© Universiti Tun Hussein Onn Malaysia Publisher's Office

**IJIE**

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

# Reduction of HARQ Latency for URLLC 5G Services Based on Network Slicing and Massive MIMO Hybrid Beamforming

**Mohammad Emad Alolaby[1*], Rudzidatul Dziyauddin[2], Liza A. Latiff [2]**

[1]MTN Syria,
P.O.Box 34474, Mazzeh, Damascus, SYRIA

[2] Razak Faculty of Technology and Informatics,
7 UTM, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2020.12.07.004
Received 30 March 2020; Accepted 3 August 2020; Available online 30 August 2020

**Abstract:** Ultra-Reliable and Low-Latency Communications (URLLC) is one of the three generic 5G services and probably the most challenging one, with strict quality of service requirements of 99.999% or more reliability and less than 1 milliseconds (ms) radio latency. To achieve latency targets, contributors to latency need to be addressed. Hybrid automatic repeat request (HARQ) retransmissions are major contributor to latency and need to be limited. The objective of this paper is to study the benefit of using Massive MIMO (M-MIMIO) along with radio network slicing to reduce number of HARQ retransmissions. A practical type of M-MIMO beamforming named hybrid beamforming is used. The performance of the proposed system is evaluated with slicing, without slicing and by alternating number of data streams per user. This work highlights the importance of technology enablers, such as M-MIMO and network slicing, in addressing quality-of-service (QoS) latency requirements for URLLC applications.

**Keywords:** Ultra-reliable and low-latency communications (URLLC), network slicing, scheduling, massive MIMO (M-MIMO), latency, reliability.

## 1. Introduction

5G should be the infrastructure that helps in mobilizing humans life and facilitating the digital transfer by enhancing user experience and empowering industries with information and communications technology (ICT), and the Internet of Things (IoT). So 5G vision can be summarized as follows: access to information and sharing of data are possible anywhere and anytime to anyone and anything. With 5G, communications are not limited anymore to humans, rather being expanded to cover machines. This will lead to unimaginable social and economic changes, including improvements in sustainability, productivity, entertainment and welfare. 5G services can be categorized in three generic ones: extreme or enhanced mobile broadband (xMBB) or (eMBB), massive machine type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) also known as ultra-reliable machine-type communications (uMTC) or critical machine-type communications (cMTC) [1]. Each of these generic 5G services has a different set of requirements. Among the three categories, URLLC allows new applications do not exist before. Driven by its targeted high reliability and very low latency, URLLC will enable applications such as, industrial automation, remote surgery, and self-driving cars. URLLC requirements are 1-10 ms for end-to-end latency (<1 ms radio latency) and 99.999% for reliability [1], [2].

There are multiple contributors to the latency of current generation of mobile systems, such as HARQ retransmissions, current transmission time interval (TTI) lengths, channel coding, legacy scheduling and network

*Corresponding author: iolaby@mtn.com.sy

architecture. These contributors are detailed in [3]. Several enablers and techniques have been proposed in literature to reduce latency. In [4], opportunistic distributed multiuser scheduling in the presence of a fixed packet deadline delay constraint is addressed. They proposed a scheduling mechanism that utilizes channel gain and buffering time to make scheduling decisions with an objective to optimize power consumption while satisfying delay constraint. The paper does not directly address 5G latency challenge of 1 ms, instead it assumes packet deadline of 100ms and simply gives priority to packets with shorter time-to-live. Some researchers, and in order to reduce HARQ latency for delay sensitive applications, suggested having a pre-scheduled resource for retransmission which is shared by a group of UEs as considered in [5]. Hence, the control signaling that is used to re-schedule the transmission when it does not succeed can be eliminated and that will be reflected in less latency. Moreover, by using this methodology, they claim major saving in capacity (up to 28%) due to the benefit of sharing of the reserved resources. Their analysis shows that with the right dimensioning of groups, the probability of dispute on the shared retransmission resources is adequately low. Leading to achieving the final error probability without re-scheduling procedures. Accordingly, in this study they showed how to minimize the time and resources needed for retransmissions, however it would have been more useful if they integrated this approach with mechanisms to reduce retransmission as a first step. A more comprehensive study is done in [6], which focused on limiting the number of HARQ retransmissions to reduce delay utilizing M-MIMO. However, this study was limited to a basic setup (two users, QPSK, no precoding), and it would have been more interesting if it used a realistic multi-user MIMO-OFDM system with a definite beamforming mechanism, and if it adopted a scheduling algorithm for further enhancement. Other researches, such as [7], focused on physical layer mechanisms. They started from the air interface of Long Term Evolution (LTE) networks, and suggested several modifications to allow lower delays and higher reliability for URLLC applications. They suggested reduced TTI length, shorter OFDM symbol durations, usage of convolutional codes instead of turbo codes, high diversity levels and physical channels design that enables early channel estimation and reliable transmission. They proved that it is possible to have an OFDM based 5G radio interface that satisfies the requirements of delay sensitive and highly reliable applications, such as URLLC. The study is one of the best studies in this domain. It would have been more comprehensive if the author had considered other schemes, such as cross layer scheduling and other physical layer mechanisms such as M-MIMO. In [8], due to the nature of machine type communications (MTC) data volume (including URLLC), they suggest that it is mandatory to do traffic aggregation of multiple packets over one resource block. This accordingly needs relay node for aggregation (RN). Also, It is suggested that priority queuing (PQ) is the best for delay sensitive applications, combined with proper network slicing based on traffic types. The idea of aggregation is attractive in terms of optimal resource utilization, nevertheless it imposes challenge on latency resulted from additional intermediate nodes along with additional signaling. This makes it more convenient for applications with moderate latency restrictions such as mMTC, rather than those with strict latency constrains such as URLLC. In [9], it is proposed to have a proactive scheduling method to minimize the delay (PDMS). The authors designed a centralized context-aware scheduler to minimize, as a primary objective, the number of dropped packets due to missed deadline and accordingly minimize average latency. PDMS scheduler superiority over non-deadline-aware scheduler on the level of dropped packets' number, especially in high system loads, was shown. In addition, it was demonstrated that the average packet delay can be reduced by this scheduler by more than 50%, given the knowledge of channel parameters. However, this paper does not explain how to predict channel properties in the future. Also, the work would have been more comprehensive, if physical layer mechanisms had been considered. Consequently, multiple techniques are suggested in literature as enablers to attain URLLC latency objective. A summary of some of these enablers along with their prospected impact are summarized in [1].

The main aim of this study is to develop a system that addresses URLLC QoS needs, of reduced latency by minimizing HARQ retransmissions. Utilizing M-MIMO and network slicing, the improvement that can be achieved on reliability and consequently on reduced number of HARQ retransmissions is studied. Hybrid beamforming is used for MIMIO. Slicing is used for better QoS differentiation and it is realized by scheduling part of the resources for URLLC users. The performance of the proposed system is evaluated in terms of bit error rate (BER) and mean number of HARQ retransmissions, with slicing and without slicing, and by alternating number of data streams per user.

The remainder of this paper is structured as follows. Section 2 introduces HARQ delay, M-MIMO hybrid beamforming, and network slicing concepts. Section 3 provides design considerations for the proposed ultra-reliable and low-latency radio system. Simulation parameters and assumptions are detailed in Section 4. In Section 5, the system-level simulation results on reduced number of HARQ retransmissions are presented. Finally, the main conclusions are drawn in Section 6.

## 2. HARQ delay, M-MIMO hybrid beamforming, and network slicing

### 2.1 HARQ delay

Transmissions over wireless channels is challenging because it is subject to errors, mainly due to fluctuations in received signal. These variations can be partially overcome before data transmission through link adaptation and scheduling. However, due to the random nature of the variations in the radio-link quality resulted mainly from Rayleigh fading, interference, and receiver noise, perfect adaptation is not feasible. Therefore, all wireless systems utilize a sort of forward error correction (FEC). Briefly, FEC coding relies on adding redundancy in the transmitted signal. This is done by adding additional bits, named as parity bits, to the information bits before transmission. Another method to deal with transmissions errors is to use automatic repeat request (ARQ). In this scheme, the receiver uses an error-detection code, usually a cyclic redundancy check (CRC), to detect the integrity of the received packet. If no error is detected, the received data is declared error-free and the transmitter is notified by sending a positive acknowledgement (ACK). Otherwise, if an error is detected, the receiver discards the received data and notifies the transmitter via a feedback channel by sending a negative acknowledgement (NAK), so that the transmitter retransmits the same packet again. In principle, all modern communication systems use hybrid ARQ (HARQ) which is a combination of forward error-correction coding and ARQ. Most practical HARQ schemes are built around a CRC code for error detection and convolutional or turbo codes for error correction, though any error-detection and error correction can be used. So, scheduling, link adaptation and HARQ are mechanisms to overcome instantaneous variations in radio-link quality and complement each other. While scheduling and link adaptation work before the transmission of data, HARQ works after transmission [10].

Despite its importance for reliability, HARQ imposes a major challenge considering 5G latency requirements. Actually, HARQ is a major contributor to the end to end delay. The LTE U-plane latency for a scheduled user equipment (UE) consists of the fixed node processing delays in UE and eNB (which includes radio frame alignment ~0.5 ms [11]) and 1ms TTI duration. Fig. 1 – **LTE FDD user-plane delay components** shows delay components for LTE Frequency-Division Duplexing (FDD) in user-plane, which can be formulated in milliseconds (ms) as [12]:

$$T(n)\ [ms] = 1.5 + 1 + 1.5 + 8 \times n = 4 + 8 \times n \quad (1)$$

where $n$ is the total number of HARQ retransmissions.

Table 1 shows the U-plane latency when HARQ is operated at an initial transmission error probability of 0 and 0.1, respectively [12]. This proves how important it is to strictly reduce the number of retransmissions to minimize latency. Moreover, smaller number of HARQ retransmissions leads to less processing time lost in decoding.

**Table 1 - U-plane latency analysis (estimated average)**

| **Description** | **Value (0% HARQ) ms** | **Value (10% HARQ) ms** |
|---|---|---|
| UE Processing Delay | 1.5 | 1.5 |
| TTI for UL data packet (Piggy back scheduling information) | 1 | 1 |
| HARQ Retransmission | 0 | 0.1*8 |
| eNB Processing Delay (including radio frame alignment) | 1.5 | 1.5 |
| ***Total delay*** | ***4*** | ***4.8*** |

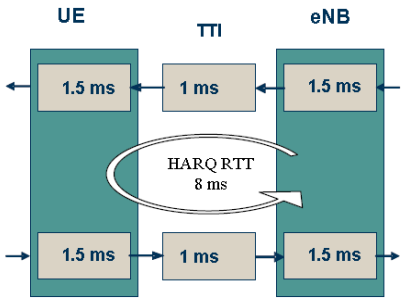


**Fig. 1 – LTE FDD user-plane delay components [12]**

## 2.2 M-MIMO hybrid beamforming

In the last few years, M-MIMO with its beamforming capability has proved to be a promising technology to improve wireless system performance, especially in capacity increase and interference reduction [6]. M-MIMO is simply traditional MIMO but with higher number of antenna elements. The number of base station (BS) antennas should be significantly larger than the number of users (usually minimum 8-10 fold). This allows the system to utilize beamforming to simultaneously communicate with multiple users utilizing same resource blocks (frequency-time resources). This has positive reflection on system performance and capacity. Several works have been done to illustrate M-MIMO benefits on capacity, spectral efficiency, and HARQ retransmissions minimization like [13], [14], [15]. M-MIMO relies on the law of large numbers and beam-forming in order to avoid fading dips and their bad impact [16]. It has the vital benefit of simplified signal processing because it creates "channel hardening" such that impact of small-scale fading is significantly suppressed due to the large beamforming gain [17].

5G systems relies on millimeter wave (mmWave) bands to benefit from their wide bandwidth. To overcome mmWave severe propagation losses, 5G systems utilize M-MIMO. The smaller wavelength of mmWave allows the usage of very large arrays at the BS, which enhances the system performance especially in terms of spectral efficiency [13], [18]. To get targeted improvement in performance from M-MIMO, there is need for perfect channel state information (CSI) at the BS. This is a challenging factor, as it needs major overhead especially when talking about mmWave frequencies, which have very short coherence time compared to legacy radio channels. Practically, in time division duplexing (TDD) systems, CSI is calculated from uplink training according to channel reciprocity [13]. Theoretically, a training sequence needs to be sent on each channel, that is, on each transmit – receive antenna pair. It is even worse in frequency division duplexing (FDD) systems, due to the fact that these systems cannot rely on channel reciprocity. In these systems, there is need for extensive overhead for downlink training and uplink feedback. In addition, hardware is another challenging factor to be considered. mmWave transceivers are quite costly with high power consumption and accordingly it is not practical to have a complete radio frequency (RF) chain for each antenna element [19], [20]. These two challenging factors triggered the research for agile methodologies such as the ones in [17], [21]. One of the most promising methodologies is to use hybrid digital/analog (HDA) beamforming structure first proposed in [22]. In this approach, hybrid transceivers use a combination of analog beamformers in the radio frequency (RF) and digital beamformers in the baseband domains, with fewer RF chains than the number of transmit elements. To realize HDA, a scheme called "Joint Spatial Division Multiplexing" (JSDM) was proposed in [23]. The first-stage, which is analog beamforming, is based on the slowly-varying second order channel statistics. This reduces the number of channels to be estimated using sounding process (downlink training reference symbol and up link feedback). In order to reduce sounding more, user equipment (UEs) are grouped in "virtual sectors" based on similar transmit channel covariance and intergroup interference is eliminated by analog precoding using block diagonalization (BD). With this virtual sectorization, downlink training can be done in different virtual sectors in parallel, and each UE only needs to feedback the intra-group channels, leading to the reduction of needed sounding overhead. The reduction is proportional to the number of virtual sectors [19].

## 2.3 Network slicing

Utilizing an architectural approach named network slicing, 5G networks will be able to offer different sub-networks for different services with competing performance requirements. Slicing addresses precise quality-of-service (QoS) requirements to each use case. Network slicing, is still in its early stages of development. It is defined in the 3GPP Release 15 specifications [24]. Further enhancements on the core network slicing will be in Release 16, and RAN slicing is targeted for the Release 17 specifications. Slices are virtual and isolated version of the network. Each user has an access to the slice subscribed based on his requirements, while other slices are not accessible for him [25]. 3GPP TS.23.501 [24] has identified four standardized Slice/Service Types (SSTs) listed in Table 2.

**Table 2 - Slice/Service Types (SSTs)**

| Slice/Service type | SST value |
|---|---|
| eMBB | 1 |
| URLLC | 2 |
| MIoT (mMTC) | 3 |

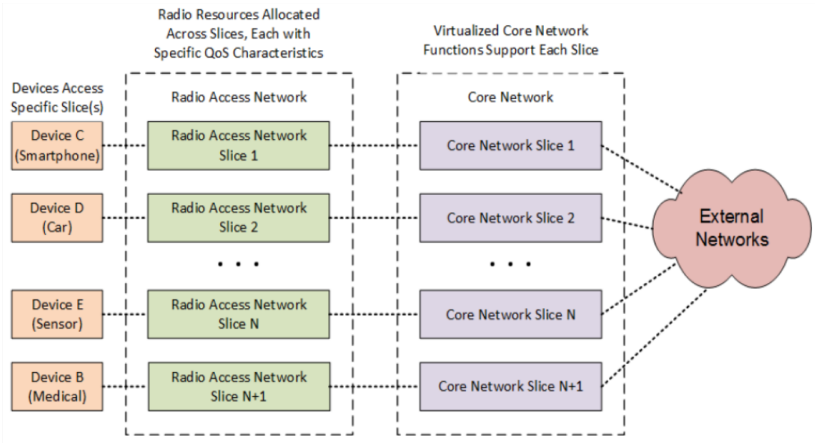


**Fig. 2 - Network slicing architecture [25]**

Fig. 2 shows network slicing architecture. Different users have access only to the slice they are subscribed to. Based on QoS needs, each slice is allocated with suitable radio resources. In the core network, radio access network slices is supported by virtualized core network functions, which connect different slices to external networks.

## 3. Proposed system design

The main aim of this study is to develop a system that addresses URLLC QoS needs, of reduced latency by minimizing HARQ retransmissions. We simulate a multi-user M-MIMO-OFDM system with Hybrid beamforming on the downlink direction. The users are mixture of URLLC users and non-URLLC ones. A network slicing is done by scheduling part of the carriers (network slice) to URLLC users while keeping non-URLLC users sharing all remaining subcarriers simultaneously. Both beamforming and slicing are expected to increase reliability, which leads to reduced HARQ retransmissions, and consequently reduced latency. Proposed system design is detailed in the following paragraphs.

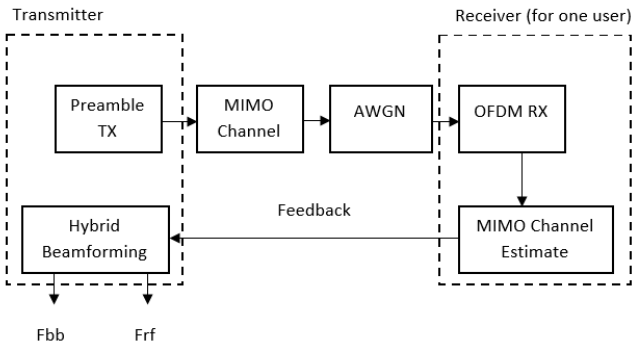


**Fig. 3 - Channel sounding**

Having good information about the channel at the transmitter end is mandatory in spatially multiplexed system in order to allow precoding. Precoding is needed to maximize the signal energy in the direction and channel of interest. Assuming a slowly varying channel, this is done by what so called sounding process. In this process, the BS transmits reference symbol. The Mobile Station (MS) uses the received reference signal to estimate the channel. Then, the MS transmits the channel estimation back to the BS, which uses it to calculate precoding weights needed for user data

transmission. In hybrid beamforming we have two kinds of precoding weights. Digital baseband $F_{bb}$ precoding weights and Radio frequency (RF) analog $F_{rf}$ precoding ones. Fig. 3 shows the processing of the channel sounding. In our simulator, reference symbols are sent over all BS antenna elements. Then each user receiver processes them. The receiver does amplification, OFDM demodulation and channel estimation for all transmitted reference symbols. Then it sends the feedback to the BS. For simplicity, we assume perfect feedback.

As for hybrid beamforming, the simulator uses Joint Spatial Division Multiplexing (JSDM) technique [19], [23] to calculate digital baseband and Radio Frequency (RF) analog precoding weights, $F_{bb}$ and $F_{rf}$ respectively. JSDM divides users with similar transmit channel covariance in different groups and eliminates the inter-group interference by an analog precoder utilizing the block diagonalization scheme [26]. In the simulator, the analog weights, $F_{rf}$, are the averaged weights over all subcarriers. Each data stream (one user can have multiple streams) maps to an individual RF chain, which is in turn connected to all antenna elements as shown in Fig. 4.

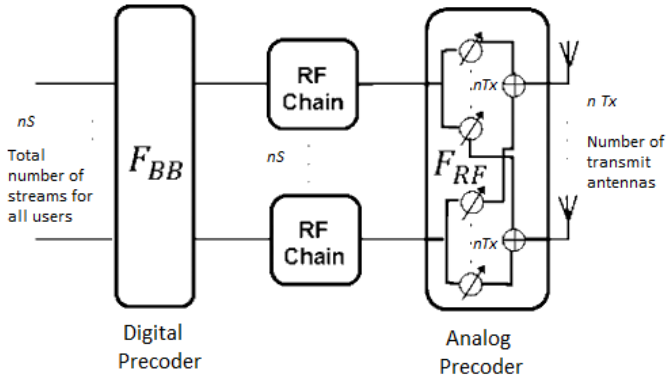


**Fig. 4 - Hybrid beamformer architecture**

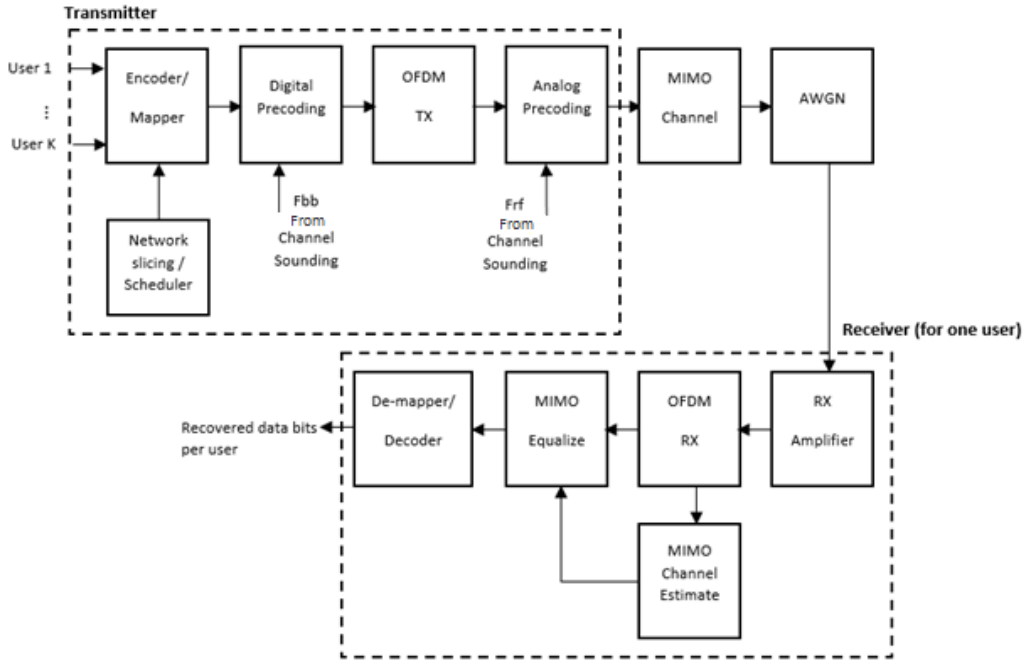


**Fig. 5 - System model for transmission and reception**

Suggested system model for data transmission and reception is shown in Fig. 5. In the transmitter end, the processing includes channel coding, QAM mapping, baseband precoding, OFDM modulation and RF analog beamforming. Convolutional coding is used rather than iterative ones, such as turbo codes, because convolutional

coding is more suitable for URLLC applications. The reason behinds that is because the performance of convolutional codes is round about level with iterative ones for short messages of a few hundred bits scale. Moreover, iterative codes may have an error floor which prevents them from achieving very low error rates (e.g. $10^{-9}$ ). On the contrary, convolutional codes do not have such an error floor and the receiver enjoys more simplicity. Furthermore, in convolutional coding the receiver can start data decoding once it is being received; to the contrary, iterative codes would decode data blocks upon completely receiving it. Thus, convolutional coding enables shorter receiver processing periods and early on-the-fly channel decoding. Therefore, reference symbols (RS) should be placed at the very beginning of a TTI. This enables having channel estimation immediately after receiving the first symbols in a TTI and then start decoding the code words, assuming that the channel will not variate significantly under very short TTI [2]. The simulator facilitates network slicing by assigning radio resources (slice) for each category of users, in our case URLLC & non-URLLC. In details, this is done by scheduling URLLC users on dedicated subcarriers. The simulator allows the choice of dedicated subcarriers for each URLLC user while remaining non-URLLC users are not scheduled and they occupy remaining subcarriers simultaneously while beamforming will take care of interference. Allocation is done either persistently (statistically) by assigning fixed arbitrary subcarriers for URLLC users, or in a dynamic manner by reviewing the assignment periodically based on subcarriers' channel conditions for each URLLC user.

The signal is then propagated over spatial MIMO. The scattering model uses a single-bounce ray tracing approximation. Scatterers are placed randomly around the receiver.

The receiver per user, starts by amplifying the signal to compensate path loss. Then it does OFDM demodulation, MIMO equalization, QAM de-mapping, channel decoding and Hybrid ARQ (HARQ) (not shown on system model for simplicity).

To define the probability of *i* retransmissions, we first define probability of accepted packet after decoding *Pacc*, that is, all its bits are received without error:

$$P_{acc} = (1 - P_b)^s \tag{2}$$

Where $P_b$ is the bit error probability. BER is quite close to $P_b$ for large number of simulation bits. *s* is the packet size. Accordingly, the probability of *i* retransmissions could be defined as:

$$P_i = P_{acc}(1 - P_{acc})^i \tag{3}$$

## 4. Simulation Parameters

Simulation studies are carried out using Matlab in outdoor scattering environment with 100 scatterers. The chosen frequency is 28MHz, which is one of the promising mmWave bands for 5G. There is one BS positioned at the side of the area with four users distributed randomly within 1000 m range of the BS. ETSI in its 5G new radio (5G NR) technical specifications: 5G NR; Physical channels and modulation (TS 138 211), specifies different transmission numerologies [27]. Transmission numerology μ = 4 with 240Khz subcarrier spacing and normal prefix is used in the simulator. This numerology corresponds to smallest access slot (TTI), which is desirable for low latency applications [2].

Each user can have multiple data streams to enhance system capacity. We assume number of receive antennas per user to be four times the number of data streams for the corresponding user. On Base station (BS) end, to satisfy M-MIMO condition to have large number of antennas compared to user number, it is assumed that the number of BS antennas is eight times the number of the total data streams of all users.

For MIMO channel, the scattering model uses 100 scatterers placed randomly around the receiver. We assume non-LOS environment and rectangular antenna array with isotropic antenna elements. Then, the transmitted signal is passed over AWGN channel.

As for packet size adopted in the simulation, it is defined as number of bits in a slot, that could be calculated as follows:

$$s = N_{symb}^{slot}.N_{SC}^{RB}.M_{RB}.Q_m.CR \tag{4}$$

Where $N_{symb}^{slot}$ is the number of symbols per slot which equals 14 in 5G NR, $N_{SC}^{RB}$ is the number of subcarriers per resource block which equals 12 in 5G NR, $M_{RB}$ is the number of resource blocks of the available downlink bandwidth. Based on subcarrier spacing configuration μ = 4, and channel bandwidth 50 MHz, this value is set to 17. *Qm* is the

modulation order. It is assumed to be 4 which corresponds to 16-QAM. *CR* is convolutional coding code rate. It is assumed to be 1/3 in the simulation. Consequently based on these values, adopted packet size (*s*) is 3808 bits.

Design parameters, as well as the evaluation assumptions are summarized in Table 3.

**Table 3 - Evaluation assumptions**

| | |
|---|---|
| **Carrier frequency** | 28 GHz |
| **Subcarrier spacing** | 240 kHz |
| **OFDM FFT length** | 256 subcarriers |
| **Cyclic prefix length** | 64 subcarriers |
| **Number of BS antennas** | Rectangular array 8 x Total number of user streams;<br>Element spacing: 0.5 lambda;<br>Antenna element: Isotropic |
| **Number of device (User) antennas** | 4 x number of corresponding user's streams; Antenna element: Isotropic |
| **Number of users** | 4 |
| **User Locations** | Random within 1000 meters of BS, specified as azimuth / elevation angles referenced to BS:<br>- Azimuth in range [-180 180];<br>- Elevation in range [-90 90]. |
| **Environment** | Outdoor |
| **Subcarrier modulation** | 16 QAM |
| **Channel** | Single-bounce ray tracing model;<br>100 scatterers. Scatterers are placed randomly within a sphere around the receiver |
| **Transmission numerology (μ)** | 4 |
| **Coding** | Convolutional coding with code rate 1/3 |

## 5. Results and Discussion

First, array response pattern resulted from beamforming is shown in Fig. 6. It is clear that the stronger lobes point at distinct users. These lobes show the amount of separation realized by beamforming.

The simulation results focus on BER and mean number of HARQ retransmissions when changing signal to noise ratio (SNR), and selection of subcarriers for network slicing.

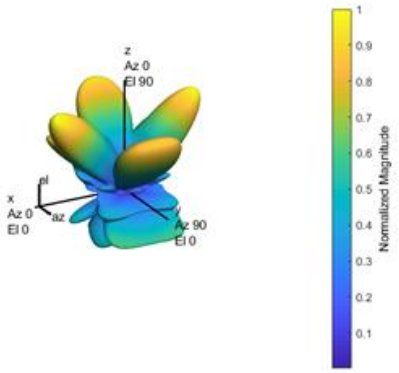


**Fig. 6 - 3D response pattern for one of the subcarriers (subcarrier 100 in this example)**

## 5.1 Impact of proposed scheduling on reliability and HARQ retransmissions

To understand the impact of proposed scheduling on reliability and accordingly on the reduction of HARQ retransmission, we first simulate the case without proposed scheduling. That is, no differentiation has been considered for URLLC user and all users are sharing all time/ frequency resources. Four users are considered in this case, while

user 1 is assumed to be URLLC user. The results of BER for different values of SNR have been shown in the dotted lines of Fig 7. The results indicate that hybrid beamforming helped in increasing system's overall capacity by allowing multiple users to share the same time/ frequency resources with acceptable level of reliability reflected in BER. Also, all users have almost same performance no matter their location and distance from Base Station (BS) is. These results are directly related to beamforming capability of maximizing the signal energy in the targeted direction and channel. However, this setup is still not suitable for URLLC users, who target higher level of reliability at lower SNRs.

Average number of HARQ retransmissions capped at a maximum of ten retransmissions is shown in the dotted lines of Fig. 8. Again, it is clear that unless for high SNR the users still suffer from considerable number of retransmissions, which make it not suitable for URLLC applications, where minimal number of retransmissions or no retransmissions at all are targeted.

Now, considering same setup, proposed network slicing is introduced by scheduling dedicated subcarriers to URLLC. First an arbitrary (random) resource block (12 subcarriers) is assigned in a persistent manner, later in the next section, best subcarriers in terms of channel gain will be evaluated. For the first case which is being analyzed in this section, first resource block is assumed to be assigned to URLLC user (User 1), while other non-URLLC users share all remaining subcarriers simultaneously. As shown in the solid lines of Fig. 7, reliability reflected in BER has been improved significantly for URLLC user.

Fig. 8 shows average number of retransmissions in solid lines. It is clear that retransmissions are almost eliminated for URLLC user even at low SNR values around 0 dB.

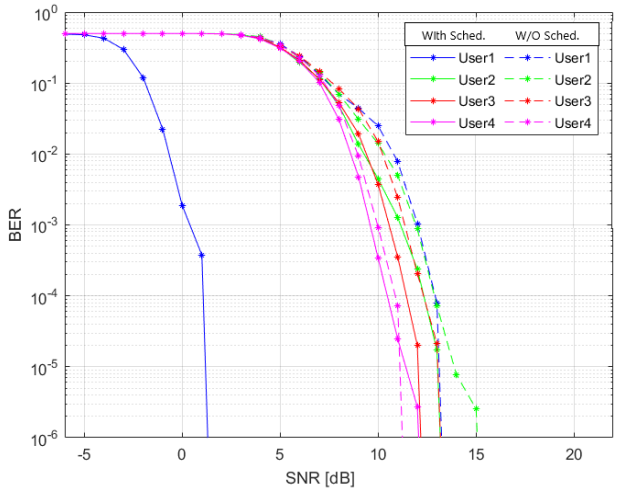


**Fig. 7 - BER In both cases, with proposed scheduling and without it**

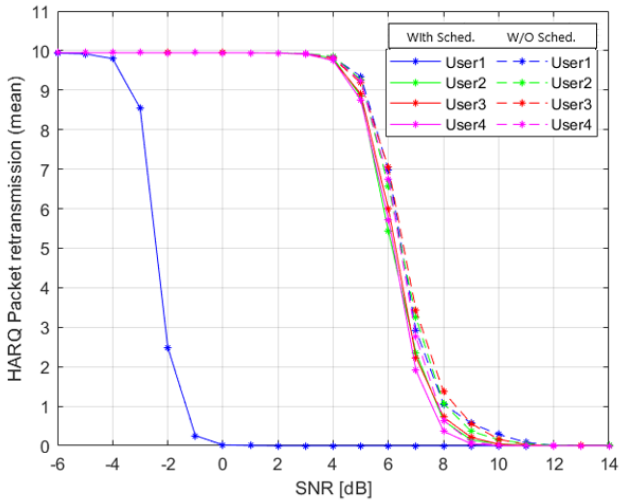


**Fig. 8 - Mean HARQ packet retransmission with and without scheduling**

## 5.2 Impact of subcarriers selection

To understand the impact of changing the subcarriers assigned to URLLC user, instead of assigning first resource block (12 sub-carriers) in a persistent manner, we dynamically assign the best twelve subcarriers for URLLC user. The

best subcarriers are those carriers who have the maximum mean channel gains over all channels related to the corresponding user, while a channel corresponds to single transmit antenna element and single receive antenna element pair. Fig. 9 and Fig. 10 compares BER and mean number of HARQ retransmissions (capped at ten retransmissions) for both cases, the assignment of best subcarriers versus the assignment of arbitrary subcarriers (persistent one). As inferred from the figures, there is some improvement in performance in the case of choosing best subcarriers. However, it is at the price of a higher computational and signaling overheads, which may not be tolerable in URLLC cases because it will lead to further delay by increasing the delay-to-access component of latency [3]. Accordingly, throughout the remaining of this paper, we will consider the usage of arbitrary assignment.

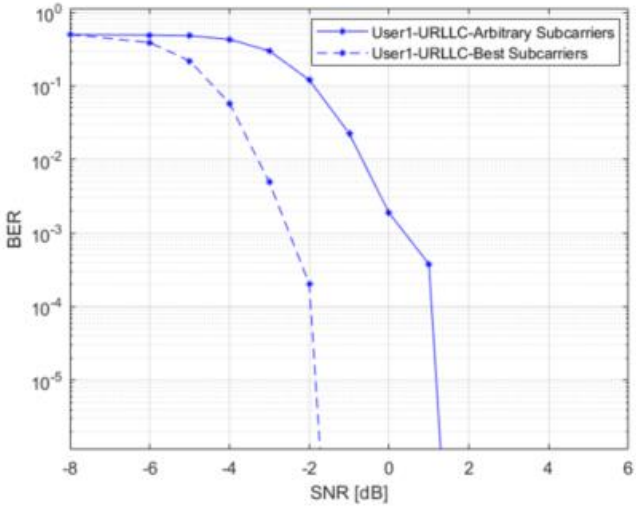


**Fig. 9 BER for both cases, assignment of best subcarriers versus the assignment of arbitrary subcarriers**

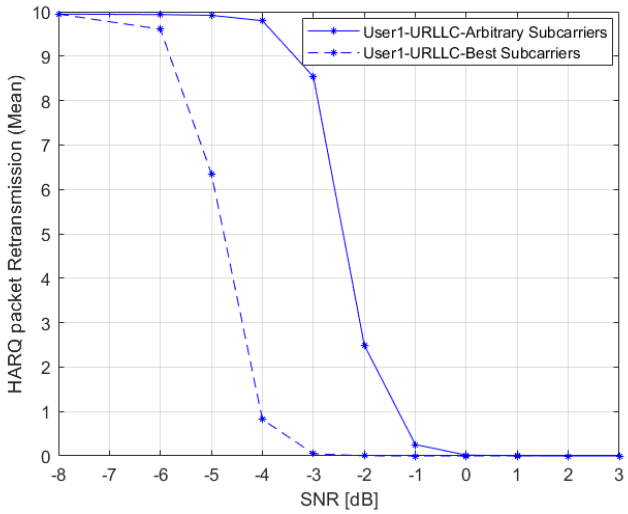


**Fig. 10 Mean HARQ packet retransmission for both cases, assignment of best subcarriers versus the assignment of arbitrary subcarriers**

## 5.3 Impact of changing number of streams per user

M-MIMO improves spectral efficiency in two forms: (i) a base station (BS) can communicate simultaneously with multiple user equipment (UEs) on the same time-frequency resources, (ii) multiple data streams can be sent between the BS and each UE. Focusing on the latter case, the effect of increasing number of data streams is simulated. The case of same four users while user 1 is URLLC and remaining users are non-URLLC is simulated after allowing each user to send two streams simultaneously. Fig. 11 shows the results comparing it to the case of one stream per user. For simplicity two users are shown. User 1 which is a URLLC user with proposed scheduling, and user 2 which is a non-URLLC user. The case of two streams is plotted in dotted lines, while the case of one stream is plotted in solid lines.

As shown in the figure there is major deterioration in BER in case we increase number of streams per user. This is logic as the user is interfering on itself. M-MIMO could alleviate the impact of this interference to a certain extent, but still it is major when talking about URLLC applications. Accordingly it is recommended not to push multiple streams for URLLC users.

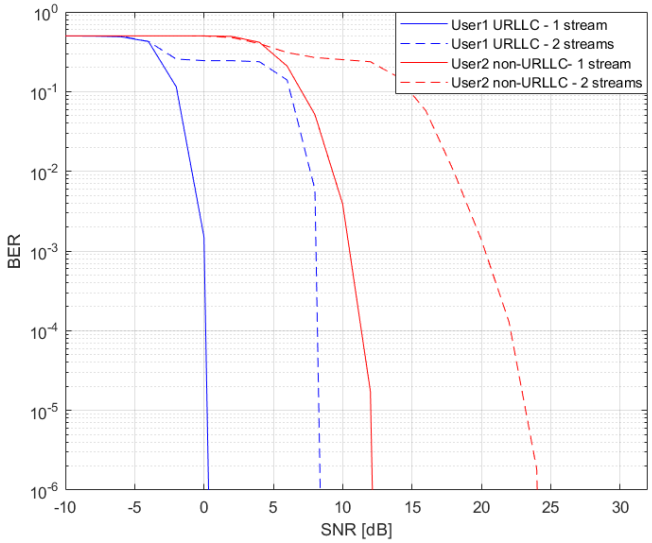


**Fig. 11 - BER for one and two streams per user, with proposed scheduling**

## 6. Conclusion

This paper addressed quality-of-service (QoS) latency requirements for URLLC applications, by studying the benefit of exploiting M-MIMO along with network slicing to reduce HARQ retransmissions and consequently latency. Hybrid beamforming is used in M-MIMO because of its advantages of less training overhead and more practical and simpler hardware compared to digital only beamformers. BER and mean number of retransmissions were simulated for different values of SNR. The results show that hybrid beamforming could accommodate all users on same frequency and time resources resulting in increase of system capacity with acceptable reliability reflected in acceptable level of BER. However, higher reliability is expected for URLLC applications as lower reliability leads to higher number of HARQ retransmissions and consequently higher latency. Accordingly, a slice of radio resources (resource blocks) was reserved for URLLC users through simple persistent scheduling leading for a major improvement in reliability reflected in lower BER and accordingly in lower retransmissions. HARQ Retransmissions are almost eliminated even at low SNR values. Further complication by scheduling on best subcarriers led to further enhancement on reliability. However, this is at the price of a higher computational and signalling overhead, which cannot be tolerated in URLLC applications. Also, the results show that having multiple streams per users is not feasible for URLLC applications because of the major increase in BER leading to higher number of HARQ retransmissions.
Based on this study, M-MIMO and network slicing are considered key technology enablers to attain latency requirements for URLLC applications. In the future, we plan to consider higher number of users, and work on machine learning for the latency improvement.

## Acknowledgment

This work is funded by Universiti Teknologi Malaysia (Grant No. Q.K130000.3556.05G14).

## References

[1] H. Tullberg, P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, et al. (2016). The METIS 5G System Concept: Meeting the 5G Requirements, IEEE Communications Magazine, vol. 54, pp. 132-139
[2] A. O. J. F. M. P. Marsch. (2016) 5G Mobile and Wireless Communications Technology vol. 1st. New York, NY, USA: Cambridge University Press
[3] X. Jiang, H. Shokri-Ghadikolaei, G. Fodor, E. Modiano, Z. Pang, M. Zorzi, et al. (2019). Low-Latency Networking: Where Latency Lurks and How to Tame It, Proceedings of the IEEE, vol. 107, pp. 280-306
[4] M. M. Butt, K. Kansanen, and R. R. Muller. (2011). Individual Packet Deadline Constrained Opportunistic Scheduling for a Multiuser System, IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1-5
[5] R. Abreu, P. Mogensen, and K. I. Pedersen, (2017). Pre-Scheduled Resources for Retransmissions in Ultra-Reliable and Low Latency Communications, IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-5
[6] N. Zarifeh, A. Kabbani, M. El-Absi, T. Kreul, and T. Kaiser. (2016). Massive MIMO exploitation to reduce HARQ delay in wireless communication system, IEEE Middle East Conference on Antennas and Propagation (MECAP), pp. 1-5

[7] O. N. C. Yilmaz, Y. P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs. (2015). Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case, IEEE International Conference on Communication Workshop (ICCW), pp. 1190-1195
[8] M. Dighriri, A. S. D. Alfoudi, G. M. Lee, T. Baker, and R. Pereira. (2017) Comparison Data Traffic Scheduling Techniques for Classifying QoS over 5G Mobile Networks, 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 492-497
[9] R. Holakouei and P. Marsch. (2015). Proactive Delay-Minimizing Scheduling for 5G Ultra Dense Deployments, IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), pp. 1-5
[10] S. P. Erik Dahlman, Johan Sköld. (2011). 4G LTE/LTE-Advanced for Mobile Broadband. UK: ELSEVIER
[11] 3GPP. (2009). Universal Mobile Telecommunications System (UMTS); LTE; Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN), Cedex - FRANCE
[12] 3GPP. (2014). LTE; Feasibility study for Further Advancements for E-UTRA (LTE-Advanced), Cedex - FRANCE.
[13] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta. (2014). Massive MIMO for next generation wireless systems. IEEE Communications Magazine, vol. 52, pp. 186-195
[14] T. Marzetta. (2014). MASSIVE MIMO AND BEYOND, Munich Workshop on Massive MIMO
[15] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang. (2014). An Overview of Massive MIMO: Benefits and Challenges, IEEE Journal of Selected Topics in Signal Processing, vol. 8, pp. 742-758
[16] E. G. L. Thomas L. Marzetta, Hong Yang, Hien Quoc Ngo. (2016). Fundamentals of Massive MIMO: Cambridge University Press
[17] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, et al. (2017). Hybrid Beamforming for Massive MIMO: A Survey. IEEE Communications Magazine, vol. 55, pp. 134-141
[18] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, et al. (2013). Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays. IEEE Signal Processing Magazine, vol. 30, pp. 40-60
[19] Z. Li, S. Han, and A. F. Molisch. (2016). Hybrid beamforming design for millimeter-wave multi-user massive MIMO downlink, IEEE International Conference on Communications (ICC), pp. 1-6
[20] L. A. Adeeb Salh, Nor Shahida M Shah, and Shipun A Hamzah. (2017). Adaptive Antenna Selection and Power Allocation in Downlink Massive MIMO Systems, International Journal of Electrical and Computer Engineering (IJECE), p. 3521-3528
[21] B. M. S. Yasmine M. Tabra. (2019). Hybrid MVDR-LMS beamforming for massive MIMO, Indonesian Journal of Electrical Engineering and Computer Science, pp. 715-723
[22] Z. Xinying, A. F. Molisch, and K. Sun-Yuan. (2005). Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection, IEEE Transactions on Signal Processing, vol. 53, pp. 4091-4103
[23] A. Adhikary, J. Nam, J. Ahn, and G. Caire. (2013). Joint Spatial Division and Multiplexing—The Large-Scale Array Regime, IEEE Transactions on Information Theory, vol. 59, pp. 6441-6463
[24] 3GPP. (2019). Technical Specification Group Services and System Aspects; System Architecture for the 5G System (5GS); (Release 15), in 3GPP TS 23.501 V15.7.0 vol. 3GPP TS 23.501 V15.7.0, ed. Valbonne - FRANCE: 3GPP
[25] Rysavy Research. (2019). Global 5G: Implications of a Transformational Technology. September
[26] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt. (2004). Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels, IEEE Transactions on Signal Processing, vol. 52, pp. 461-471
[27] ETSI. 5G NR. (2018). Physical channels and modulation. vol. 3GPP TS 38.211 version 15.2.0 Release 15, ed: ETSI