (ICC Workshops), 2019, pp. 1-6, doi: 10.1109/ICCW.2019.8757174.

[29] X. Guo et al., "A Novel User Selection Massive MIMO Scheduling Algorithm via Real Time DDPG," GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020, pp. 1-6, doi: 10.1109/GLOBECOM42002.2020.9322383.

[30] H. Chen et al., "Joint User Scheduling and Transmit Precoder Selection Based on DDPG for Uplink Multi-User MIMO Systems," 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), 2021, pp. 1-5, doi: 10.1109/VTC2021-Fall52928.2021.9625046.

[31] System architecture for the 5G System (Release 17) TS 23.501, V17.5.0, 3GPP Technical Specification Technical Specification Group Services and System Aspects, June 15, 2022, Available Online: https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/23501-h50.zip.

[32] Study on XR (Extended Reality) Evaluations for NR (Release 17) TR 38.838, V17.0.0, 3GPP Technical Specification Technical Specification Group Radio Access Network, January 4, 2022, Available Online: https://www.3gpp.org/ftp/Specs/archive/38_series/38.838/38838-h00.zip.

[33] "Augmented reality headsets market: Industry Report, 2019-2025," Augmented Reality Headsets Market | Industry Report, 2019-2025. Available Online: https://www.grandviewresearch.com/industry-analysis/augmented-reality-ar-headsets-market [Accessed: 28-Sep-2022].

[34] Physical layer procedures (Release 16) TS 36.213, V16.8.0, 3GPP Technical Specification Group Radio Access Network, January 5, 2022, Available Online: https://www.3gpp.org/ftp/Specs/archive/36_series/36.213/36213-g80.zip.

[35] A. Kosta, N. Pappas, V. Angelakis et al., "Age of information: A new concept metric and tool", Foundations and Trends® in Networking, vol. 12, no. 3, pp. 162-259, 2017.

[36] S. Kaul, M. Gruteser, V. Rai and J. Kenney, "Minimizing age of information in vehicular networks," 2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, 2011, pp. 350-358, doi: 10.1109/SAHCN.2011.5984917.

[37] Y. Gu, H. Chen, C. Zhai, Y. Li and B. Vucetic, "Minimizing Age of Information in Cognitive Radio-Based IoT Systems: Underlay or Overlay?," in IEEE Internet of Things Journal, vol. 6, no. 6, pp. 10273-10288, Dec. 2019, doi: 10.1109/JIOT.2019.2937334.

[38] Y. Gu, H. Chen, Y. Zhou, Y. Li and B. Vucetic, "Timely Status Update in Internet of Things Monitoring Systems: An Age-Energy Tradeoff," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 5324-5335, June 2019, doi: 10.1109/JIOT.2019.2900528.

[39] Study on channel model for frequencies from 0.5 to 100 GHz (Release 16) 3GPP TR 38.901 V16.1.0, 3GPP Technical Specification Group Radio Access Network, January 11, 2020, Available Online: https://www.3gpp.org/ftp//Specs/archive/38_series/38.901/38901-g10.zip.