*Computers, Materials & Continua*
DOI: 10.32604/cmc.2022.024824
***Article***

Tech Science Press

# HARQ Optimization for PDCP Duplication-Based 5G URLLC Dual Connectivity

**Changsung Lee[1,3], Junsung Kim[2,3], Jaewook Jung[3], Jungsuk Baik[3] and Jong-Moon Chung[3,*]**

[1]Samsung Research, Seoul, 06765, Korea
[2]Samsung Electronics, Suwon, 16677, Korea
[3]School of Electrical and Electronic Engineering, Yonsei University, Seoul, 03722, Korea
*Corresponding Author: Jong-Moon Chung. Email: jmc@yonsei.ac.kr
Received: 01 November 2021; Accepted: 07 December 2021

**Abstract:** Packet duplication (PD) with dual connectivity (DC) was newly introduced in the 5G New Radio (NR) specifications to meet the stringent ultra reliable low latency communication (URLLC) requirements. PD technology uses duplicated packets in the packet data convergence protocol (PDCP) layer that are transmitted via two different access nodes (ANs) to the user equipment (UE) in order to enhance the reliability performance. However, PD can result in unnecessary retransmissions in the lower layers since the hybrid automatic retransmission request (HARQ) operation is unaware of the transmission success achieved through the alternate DC link to the UE. To overcome this issue, in this paper, a novel duplication-aware retransmission optimization (DRO) scheme is proposed to reduce the resource usage induced by unnecessary HARQ retransmissions. The proposed DRO scheme can minimize the average channel use while satisfying the URLLC requirements. The proposed DRO scheme derives the optimal HARQ retransmission attempts for different ANs by solving a nonlinear integer programming (NLIP) problem. The performance of the proposed DRO scheme was evaluated using MATLAB simulation and is compared to the existing 5G HARQ support schemes. The simulations results show that the proposed DRO scheme can provide a 14.71% and 15.11% reduced average channel use gain compared to the selective data duplication upon failure (SDUF) scheme and latency-aware dynamic multi-connectivity algorithm (LADMA) scheme, respectively, which are the existing 5G PD schemes that use HARQ.

**Keywords:** 5G networks; URLLC; dual connectivity; retransmission; packet duplication

## 1 Introduction

Ultra reliable low latency communication (URLLC) is one of the fifth generation (5G) mobile service modes that can support high reliability wireless networking with very low time delays. Although different design goals are required for different URLLC applications (e.g., autonomous driving,

drone control, factory automation, remote surgery), one of the stringent quality of service (QoS) requirements for URLLC traffic is the 99.999% reliability within 1 ms packet latency [1]. To support URLLC traffic in 5G networks, novel design concepts have been proposed in the physical (PHY) and medium access control (MAC) layers. For example, a short transmission time interval (sTTI) and faster processing speeds were applied to achieve the low latency requirements, and hybrid automatic repeat request (HARQ) was enhanced to achieve the high reliability requirements. However, in addition to the novel PHY and MAC designs, changes in the upper layers are also required to meet the stringent URLLC service requirements.

Dual connectivity (DC) was first introduced in the 3GPP Release 12 specifications in order to enable user equipment (UE) to access two different access nodes (ANs) simultaneously. The initial objective of DC was to achieve downlink (DL) data rate enhancement by splitting the user plane (UP) data, however, DC was recently extended to support reliability enhancements by exploiting packet duplication (PD). In the 3GPP Release 15 specifications, the PD functionality was adopted as a part of the packet data convergence protocol (PDCP) layer defined under the name PDCP duplication [2]. By duplicating URLLC packets in the PDCP layer and transmitting them through different ANs, an UE can obtain macro diversity. As a result, PD enables an UE to obtain not only higher transmission reliability, but reduced latency since it can alleviate the potential need for a retransmission.

However, PDCP duplication can give rise to unnecessary retransmissions if HARQ is operated in the lower layers. When duplicated packets are transmitted via different ANs, the packet delivered through one AN might be successful and the other packet sent through the other AN may be unsuccessful. In this case, the HARQ process will not be needed, but HARQ will be executed since the HARQ entity would be unaware of the successful packet delivery through the other AN. This redundant retransmission of HARQ will lead to a waste of radio resources. As the number of data traffic and connected devices in the mobile network increase, efficient management of radio resources become more essential in 5G networks [3,4]. In order to overcome this problem, a novel HARQ optimization scheme that cooperates with the PDCP duplication process is proposed. The original contributions of this paper are summarized as follows.

- To minimize redundant wireless resource usage in PDCP duplication, the duplication-aware retransmission optimization (DRO) scheme is proposed.
- The expected channel use model and the corresponding optimization statement is derived considering the HARQ latency and packet reliability model.
- The optimal number of HARQ retransmission attempts for different ANs is derived by solving a nonlinear integer programming (NLIP) problem based on a gradient descent method.
- The performance of DRO is compared to non-optimal retransmission control schemes such as selective data duplication upon failure (SDUF) and latency-aware dynamic multi-connectivity algorithm (LADMA), and the simulation results show that DRO can provide the lowest average channel use while satisfying the URLLC requirements.

## 2 Related Work

Several papers have focused on the reliability enhancement of PDCP duplication in 5G networks. In [5], the feasibility and effectiveness of PDCP duplication were investigated, where the numerical results show that PD can help satisfy the latency and reliability requirement of URLLC. In [6], a reliability-oriented multi-connectivity (MC) concept was proposed based on PDCP duplication, which includes a novel admission mechanism to control the number of MC users. In [7], an analytical study on the reliability of MC was conducted while considering HARQ retransmission. In [8], the LADMA

scheme was proposed to reduce the resource usage while satisfying the latency requirement, which activates MC only for users with a high latency violation probability. In [9], the SDUF scheme was proposed to reduce the radio resource consumption by activating DC only upon a transmission failure on the primary link. However, the number of HARQ retransmissions and flexible numerology which have an effect on the overall latency are not considered in the previous PDCP duplication works.

Other papers have studied on the HARQ retransmission aspect in 5G networks. These works focused on HARQ and efficent resource allocation for single connected users. In [10], an optimal resource allocation and repetition coding based HARQ retransmission scheme was proposed to minimize the necessary bandwidth required for URLLC traffic. In [11], a novel resource allocation scheme for retransmissions was proposed, where corresponding analytical models for loss rates were derived. In [12], a Pollaczek–Khinchine (P-K) formula based quadratic optimization (PFQO) scheme was proposed, which can minimize the URLLC bandwidth requirement based on the queueing analysis of HARQ.

## 3 System Model

Consider an UE which is connected to the master gNB (M-gNB) and the secondary gNB (S-gNB) simultaneously in the NR-NR DC (NR-DC) architecture [13], as shown in Fig. 1. The index $i \in \{1, 2\}$ denotes one of the dual-connected gNBs, where $i = 1$ represents the M-gNB and $i = 2$ represents the S-gNB. Incoming URLLC packets of the M-gNB are duplicated in the PDCP layer, and transmitted via the M-gNB and S-gNB. Let $\gamma_i$ be the signal-to-interference-plus-noise ratio (SINR) with respect to gNB $i$. For each gNB, incremental redundancy HARQ (IR-HARQ) is considered, where a packet is retransmitted if the UE fails to decode the received packet successfully. The index of transmission attempt for gNB $i$ is $m_i = 1, 2, \ldots, M_i$, where $M_i$ denotes the maximum number of transmission attempts for one packet. For example, $m_i = 1$ represents the initial transmission, and $m_i \geq 2$ represents a retransmission. For NR numerology, a flexible frame structure is considered. One slot consists of 14 orthogonal frequency division multiplexing (OFDM) symbols, where 1 OFDM symbol duration is $T_s = 2^{-\mu}/14$ ms and subcarrier spacing (SCS) is $\Delta f = 15 \times 2^{\mu}$ kHz ($\mu \in \{0, 1, 2, 3, 4\}$). In addition, mini-slots are considered for sTTIs, which can consist of 1∼13 OFDM symbols.

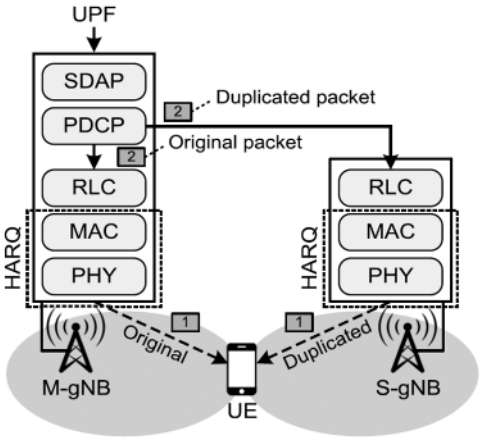


**Figure 1:** Example of the NR-DC architecture when packet duplication occurs in the PDCP layer

### 3.1 Packet Error Rate Model

A packet error occurs after a decoder fails to correct an erroneous packet, and thereby a retransmission of the erroneous packet is induced. Let $p_{i,m}$ be the packet error rate (PER) at the $m$-th transmission attempt, which can be modeled as

$$p_{i,m}(\gamma_i) = \begin{cases} 1 & \text{if } 0 \le \gamma_i < \gamma_{\text{th},m} \\ a_m \exp(-g_m \gamma_i) & \text{if } \gamma_i \ge \gamma_{\text{th},m} \end{cases} \tag{1}$$

where $a_{n,m}$ and $g_{n,m}$ are parameters of the $m$-th transmission attempt when the modulation and coding scheme (MCS) is $n$, and $\gamma_{\text{th},m}$ is the SINR threshold when the PER is 1 [14]. The example of MCS parameters is provided in Tab. 1 [15]. Consider that the channel is in the $n$-th state if the received SINR falls into the range of $[\Gamma_n, \Gamma_{n+1})$. Based on the adaptive modulation and coding (AMC) scheme, MCS $n$ is assigned for the $n$-th channel state, where the SINR threshold to the $n$-th channel state $\Gamma_n$ can be expressed as below [15].

$$\Gamma_n = \frac{1}{\sum_{m=1}^{M_i} g_{n,m}} \ln\left(\frac{\prod_{m=1}^{M_i} a_{n,m}}{P_{\text{target}}}\right) \tag{2}$$

**Table 1:** Example of MCS parameters

| | MCS 1 | MCS 2 | MCS 3 | MCS 4 | MCS 5 |
|---|---|---|---|---|---|
| Modulation | BPSK | QPSK | QPSK | 16QAM | 16QAM |
| Code rate | 1/2 | 2/3 | 5/6 | 2/3 | 5/6 |
| Bits/symbol | 1/2 | 4/3 | 5/3 | 8/3 | 10/3 |
| $a_1$ | 4447.4 | 2068.5 | 514.7 | 850.9 | 142.9 |
| $a_2$ | 2298.6 | 1344.3 | 3297.8 | 372.3 | 895.7 |
| $a_3$ | 5944.9 | 1428.4 | 7247.9 | 3567.3 | 1057.1 |
| $g_1$ | 11.104 | 3.315 | 1.759 | 0.816 | 0.339 |
| $g_2$. | 21.012 | 6.997 | 6.196 | 1.895 | 1.657 |
| $g_3$ | 34.203 | 10.569 | 10.224 | 3.378 | 2.706 |
| $\gamma_{p,1}$ | −1.212 | 3.623 | 5.502 | 9.172 | 11.660 |
| $\gamma_{p,2}$ | −4.337 | 127 | 1.164 | 4.946 | 6.131 |
| $\gamma_{p,3}$ | −5.950 | −1.629 | −0.608 | 3.841 | 4.105 |

### 3.2 Finite Block Length Model

Since URLLC packets are small in general, finite blocklength theory can be used to derive the amount of resource usage [16]. An URLLC packet of $L$ bits can be transmitted with the PER of $p_{i,m}$ based on channel use $r_i$ with a given SINR $\gamma_i$ as

$$L = r_i C_i - Q^{-1}(p_{i,m})\sqrt{r_i V(\gamma_i)} + O(\log_2(r_i)) \tag{3}$$

where $C_i = \log_2(1+\gamma_i)$ is the Shannon capacity based on the infinite blocklength and $V(\gamma_i) = (\log_2(e))^2(1-1/(1+\gamma_i)^2)$. As a result, the channel use $r_i$ can be approximated as in (4).

$$r_i = \frac{L}{C_i} + \frac{(Q^{-1}(p_{i,m}))^2 V(\gamma_i)}{2C_i^2}\left(1+\sqrt{1+\frac{4LC_i}{V(\gamma_i)(Q^{-1}(p_{i,m}))^2}}\right) \tag{4}$$

### 3.3 Packet Reliability Model

Since PDCP duplication is used in the NR-DC architecture, the reliability of the URLLC packet can be enhanced. Let $P_{\text{out},i}$. be the packet loss rate (PLR) of gNB $i$. Since a packet is considered to be lost after $M_i$ transmission attempts, the packet delivery failure probability $P_{\text{out},i}$ can be defined as in (5).

$$P_{\text{out},i} = \prod_{m=1}^{M_i} p_{i,m} \tag{5}$$

In the NR-DC architecture, the URLLC packet fails if both the original packet and the duplicated packet are lost at the M-gNB and S-gNB. Consequently, the packet reliability in the NR-DC architecture $R_{\text{DC}}$ can be modeled as in (6).

$$R_{\text{DC}} = 1 - P_{\text{out},1}P_{\text{out},2} \tag{6}$$

### 3.4 Latency Model

Let $T_{\text{one},i}$ be the one-shot transmission latency from gNB $i$, which is defined as the duration from the time that a packet arrives at the lower layer of gNB $i$, to the time that a packet is decoded at the UE. Therefore, $T_{\text{one},i}$ can be expressed as

$$T_{\text{one},i} = T_{\text{gNB\_proc}} + T_{\text{DL\_align}} + T_{\text{DL\_Tx},i} + T_{\text{UE\_proc}} \tag{7}$$

where $T_{\text{gNB\_proc}}$ is the gNB processing time (i.e., time to prepare transmission and make a scheduling decision) and $T_{\text{DL\_align}}$ is the packet alignment time (i.e., time until the next monitoring occasion (MO)). In addition, $T_{\text{DL\_Tx},i}$ is the packet transmission time and $T_{\text{UE\_proc}}$ is the UE processing time (i.e., time it takes to decode a received packet and prepare an ACK). Let $T_{\text{HARQ},i}$ be the maximum latency after $M_i$ transmission attempts of gNB $i$. Since a packet is retransmitted by HARQ after a decoding failure, $T_{\text{HARQ},i}$ can be expressed as

$$T_{\text{HARQ},i} = T_{\text{one},i} + (M_i - 1)(T_{\text{ACK\_align}} + T_{\text{ACK\_Tx}} + T_{\text{one},i}) \tag{8}$$

where $T_{\text{ACK\_align}}$ is the ACK alignment time, and $T_{\text{ACK\_Tx}}$ is the ACK transmission time. In Fig. 2, an example of the latency timeline is presented based on HARQ retransmission.

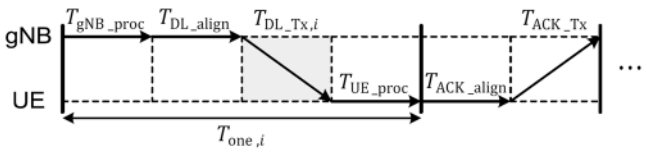


**Figure 2:** URLLC latency timeline based on HARQ retransmission mechanism

## 4 Proposed Scheme

The proposed DRO technique is a selective duplication scheme [8,9], where a duplicated packet of the S-gNB is transmitted only when an initial transmission of the M-gNB fails. Therefore, unnecessary transmissions of the duplicated packets are prevented and the total number of transmission attempts can be effectively reduced. Fig. 3 shows an example of the selective duplication scheme when an initial transmission of the M-gNB fails. An incoming packet is duplicated in the PDCP layer of the M-gNB and the duplicated packet is forwarded to the S-gNB only after HARQ NACK reception. Otherwise, if the initial transmission of the M-gNB is successful, ACK is transmitted instead of NACK and the duplicated packet in the M-gNB is discarded. In addition to selective duplication, DRO derives the optimal number of maximum transmission attempts to minimize the radio resource consumption.

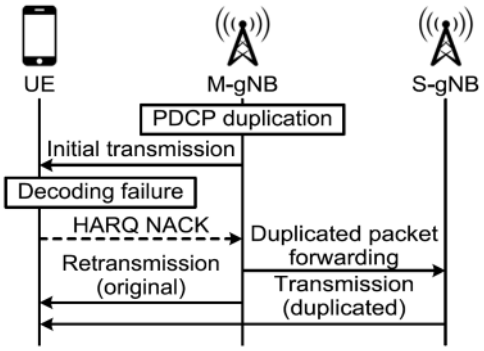


**Figure 3:** Signaling flow of selective duplication, where duplicated packet forwarding and transmission is carried out only after HARQ NACK reception

### *4.1 Problem Formulation*

Let $r_{\mathrm{DC}}$ be the total channel use to transmit an URLLC packet in the NR-DC system. The objective of the proposed DRO scheme is to derive the optimal number of maximum transmission attempts for each gNB that can minimize the expected channel use $\mathbb{E}[r_{\mathrm{DC}}]$ under the URLLC QoS requirements. Therefore, the optimization statement can be formulated as

$$\underset{M_1,M_2}{\text{minimize}} \quad \mathbb{E}[r_{\mathrm{DC}}] \tag{9}$$

$$\begin{aligned} \text{subject to} \quad & (\text{C1}): \ R_{\mathrm{DC}} \geq 1-\delta \\ & (\text{C2}): \ T_{\mathrm{HARQ},i} \leq d \qquad \forall i \in \{1,2\} \\ & (\text{C3}): \ r_i = k h_i s_i \qquad \forall i \in \{1,2\} \\ & (\text{C4}): \ h_i \leq W_i \qquad \forall i \in \{1,2\} \end{aligned}$$

where (C1) represents the URLLC reliability condition, in which the reliability $R_{\mathrm{DC}}$ should be greater than the reliability requirement $1-\delta$, (C2) represents the URLLC latency condition, in which the maximum transmission delay $T_{\mathrm{HARQ},i}$ should be smaller than the latency requirement $d$. Due to the selective duplication scheme, packet transmission from the S-gNB is delayed for one-shot transmission and the feedback time from the M-gNB and additional signaling delay $T_{\mathrm{PD\_delay}}$. Therefore, the maximum latency of each gNB can be rewritten as (10) and (11).

$$T_{\mathrm{HARQ},1} = T_{\mathrm{one},1} + (M_1 - 1)(T_{\mathrm{ACK\_align}} + T_{\mathrm{ACK\_Tx}} + T_{\mathrm{one},1}) \tag{10}$$

$$T_{\text{HARQ},2} = T_{\text{one},1} + T_{\text{ACK\_align}} + T_{\text{ACK\_Tx}} + T_{\text{PD\_Delay}} + T_{\text{one},2} + (M_2 - 1)(T_{\text{ACK\_align}} + T_{\text{ACK\_Tx}} + T_{\text{one},2}) \tag{11}$$

In addition, (C3) represents the channel use condition. The channel use in the time-frequency plane can be derived by the required bandwidth resource $h_i$, the required time resource $s_i$, and the number of channel use per unit time per unit bandwidth $k$ [10]. Since the packet transmission time is allocated based on the OFDM symbol duration, $T_{\text{DL\_Tx},i}$ can be expressed as

$$T_{\text{DL\_Tx},i} = \left\lceil \frac{s_i}{2^{-\mu}/14} \right\rceil \frac{2^{-\mu}}{14} \tag{12}$$

where $\lceil \cdot \rceil$ is the ceiling function. In addition, (C4) represents the bandwidth constraint, where an URLLC packet can exploit the $W_i$ bandwidth.

**Lemma 1.** *Suppose* $p_i = p_{i,1} = p_{i,2} = \cdots = p_{i,M_i}$ *for all* $i \in \{1,2\}$. *If the maximum number of transmission attempts on the i-th connection* $M_i$ *increases from* $M_i$ *to* $M_i + 1$, *then the expected channel use* $\mathbb{E}[r_{\text{DC}}]$. *increases by* $-r_1 \ln p_1 (p_1)^{M_1}/(1-p_1)$ *for* $i = 1$, *and* $-p_1 r_2 \ln p_2 (p_2)^{M_2}/(1-p_2)$ *for* $i = 2$. □

*Proof*. The expected channel use $\mathbb{E}[r_{\text{DC}}]$ can be expressed using the expected number of transmission attempts $\mathbb{E}[m_i]$ as follows.

$$\mathbb{E}[r_{\text{DC}}] = \mathbb{E}[m_1] r_1 + \mathbb{E}[m_2] r_2 \tag{13}$$

The expected number of transmission attempts for M-gNB can be derived as

**Algorithm 1:** DRO algorithm

DRO Algorithm

**Input:** SINR measurement $\gamma_1$, $\gamma_2$
**Output:** maximum transmission $M_1$, $M_2$

**Initialize** $M_1 = 1$, $M_2 = 1$
**while do**
  **for** gNB $i \in \{1,2\}$ **do**
    **Compute** $p_i$, $r_i$, $T_{\text{DL\_Tx},i}$
    **Compute** maximum latency $T_{\text{HARQ},i}$
  **end**
  **Compute** packet reliability $R_{\text{DC}}$
  **if** $R_{\text{DC}} \geq 1 - \delta$ and $T_{\text{HARQ},i} \leq d$ **then**
    **Terminate** the loop (optimal solution)
  **else**
    **Compute** $\partial \mathbb{E}[r_{\text{DC}}]/\partial M_i$ based on (17)
    **Select** $j = \arg\min_i (\partial \mathbb{E}[r_{\text{DC}}]/\partial M_i)$
    **Increase** $M_j \leftarrow M_j + 1$
    **Update** maximum latency $T_{\text{HARQ},j}$
    **if** $T_{\text{HARQ},j} > d$ **then**
      **Decrease** $M_j \leftarrow M_j - 1$
      **Increase** $M_k \leftarrow M_k + 1$ for $k \neq j$
      **Update** maximum latency $T_{\text{HARQ},k}$

(Continued)

**Algorithm 1:** Continued

```
19        if T_HARQ,k > d then
20            Terminate the loop (there exist no feasible solutions)
21        end
22      end
23    end
24  end
```

$$\mathbb{E}[m_1] = (1-p_1) + 2p_1(1-p_1) + \cdots + M_1 p_1^{M_1-1} = 1 + p_1 + p_1^2 + \cdots + p_1^{M_1-1} \tag{14}$$

However, the expected number of transmission attempts for the S-gNB can be derived as

$$\mathbb{E}[m_2] = p_1[(1-p_2) + 2p_2(1-p_2) + \cdots + M_2 p_2^{M_2-1}] = p_1(1 + p_2 + p_2^2 + \cdots + p_2^{M_2-1}) \tag{15}$$

since the S-gNB transmits a duplicated packet after transmission failure of the M-gNB. Therefore, the expected channel use $\mathbb{E}[r_{\mathrm{DC}}]$ can be derived as

$$\mathbb{E}[r_{\mathrm{DC}}] = \frac{1-p_1^{M_1}}{1-p_1} r_1 + \frac{1-p_2^{M_2}}{1-p_2} p_1 r_2 = \frac{r_1}{1-p_1} p_1^{M_1} - \frac{p_1 r_2}{1-p_2} p_2^{M_2} + \left( \frac{r_1}{1-p_1} + \frac{p_1 r_2}{1-p_2} \right) \tag{16}$$

Therefore, the partial differentials of $\mathbb{E}[r_{\mathrm{DC}}]$ can be derived as

$$\frac{\partial \mathbb{E}[r_{\mathrm{DC}}]}{\partial M_1} = -\frac{r_1 \ln p_1 (p_1)^{M_1}}{1-p_1}, \frac{\partial \mathbb{E}[r_{\mathrm{DC}}]}{\partial M_2} = -\frac{p_1 r_2 \ln p_2 (p_2)^{M_2}}{1-p_2} \tag{17}$$

In addition, the partial differential $\partial \mathbb{E}[r_{\mathrm{DC}}]/\partial M_i > 0$ is satisfied since $r_i > 0$ and $0 < p_i < 1$. ■

The approximation gap of $p_i = p_{i,1} = p_{i,2} = \cdots = p_{i,M_i}$ was neglected from the computation for simplicity since it has almost no effect on the expected channel use and optimal solution (0.953% error on average) based on the simulation results.

### *4.2 DRO Scheme*

Since $M_1$ and $M_2$ are integer values, the optimization statement of (9) becomes a NLIP problem. In addition, the expected channel use. $\mathbb{E}[r_{\mathrm{DC}}]$ is a monotonically increasing function since $\partial \mathbb{E}[r_{\mathrm{DC}}]/\partial M_i > 0$. Then, the minimum channel use can be achieved by increasing the maximum number of transmission attempts of the gNB which has the least gradient value $\partial \mathbb{E}[r_{\mathrm{DC}}]/\partial M_i$ until satisfying the reliability condition while not violating the latency condition. Therefore, the optimal number of maximum transmission attempts can be found iteratively based on Lemma 1. Basically, the DRO scheme is conducted by the M-gNB periodically to update the maximum transmission attempts according to the dynamic channel condition. The pseudo code of the DRO sc. he is presented in Algorithm 1.

As the input, the SINR measurements with respect to the reference signal are used, which are reported periodically from the UE. First, the maximum number of transmission attempts are initialized to 1 in line 1, then the DRO scheme iterates a loop until it achieves the URLLC requirements in line 2. The maximum latency $T_{\mathrm{HARQ},i}$ and the reliability $R_{\mathrm{DC}}$ are computed based on $p_i$, $r_i$, $T_{\mathrm{DL\_Tx},i}$ in lines 3–7. If the DRO scheme satisfies the URLLC requirements, it terminates the loop in lines 8–9. Otherwise, the partial differentials $\partial \mathbb{E}[r_{\mathrm{DC}}]/\partial M_i > 0$ are computed based on (17) in lines 10–11. Then, the maximum number of transmission attempts for gNB $j$ is increased by 1, where $j$ is an index of the gNB which least increases the expected channel use compared to the other gNBs in line 12–13. To examine if any violation of the latency requirements occurs, the maximum latency $T_{\mathrm{HARQ},i}$ is updated in line 14. If the latency requirement is violated, $M_j$ is decreased by 1 and the maximum transmission attempts

of the other gNB $M_k$ is increased by 1 instead in lines 15–17. If the latency requirement is violated after updating $T_{\text{HARQ},k}$, the loop is terminated since there exists no solution to satisfy the URLLC requirements in lines 18–20. Otherwise, the iteration is repeated based on line 2.

## 5 Performance Analysis

In this section, the performance of the DRO scheme is evaluated using MATLAB simulation, where the URLLC packet size was set to $L = 32$ bytes. The QoS parameters for URLLC packet transmission include the reliability requirement that was set to $\delta = 10^{-5}$, and the latency requirement which was set to $d = 1$ ms. Based on UE capability 2 and NR numerology $\mu = 2$, the UE processing time $T_{\text{UE\_proc}}$ was set to $N_1 = 0.161$ ms and the gNB processing time $T_{\text{gNB\_proc}}$ was set to $N_2/2 = 0.098$ ms [17]. The packet alignment time $T_{\text{DL\_align}}$ was set to a 0∼1 OFDM symbol duration since 7 MOs per slot was considered. In addition, an ideal backhaul is assumed and the ACK transmission time $T_{\text{ACK\_Tx}}$ was set to 1 OFDM symbol duration, where the ACK alignment time $T_{\text{ACK\_align}}$ was considered negligible.

In the simulation, NR gNBs form a hexagonal-shaped cellular coverage. The M-gNB is located at (−50, 0), and the S-gNB is located at (50), where the carrier frequency is set to 3.5 GHz. Since DC is activated in the cell edge regions and cell range extension (CRE) is considered, the UE is initially located at (−10, 15) and moves toward (20, 15). The performance of the DRO scheme was compared with LADMA [8] and SDUF [9] when the maximum transmission attempts $M = M_1 = M_2$ is fixed to 2 and 3, respectively. A performance comparison of the reliability and the maximum latency performance (which are the two major URLLC QoS requirements) are presented in Figs. 4 and 5, respectively. In addition, a performance comparison of the average channel use (which is an objective performance metric of the DRO scheme) is presented in Fig. 6.

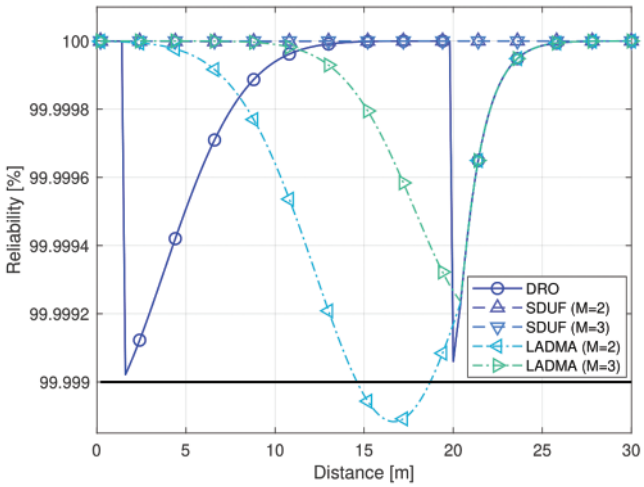


**Figure 4:** Simulation results with respect to the reliability performance

In Fig. 4, the reliability of the DRO scheme is presented, where the black bold line represents the reliability condition ($\delta = 10^{-5}$). The distance of the horizontal-axis represents the travel distance of the UE from the initial position. The simulation results show that the DRO scheme can satisfy the reliability condition regardless of the distance of the UE movement, for the range of interest. Since the DRO scheme reduces the maximum transmission attempts adaptively to minimize the average channel use, the reliability performance is degraded when the distance is 2 and 20, but the satisfaction of the

reliability condition is always guaranteed within the range of interest. In addition, if the maximum transmission attempts for the S-gNB is set to be larger than the M-gNB, the reliability of DRO increases as the UE moves further away from the M-gNB. While SDUF can satisfy the reliability condition for all distances regardless of the maximum transmission attempts, LADMA can satisfy the reliability condition only when the maximum transmission attempts are 3. Since LADMA relies on packet retransmissions of the M-gNB until the latency budget condition, the reliability of LADMA decreases as the UE moves further away from the M-gNB.

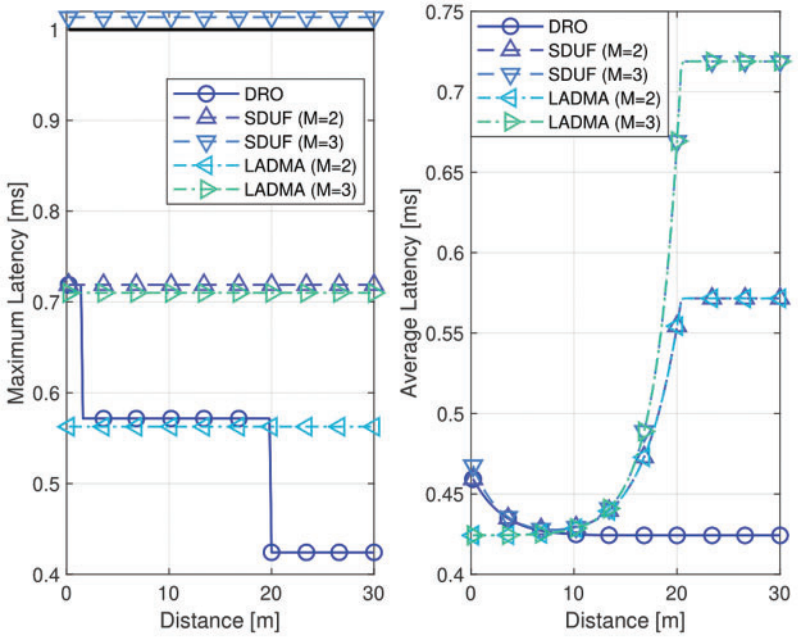


**Figure 5:** Simulation results with respect to the latency performance

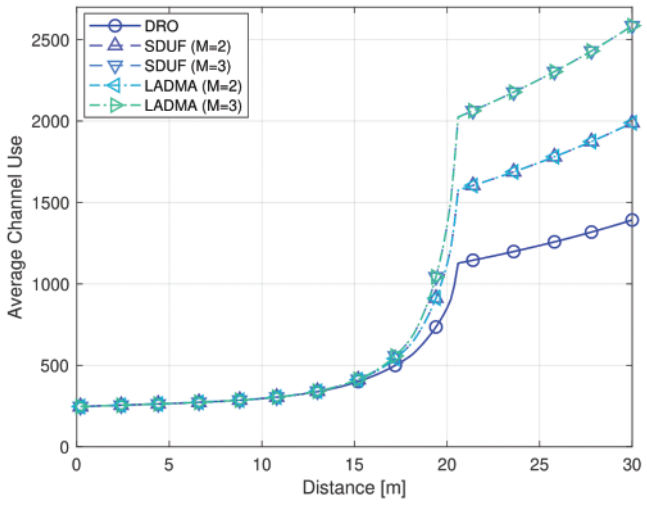


**Figure 6:** Simulation results with respect to the average channel use performance

In Fig. 5, the maximum latency and average latency of the DRO scheme is presented, where the black bold line in (a) represents the latency condition ($d = 1$). The simulation results show that the DRO scheme can satisfy the latency condition regardless of the distance of the UE movement. Since the DRO scheme can reduce the maximum transmission attempts when the distance is increased, the maximum latency is reduced. In addition, DRO shows the lowest average latency performance for the cell edge region compared to the other schemes by reducing redundant retransmissions from the unreliable M-gNB. In terms of the average latency, DRO can provide a 11.41% and 17.54% gain when compared to SDUF with $M = 2$ and $M = 3$, respectively, and can provide a 10.64% and 16.67% gain when compared to LADMA with $M = 2$ and $M = 3$, respectively.

In Fig. 6, the average channel use of the DRO scheme is presented with respect to the distance of the UE movement. Based on the simulation results of Figs. 4 and 5, the proposed DRO scheme can satisfy the URLLC QoS conditions for the entire range of interest. Furthermore, Fig. 6 shows that the DRO scheme can provide the lowest average channel use compared to LADMA and SDUF for the entire range of interest. Under the feasible region of the URLLC QoS conditions, the DRO scheme can provide a 11.64% and 17.78% reduced average channel use when compared to SDUF with $M = 2$ and $M = 3$, respectively, and can provide a 12.44% and 17.77% reduced average channel use when compared to LADMA with $M = 2$ and $M = 3$, respectively.

## 6 Conclusion

In this paper, a PD aware HARQ retransmission scheme named DRO is proposed, which can minimize the average resource usage while satisfying the URLLC requirements in the NR-DC architecture. The proposed DRO scheme derives the optimal value of the maximum transmission attempts for each gNB based on the average channel use. The simulation results show that the proposed DRO scheme can provide a significant performance gain compared to the existing PD schemes, which include LADMA and SDUF.

**Funding Statement:** This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the South Korea government (MSIT, 2021-0-00040, Development of intelligent stealth technology for information and communication resources for public affairs and missions).

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.
[2] 3GPP, "NR and NG-RAN overall description (Release 16)," 3GPP TS 38.300 v16.2.0, Jul. 2020.
[3] F. Xu, P. Zou, H. Wang, H. Cao, X. Fang *et al.,* "Resource allocation for D2D communication in cellular networks based on stochastic geometry and graph-coloring theory," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 12, pp. 4946–4960, 2020.
[4] Z. Li, X. Wang, J. Zhang, W. Huang and Y. Tian, "Temporal and spatial traffic analysis based on human mobility for energy efficient cellular network," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 1, pp. 114–130, 2021.
[5] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis," *IEEE Network*, vol. 32, no. 2, pp. 32–40, 2018.

[6] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen and G. Berardinelli, "Reliability oriented dual connectivity for URLLC services in 5G new Radio," in *2018 15th Int. Symp. on Wireless Communication Systems (ISWCS)*, Lisbon, Portugal, pp. 1–6, Aug. 2018.
[7] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G new Radio," in *2019 IEEE Wireless Communications and Networking Conf. Workshops (WCNCW)*, Marrakech, Morocco, pp. 1–6, Apr. 2019.
[8] N. H. Mahmood and H. Alves, "Dynamic multi-connectivity activation for ultra-reliable and low-latency communication," in *2019 16th Int. Symp. on Wireless Communication Systems (ISWCS)*, Oulu, Finland, pp. 1–5, Aug. 2019.
[9] M. Centenaro, D. Laselva, J. Steiner, K. Pedersen and P. Mogensen, "System-level study of data duplication enhancements for 5G downlink URLLC," *IEEE Access*, vol. 8, pp. 565–578, 2020.
[10] A. Anand and G. d. Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.
[11] S. E. Elayoubi, P. Brown, M. Deghel and A. Galindo-Serrano, "Radio resource allocation and retransmission schemes for URLLC over 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 896–904, 2019.
[12] H. Jang, J. Kim, W. Yoo and J. -M. Chung, "URLLC mode optimal resource allocation to support HARQ in 5G wireless networks," *IEEE Access*, vol. 8, pp. 126797–126804, 2020.
[13] 3GPP, "Multi-connectivity; Overall description (Release 16)," 3GPP TS 37.340 v16.7.0, Sep. 2021.
[14] J. Ramis and G. Femenias, "Cross-layer design of adaptive multirate wireless networks using truncated HARQ," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 944–954, 2011.
[15] H. D. Le, C. T. Nguyen, V. V. Mai and A. T. Pham, "On the throughput performance of TCP cubic in millimeter-wave cellular networks," *IEEE Access*, vol. 7, pp. 178618–178630, 2019.
[16] Y. Polyanskiy, H. V. Poor and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
[17] 3GPP, "Email discussion/approval on converging the proposals for eURLLC processing timeline," 3GPP R1-1901472, Jan. 2019.