URLLC Mode Optimal Resource Allocation to Support HARQ in 5G Wireless Networks

HYEONDEOK JANG*, JUNSUNG KIM*, WONSUK YOO*, AND JONG-MOON CHUNG*, (Senior Member, IEEE)
School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
Corresponding author: Jong-Moon Chung (jmc@yonsei.ac.kr)

ABSTRACT Ultra Reliable Low Latency Communication (URLLC) mode is a promising 5G technology for various real-time applications, such as, autonomous vehicles, augmented reality, and factory automation. Several papers on URLLC exist, however, most papers only focus on ways to achieve high reliability, and lack mathematical analysis of delay related to packet errors, retransmissions, and bandwidth. In this paper, a M/G/1 queuing model is applied to analyze the erroneous transmission recovery delay of URLLC multi-user services, and using this model the minimum required bandwidth is derived and applied into an adaptive control scheme. The proposed Pollaczek-Khinchine (P-K) formula based quadratic optimization (PFQO) scheme optimizes the bandwidth requirement by controlling the maximum retransmission parameter of the hybrid automatic repeated request (HARQ) mechanism in URLLC. Simulation results show the bandwidth saving effect of the proposed PFQO scheme based on various signal to interference plus noise ratios (SINRs) and packet length distributions.

INDEX TERMS 5G, URLLC, HARQ, queuing delay, latency.

I. INTRODUCTION
Ultra Reliable Low Latency Communication (URLLC) 5G technology was designed to support packet delivery with minimum latency and high reliability, which is required for real-time automation applications that must satisfy high levels of quality of service (QoS) to be precise and maintain stability. This makes URLLC appropriate for autonomous system services, such as, drone control, autonomous vehicle platooning, and factory automation. In URLLC systems, the end to end (E2E) delay should be less than 1 ms and the packet error rate (PER) needs to be less than $10^{-5}$ [1]. In order to meet the requirements, many technologies are suggested. The 5G New Radio (NR) is designed to use flexible numerology (e.g., sub-carrier spacing and mini-slots) to allocate resources efficiently for various services like URLLC, Enhanced Mobile Broadband (eMBB), and Massive Machine Type Communications (mMTC). In 5G, resources can be allocated at the mini-slots, which are much smaller units compared to the slots allocated in LTE [2]. In addition, feedback signals can be sent within the same slot using self-contained slots in NR to reduce the delay. Similar to Long-Term Evolution (LTE), hybrid automatic repeated request (HARQ) can be applied in NR for reliability improvement. Many papers on URLLC have proposed schemes to improve the reliability and latency simultaneously using techniques like early HARQ feedback prediction with machine learning [3] and blind retransmission HARQ [4], where results show that the feedback-delay waiting for an ACK/NACK can become too large to satisfy the low-latency constraints. Therefore, papers propose schemes that send retransmission messages in advance using channel state information prediction to reduce the latency in HARQ. In order to satisfy the QoS requirements, URLLC packets have higher priority over other applications, which are implemented using schemes like puncturing eMBB transmission [5]. The studies of [6] and [7] propose schemes where URLLC transmissions puncture eMBB transmissions in a way that has little impact to eMBB transmissions. In [6], Conditional Value at Risk (CVaR) is used to measure the risk of eMBB transmissions, and Markov inequality is used to maintain the URLLC reliability. Research in [7] and [8] propose a scheme to decompose the resource allocation...
problem into two sub-optimal problems of URLLC and eMBB transmission and solve the sub-problems. Applying network slice technology for URLLC applications makes maintaining URLLC and eMBB QoS more effective. In [9], by controlling the modulation coding scheme (MCS) and retransmission resource parameters, the authors minimize the total resource consumption based on reliability constraints. The authors of [10] divide the use of one-shot transmission and the HARQ scheme considering the reliability, block error rate (BLER), and required channel use. In [11], a scheme that controls the maximum number of transmissions and minimizes the bandwidth (based on a finite block length assumption of the system operation) using the square root staffing rule is proposed. However, these papers do not perform an URLLC protocol based queuing delay analysis or consider the scenario where the URLLC packets can be transmitted on demand. These factors have to be considered when there are multiple URLLC users, such as in control of drone swarms or autonomous vehicle platoons. This paper deals with the formulation is based on the HARQ end-to-end delay, which includes the queuing delay and transmission delay. This is used to transmit the packets to multi-class URLLC users and resources are allocated through the slot based NR numerology system. The gNB transmits a finite block length packets of $L_i$ bits to the $i^{th}$ user equipment (UE) (among $M$ users) denoted as $U_i$ ($i = 1, 2, \ldots, M$). In order to transmit a packet of $L_i$ information bits, the channel use factor $r_i$ is required to access a finite block size transmission of

$$r_i = \frac{L_i}{C(\psi_i)} + \frac{(Q^{-1}(p))^2 V(\psi_i)}{2C(\psi_i)^2} \left(1 + \sqrt{1 + \frac{4L_i C(\psi_i)}{V(\psi_i)(Q^{-1}(p))^2}}\right)$$

(1)

where, $C(\psi) = \log_2(1 + \psi)$ is the channel capacity based on an infinite block length, $V(\psi) = (\log_2(e))^2(1 - \frac{1}{(1+\psi)^2})$ is the channel dispersion, $\psi_i$ is signal to interference plus noise ratio (SINR) for each user, $p$ is the decoding failure probability, and $Q^{-1}(x)$ is the inverse $Q$ function of $x$ [11]. It is assumed that there are discrete $C$ classes of users, where packets of each class have (nearly) the same SINR [11]. As the gNB analyzes the Channel State Information (CSI), each packet of the classes are stacked and serviced based on queuing order transmissions (corresponding to the NR numerology), as shown in Fig. 1.

In order to deal with the HARQ packet length influence, uniform and beta distributions are considered in the 5G gNB antenna resource allocation M/G/1 model. Based on use of network slice technology, derivations to determine how much bandwidth is required to allocate an URLLC slice, considering NR specifications and queuing delay are provided in this paper, which could be implemented into future 5G schedulers to enhance the URLLC performance.

The performance of PFQO is compared to the bandwidth square root staffing rule based optimization (SSRO) scheme of [11], which uses the square root staffing rule to optimize the URLLC bandwidth.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In the analysis, 5G gNB base stations use the 3 GHz and mmWave frequency bands (based on the Release 16 3GPP standards on URLLC [1]) to transmit the packets to multi-users and resources are allocated through the slot based NR numerology system. The gNB transmits a finite block length packets of $L_i$ bits to the $i^{th}$ user equipment (UE) (among $M$ users) denoted as $U_i$ ($i = 1, 2, \ldots, M$). In order to transmit a packet of $L_i$ information bits, the channel use factor $r_i$ is required to access a finite block size transmission of

$$r_i = \frac{L_i}{C(\psi_i)} + \frac{(Q^{-1}(p))^2 V(\psi_i)}{2C(\psi_i)^2} \left(1 + \sqrt{1 + \frac{4L_i C(\psi_i)}{V(\psi_i)(Q^{-1}(p))^2}}\right)$$

(1)

where, $C(\psi) = \log_2(1 + \psi)$ is the channel capacity based on an infinite block length, $V(\psi) = (\log_2(e))^2(1 - \frac{1}{(1+\psi)^2})$ is the channel dispersion, $\psi_i$ is signal to interference plus noise ratio (SINR) for each user, $p$ is the decoding failure probability, and $Q^{-1}(x)$ is the inverse $Q$ function of $x$ [11]. It is assumed that there are discrete $C$ classes of users, where packets of each class have (nearly) the same SINR [11]. As the gNB analyzes the Channel State Information (CSI), each packet of the classes are stacked and serviced based on queuing order transmissions (corresponding to the NR numerology), as shown in Fig. 1.

A. NR NUMEROLOGY

The sub-carrier spacing and time slot duration can be adjusted with flexible numerology in 5G systems. In the system model, the channel use factor $r_i = k_i s_i h_i$ is required to send a packet, where, $k_i$ is the number of channel use per unit bandwidth and per unit time, $s_i$ is the period of the transmission time,
and \( h_i \) is the allocated bandwidth for transmission [11]. At the gNB, a bandwidth of \( W = E[k_i h_i] \) is required to transmit the packets within the time period. The data rate can be obtained from \( \eta W \), where \( \eta \) is the spectral efficiency in units of \( \text{bits/s/Hz} \), in which \( \eta \) is from 3.24 ∼ 11.25 \( \text{bits/s/Hz} \) in general for mmWave communications [13]. The URLLC system may adaptively reduce the spectral efficiency \( \eta \) to lower the error rate and maintain the wireless link reliability.

In a normal cyclic prefix (CP) slot, 14 symbols are allocated per slot. Compared to LTE (which allocates radio resources in blocks of 14 OFDM symbols, i.e., 1 TTI/subframe), 5G NR more efficiently allocates \( l \) (e.g., 2, 4, 7) symbol units (called a mini-slot) per slot in the downlink [14], which is shown in Table 1. While the frame duration and subframe durations are fixed regardless of the Sub-Carrier Spacing (SCS), the symbol duration becomes shorter as the SCS increases. Correspondingly, the duration to transmit a packet is represented as \( s_i = \frac{l}{15(2^{-m})} \) ms and the subcarrier spacing is \( \Delta f = 15(2^m) \text{kHz} \), where \( m = 0, 1, 2, 3, 4 \) [15].

### B. QUEUING MODEL

It is assumed that, at a gNB, the packets to be sent follow a Poisson distribution, and because each user uses a different application with a different packet length, the service time needs to be modeled using a general distribution [11]. Since packets are transmitted one by one, the system is modeled using a M/G/1 queue model like in [16]. M/G/1 is a queuing model that applies a packet arrival rate pattern that follows a Poisson distribution, service rate follows a general distribution, and packets are serviced one by one. In Fig. 2, when a packet is generated, the packet is placed in the queue of the gNB. The queuing delay and the transmission delay occur at the gNB. If the receiver fails to decode the packet, it returns a NACK to request a retransmission up to a maximum of \( \phi \) times, and feedback delay occurs on each round. The total delay should be less than the delay constraint, and can be expressed as (2), which does not include the time spent in the retransmissions.

\[
T_{\text{total}} = T_{\text{queue}} + T_{\text{trans}} + T_{\text{prop}} + T_{\text{proc}} \tag{2}
\]

In (2), \( T_{\text{prop}} \) is the propagation delay and \( T_{\text{proc}} \) is the packet processing time, which can be modeled as a deterministic value. When the uplink resources are sufficient, the feedback delay can also be considered a constant value [11].

Since the gNB is maintained as a steady stable system, the waiting time and the arrival rate can be considered as average values. Furthermore, the number of packets in the queue can be calculated based on Little’s theory [12]. At the gNB, \( \lambda \) packets/ms are generated and sent to \( M \) users. In the HARQ scheme, packet decoding can fail at the receiver, which leads to a retransmission. The number of transmission attempts is limited by the maximum number of retransmissions \( \phi \), based on the delay constraints. There are various HARQ schemes, such as, Incremental Redundancy HARQ (IR-HARQ) and Chase Combining HARQ (CC-HARQ) that can be considered.

In this paper, the chase-combining model is analyzed based on the assumption that the gNB retransmits the same packet if erroneously received. The total average downlink arrival rate at the gNB is \( \lambda_{\text{down}} = \lambda (1 + \sum_{k=1}^{\phi} \prod_{i=1}^{k} p_i) \), where \( \phi \) is the number of maximum retransmission attempts and \( p_k \) is the

| Sub-carrier spacing (kHz) | 15  | 30  | 60  | 120 |
|---------------------------|-----|-----|-----|-----|
| frame duration (ms)       | 10  | 10  |     |     |
| Subframe per frame        |     |     |     |     |
| Subframe duration (ms)    | 1   | 2   | 4   | 8   |
| Slots per subframe        |     |     |     |     |
| Symbols per slot          | 14  |     |     |     |
| Symbols duration (ms)     | 66.7| 33.3| 16.7| 8.33|
| Mini-slot length (symbols)| PDSCH : 2, 4, 7 / PUSCH : 1 - 14 (Type B) |
Packet Error Rate (PER) at the $k^{th}$ transmission ($p_\phi < \ldots < p_2 < p_1$). Based on the block-fading channel model, $p_k$ is represented [17] as

$$p_k(\psi_{k,a}) = \begin{cases} 1 & \text{if } 0 < \psi_{k,a} < \psi_M \\ \exp(-g(\psi_{k,a} - \psi_M)) & \text{if } \psi_{k,a} \geq \psi_M \end{cases} \quad (3)$$

where $g$ is the parameter derived according to the Modulation and Coding Scheme (MCS) and $\psi_M$ is the maximum SINR when the decoding failure probability is 1, which can be calculated depending on the MCS. In the CC-HARQ scheme, when the decoding failure probability is 1, which can be approximated using the P-K formula in the M/G/1 system [18],

$$E_r \approx \frac{E[L_i]}{C(\psi)} + \frac{(Q^{-1}(p))^2V(\psi)}{2C(\psi)^2} + \frac{Q^{-1}(p)V(\psi)}{C(\psi)^{3/2}}E[\sqrt{L_i}] \quad (5)$$

$$E_r \approx \frac{E[L_i^2]}{C(\psi)^2} + \frac{(Q^{-1}(p))^2V(\psi)}{4C(\psi)^4} + \frac{(Q^{-1}(p))^2V(\psi)E[L_i]}{C(\psi)^3} + \frac{(Q^{-1}(p))^2V(\psi)^{3/2}E[\sqrt{L_i}]}{C(\psi)^{5/2}} \quad (6)$$

The distribution of the packet length ($L_i$) bits can be obtained from traffic monitoring, or approximated using a packet length distribution function, thus $E[L_i]$, $E[\sqrt{L_i}]$, $E[L_i^2]$, and $E[L_i^{3/2}]$ can be obtained. Although, the transmission delay as well as the queuing delay to analyze the overall delay, which can be simplified as $T_{\text{trans}} = \bar{X}$, needs to be derived.

B. VACATION

In order to model the multi-user system’s queuing delay, the time durations in which a user cannot access the channel due to other users transmitting, is modeled using vacation intervals. The vacation pattern of the downlink system can be modeled based on the slot configuration, which is determined when allocating downlink and uplink symbols. When $\tau$ symbols of the uplink mini-slots are set in slot format, the vacation of the system can be represented as $\bar{V} = \frac{\tau}{\bar{T}} \epsilon \text{ ms (} \tau = 0, 1, 2, \ldots, 14\text{)}$, where $\epsilon$ is the probability that the uplink request is idle. The slot format supported in the 5G system is defined in the 3GPP standardization and can be assumed to follow fixed slot formats. Therefore, the uplink request can be assumed to be periodic, where the constant vacation intervals lead to $\alpha^2 = 0$.

C. DELAY BOUNDS BASED ON THE URLLC REQUIREMENTS

To satisfy the URLLC latency and reliability,

$$T_{\text{total}} \leq T_d \quad (7a)$$

$$\left\{ \begin{array}{l}
\phi + 1 \leq \sum_{k=1}^{\phi+1} p_k \leq \delta
\end{array} \right\} \quad (7b)$$

$$T_{\text{total}} = (\phi + 1) \left( T_{\text{queue}} + T_{\text{trans}} + T_{\text{prop}} + T_{\text{proc}} \right) \quad (7c)$$

where $\phi$ is the number of maximum retransmission attempts, $p_k$ is the decoding failure probability at the $k^{th}$ transmission, $T_d$ is the delay constraint, and $\delta$ is the PER constraint. Latency and reliability constraints are represented in (7a) and (7b). The total delay (including the transmission time, queuing delay, and feedback delay, multiplied by $(\phi + 1)$ in (7c)) needs
to be less than the delay constraint 1ms. Since $T_{prop}$ and $T_{pren}$ are deterministic values, the feedback delay $T_f$ is also constant, and therefore, the inequality of (8) is obtained.

$$T_{queue} + T_{trans} \leq T_{target} = \frac{T_d}{\phi + 1} - T_f$$  \hspace{1cm} (8)

### IV. MINIMUM BANDWIDTH REQUIREMENT OPTIMIZATION

In order to derive the minimum bandwidth requirement of $W$ based on the maximum retransmission attempt $\phi$ and system constraints, the optimization problem is formulated as (9).

$$\phi^* = \arg \min W$$  \hspace{1cm} (9a)

s.t. C1: $T_{target} \geq T_{queue} + T_{trans}$

$$\phi + 1$$  \hspace{1cm} (9b)

C2: $\prod_{k=1}^{\lambda} p_k \leq \delta$

$$k = 1$$  \hspace{1cm} (9c)

C3: $r = k\phi$

$$C$$  \hspace{1cm} (9d)

C4: $\rho = \lambda_{down, \bar{X}} = \frac{\lambda_{down, \bar{X}} [r_i]}{W} < 1$

$$\rho$$  \hspace{1cm} (9e)

From (9e), the bandwidth lower bound becomes $\lambda_{down, \bar{X} [r_i]} < W$ to maintain stability. Therefore, based on $\Delta f = 15(2^m)$ kHz, the lower bound for a stable queuing system can be expressed as in (10).

$$W_m = \left[ \frac{\lambda_{down, \bar{X} [r_i]}}{15 (2^m)} \right] 15 (2^m) \text { kHz (} m = 0, 1, \ldots, n) \hspace{1cm} (10)$$

In order to simplify the formulation, a mapping of

$$A = (M \bar{V} - 2T_{target})$$  \hspace{1cm} (11a)

$$B = (2E [r_i] - \lambda_{down, \bar{V} E [r_i] + 2T_{target} \lambda_{down, E [r_i]}])$$  \hspace{1cm} (11b)

$$C = (\lambda_{down, \bar{V} [r_i]} \lambda_{down [r_i]}]^2 - 2 \lambda_{down [r_i]})$$  \hspace{1cm} (11c)

is applied to the optimization function based on the determination equation $D = B^2 - 4AC$, where $M$ is the number of users, $\bar{V}$ is the average of the vacation, and $\lambda_{down}$ is the arrival rate of packets at the gNB.

**Lemma 1:** If $A > 0$, there is no $W^*$ that satisfies the URLLC requirement.

**Proof:** In order to derive the optimum bandwidth $W$, two conditions should be considered. At first, according to (9), $T_{queue} + T_{trans}$ needs to be less than $T_{target}$ to satisfy the latency constraint. The other condition is $W^*$ should be larger than $\lambda_{down, \bar{X} [r_i]}$ to maintain a stable queuing system. In order to consider the first condition, $T_{queue} + T_{trans}$ is expressed in terms of $\bar{X}$, $\bar{X}^2$, and $\bar{V}$. This can be calculated in terms of $E [r_i], E [r_i^2]$, and $W$ based on (5) and (6).

$$T_{queue} + T_{trans} = \frac{\lambda_{down, \bar{X}^2}}{2 (1 - \lambda_{down, \bar{X}})} + \frac{\bar{V}}{2 (1 - \lambda_{down, \bar{X}})} + \bar{X} = \frac{\lambda_{down, E [r_i]} + \bar{M} \bar{V} - \lambda_{down, \bar{V} E [r_i]} + E [r_i]}{2 (1 - \lambda_{down, E [r_i]}) W} + \bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

Considering the stable system condition of $\rho < 1$ based on (9e), $W >= \lambda_{down, \bar{X} [r_i]} W > 0$. As $W$ increases, the service time and queuing delay decreases and $\lim_{W \rightarrow \infty} T_{queue} + T_{trans} = \frac{M \bar{V}^2}{2}$, therefore, the lower bound of $T_{queue} + T_{trans}$ is $\frac{M \bar{V}^2}{2}$. If $M \bar{V} > 2T_{target}$ (i.e., $A > 0$), the lower bound of $T_{queue} + T_{trans}$ is larger than $T_{target}$. Therefore, there is no $W^*$ that can satisfy the URLLC requirement in this case.

**Lemma 2:** If $A < 0$ & $D < 0$, then $W^* = W_m$, else if $A < 0$ & $D > 0$ & $W_m > \max \left( \frac{-B - \sqrt{D}}{2A}, \frac{-B + \sqrt{D}}{2A} \right)$, then $W^* = \max \left( W_m, \max \left( \frac{-B - \sqrt{D}}{2A}, \frac{-B + \sqrt{D}}{2A} \right) \right)$, else if $A < 0$ & $D > 0$ & $W_m \leq \min \left( \frac{-B - \sqrt{D}}{2A}, \frac{-B + \sqrt{D}}{2A} \right)$, then $W^* = W_m$.

**Proof:** Similar to the proof of Lemma 1, (12) is used to optimize the bandwidth. The quadratic relation is expressed in (13).

$$\frac{1}{2} (\sum_{i=1}^{\lambda} \lambda_{down, \bar{V} E [r_i] + 2T_{target} \lambda_{down, E [r_i]}} W$$

$$+ \bar{M} \bar{V} - 2T_{target}) + (2E [r_i] - \lambda_{down, \bar{V} E [r_i] + 2T_{target} \lambda_{down, E [r_i]}} W$$

$$+ \bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

$$\bar{X}$$

The relation of (13) can be expressed in a simplified $AW^2 + BW + C < 0$ form in terms of $W$. Since any $W$ value can satisfy the requirement when $A < 0$ & $D \leq 0$, $W^*$ only needs to be above $W_m$. Therefore, $W^* = W_m$. If $A < 0$ & $D > 0$, the optimization needs to be analyzed separately according to $W_m$ and the smaller solution of the quadratic equation. If $W_m$ is larger than the smaller solution of the quadratic equation, then $W^*$ is determined by the larger solution of the quadratic equation and $W_m$, $W^* = \max \left( W_m, \max \left( \frac{-B - \sqrt{D}}{2A}, \frac{-B + \sqrt{D}}{2A} \right) \right)$. Else, $W^*$ can be set in the range to be less than the smaller solution of the quadratic equation. Since $W^*$ only needs to be above $W_m$, $W^* = W_m$.

The derived optimal bandwidth $W^*$ is allocated to the queue of each class. In reality, since the bandwidth isn’t infinite, $W^*$ should be less than the bandwidth upper limit. Based on [19], a resource grid is used for an antenna port and the number of resource blocks is fixed, meaning that the maximum bandwidth is fixed at the classes as Table 2. As the required bandwidth is evaluated in section V, the optimal bandwidth is less than the maximum bandwidth indicated.

| $\lambda_{numerosity (\mu)}$ | $\text{Min RB}$ | $\text{Max RB}$ | $\text{Min Bandwidth (MHz)}$ | $\text{Max Bandwidth (MHz)}$ |
|-----------------------------|----------------|----------------|-----------------------------|-----------------------------|
| 0                           | 24             | 275            | 4.32                        | 49.5                        |
| 1                           | 24             | 275            | 8.64                        | 99                          |
| 2                           | 24             | 275            | 17.28                       | 198                         |
| 3                           | 24             | 275            | 34.56                       | 396                         |
| 4                           | 24             | 275            | 69.12                       | 397.44                      |
in Table 2. Furthermore, even if the channel state is worse or the packet length is longer, a larger bandwidth can be supported by Carrier Aggregation (CA). According to [20] and [21], both FR1 (from 410 MHz to 7,125 MHz) and FR2 (from 24.25 GHz to 52.6 GHz) supports CA up to 16 carriers. Since the upper bound is larger than the required bandwidth in performance evaluation, the system can be maintained stably.

**Algorithm 1 PFQO Algorithm**

**INPUT:** λ, ψ, M, V, L

**INITIALIZE** $T_f = 0.125$ ms, $T_{total} = 0$ ms, $W = []$, $φ = 0$

1: while True do
2:   compute decoding failure probability $p$ based on (3)
3:   compute $E[r]$ and $E[r^2]$ based on (5), (6)
4:   compute $T_{trans}$ and $T_{queue}$ based on (4)
5:   $T_{total} = (φ + 1)(T_{trans} + T_{queue} + T_f)$
6:   if $T_{total} > 1$ ms then
7:     break
8:   else
9:     compute $A$, $B$, $C$ based on (11)
10:    if $A > 0$ then
11:       no $W^*$ solution exists
12:    else
13:       compute $W^*$ based on Lemma 2
14:       $W[φ] = W^*$
15:    end if
16:   end if
17: end if
18: while
19: $W_{optimal} = \min W$
20: $φ_{optimal} = \arg \min W$

**OUTPUT:** $W_{optimal}$, $φ_{optimal}$

The optimization problem of (9) can be solved by Algorithm 1. After computing the decoding failure probability and the parameters including $A$, $B$, and $C$, the optimal bandwidth based on Lemma 1 and 2 is derived according to the condition that $A$ is positive. This procedure is repeated while the total delay is less than 1 ms, as $φ$ is increased by 1. Since the complexity of (11) can be calculated in terms of the number of users $M$, and the equations consists of simple additions and multiplications, the complexity of computing (11) is $O(M)$. After calculating $A$, $B$, and $C$, the problem can be solved using the quadratic formula and using the comparison in Lemmas 1 and 2, which has linear complexity. This problem should be solved at each transmission attempt. Although the maximum retransmission attempt $φ$ affects the complexity to compute the optimal solution, the number of permitted transmission attempts will be few, due to the URLLC delay constraints. Therefore, the complexity of this optimization problem is $O(M)$, which shows linear complexity characteristics.

**TABLE 3. Simulation parameters.**

| Parameters                | Value  |
|---------------------------|--------|
| Number of Users $M$       | 10     |
| Feedback Delay $T_f$      | 0.125 ms |
| Delay constraint $T_d$    | 1 ms   |
| Packet Error Rate constraint $δ$ | $10^{-6}$ |
| Number of uplink symbols in a slot $τ$ | 2 symbols |
| Idle Probability $ϵ$      | 0.01   |

**V. PERFORMANCE ANALYSIS**

In this section, the bandwidth requirement to satisfy the required QoS performance of the proposed PFQO scheme is compared to the SSRO scheme of [11] in terms of packet arrival rate, packet length, and SINR. Because the PFQO scheme uses the derived optimum bandwidth formula, accurate adaptation of the required bandwidth based on the maximum number of transmissions (limit) is applied in the operations, where this effectiveness is demonstrated in the figures below. In the evaluation, the decoding failure probability is assumed to be the same in each transmission round, namely, repetition coding and the homogenous transmissions model of [11] is applied to compare the performance with the SSRO scheme (i.e., $p_1 = p_2 = \ldots = p_{φ+1}$).

The performance of PFQO and SSRO is simulated using MATLAB (R2019a), where the maximum number of transmissions is controlled to minimize the bandwidth requirement based on the number of users $M = 10$, $T_f \approx 0.125$ ms, delay constraint $T_d = 1$ ms, PER constraint $δ = 10^{-6}$, and the number of channel use per unit time and per unit bandwidth $k = 1$. The number of maximum transmissions, including both the original transmission and all retransmissions, is set from 1 to 7 in consideration of the one-shot transmission and the delay constraint. The parameter values used in the simulation are summarized in Table 3.

The analysis in Fig. 3 was conducted while ignoring the vacations in the uplink, where the packet arrival rate range is $10 \sim 100$ packets/ms, SINR = 20 dB, and the packet length $L_i$ has an Uniform distribution of [150, 250] bits. If the number of maximum transmissions increases, the allowed decoding failure probability of each round increases which leads to a need to reduce the redundancy and channel use based on the fixed maximum allowed URLLC latency. However, since too many retransmission attempts require more expected bandwidth for each attempt, it shows convex features in the results.

URLLC has a recommended packet length limit of 32 bytes for evaluation purposes, and within this range, each application has an unique packet length distribution. Based on this consideration, Fig. 4 compares the performance results when the packet length has an Uniform or Beta distribution (based on the parameters $α$ and $β$), number of uplink symbols in a slot $τ = 2$ symbols, idle probability $ϵ = 0.01$, packet arrival rate $λ = 100$ packets/ms, and SINR = 10 dB. Since URLLC packets have a maximum and minimum packet length, the Beta distribution is more suitable to use in
approximating the packet length distribution compared to the Normal distribution, which has a long tail [22].

The analysis in Fig. 5 is based on the number of uplink symbols in a slot $\tau = 2$ symbols, idle probability is $\epsilon = 0.01$, packet arrival rate of 100 packets/ms, and the packet length $L_i$ has an Uniform distribution of $[100, 200]$ bits.

Fig. 3 shows that PFQO requires less bandwidth than SSRO while satisfying the URLLC requirements, where the performance gain of PFQO increases as the packet arrival rate increases. This is due to the difference where SSRO blocks and drops a packet if there isn’t sufficient bandwidth to support transmissions immediately, while PFQO places the packet in the transmission queue of the gNB, which has a bandwidth saving effect. Furthermore, the PFQO optimized formula enables accurate instantaneous adaptation based on the current status of the URLLC system.

Fig. 4 shows the trend of bandwidth requirement increment as the packet size increases based on the Uniform and Beta distributions. Changing $\alpha$ and $\beta$ of the Beta distribution changes the mean and variance of the channel usage, which affects the bandwidth requirements. This is because when $\alpha$ is larger than $\beta$, the probability of larger packets being generated increases, which influences the bandwidth requirement.

VI. CONCLUSION

In this paper, the proposed PFQO scheme optimizes the bandwidth considering HARQ based on 5G URLLC communication. In comparison with other papers, the proposed scheme uses the queuing delay model based on the P-K formula while including the influence of the transmission time. The analysis of this paper formulates the relation of the allocated bandwidth and total delay, and proposes a more accurate delay model of the URLLC mode. The simulation results show that the bandwidth requirement when using PFQO is less compared to when using the SSRO scheme. For example, at the 100 packets/ms rate, the simulation experiments show that the proposed PFQO scheme results in a 32.3 $\sim$ 62.9% smaller required bandwidth for the range of 1 $\sim$ 7 maximum number of transmissions, when compared to the SSRO scheme. In addition, even at a much lower rate of 10 packets/ms, the proposed PFQO scheme results in a 34.1 $\sim$ 64.8% smaller required bandwidth for the same transmission attempts range of interest, when compared to the SSRO scheme. For future research, multi-server systems can be investigated by applying a M/G/m model to the URLLC system performance analysis. Such modeling is expected to enable higher levels of performance analysis for 5G URLLC systems. In addition, detailed HARQ feedback delay like applying blind-retransmission or ACK timing in Rel-16 (e.g., K1 parameter) will be studied.

REFERENCES

[1] Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC), document 3GPP TR 38.824, v.16.0.0, Release 16, 3GPP, Mar. 2019.
[2] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, “NR: The new 5G radio access technology,” IEEE Commun. Standards Mag., vol. 1, no. 4, pp. 24–30, Dec. 2017.
[3] N. Strodthoff, B. Göktepe, T. Schierl, C. Heilge, and W. Samek, “Enhanced machine learning techniques for early HARQ feedback prediction in 5G,” IEEE J. Sel. Areas Commun., vol. 37, no. 11, pp. 2573–2587, Nov. 2019.
[4] R. Abreu, G. Berardinelli, T. Jacobsen, K. Pedersen, and P. Mogensen, “A blind retransmission scheme for ultra-reliable and low latency communications,” in *Proc. IEEE 87th Veh. Technol. Conf.*, Jun. 2018, pp. 1–5.

[5] A. Anand, G. De Veciana, and S. Shakkottai, “Joint scheduling of URLLC and eMBB traffic in 5G wireless networks,” in *Proc. IEEE INFOCOM*, Apr. 2018, pp. 1970–1978.

[6] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, and C. S. Hong, “A matching based coexistence mechanism between eMBB and uRLLC in 5G wireless networks,” in *Proc. 34th ACM SIGAPP SAC*, Apr. 2019, pp. 2377–2384.

[7] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, “EMBB-URLLC resource slicing: A risk-sensitive approach,” *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.

[8] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, “Coexistence mechanism between eMBB and uRLLC in 5G wireless networks,” *CoRR*, vol. abs/2003.04551, pp. 1–30, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.04551

[9] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, “Radio resource allocation and retransmission schemes for URLLC over 5G networks,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 896–904, Apr. 2019.

[10] H. Shariatmadari, S. Iraji, and R. Jäntti, “Analysis of transmission methods for ultra-reliable communications,” in *Proc. IEEE 26th PIMRC*, Aug./Sep. 2015, pp. 2303–2308.

[11] X. Anand and G. de Veciana, “Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2411–2421, Nov. 2018.

[12] D. P. Bertsekas, R. G. Gallager, and P. Humblet, *Data Networks*, vol. 2, Upper Saddle River, NJ, USA: Prentice-Hall, 1992.

[13] V. Wong, R. Schober, D. Ng, and L. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[14] H. Jang, *Numerology Requirements*, document R1-1706173, Qualcomm, 3GPP TSG RAN WG1 Meeting #84, Busan, South Korea, Apr. 2016.

[15] Y. Kim and S. Park, “Calculation method of spectrum requirement for IMT-2020 eMBB and URLLC with puncturing based on M/G/1 priority queuing model,” *IEEE Access*, vol. 8, pp. 25027–25040, 2020.

[16] X. Lagrange, “Performance analysis of HARQ protocols with link adaptation on fading channels,” *Annales des Telecommun.*, vol. 66, pp. 695–705, Mar. 2011.

[17] L. Kleinrock, *Queuing Systems: Computer Applications*, vol. 2, New York, NY, USA: Wiley, 1976.

[18] *Physical Channels and Modulation*, document 3GPP TS 38.211, v.16.1.0, Release 16, 3GPP, Dec. 2019.

[19] *Numerology Requirements*, document R1-1706173, Qualcomm, 3GPP TSG RAN WG1 Meeting #84, Busan, South Korea, Apr. 2016.

[20] *Physical Channels and Modulation*, document 3GPP TS 38.211, v.16.1.0, Release 16, 3GPP, Mar. 2020.

[21] *User Equipment (UE) Radio Transmission and Reception Part 1: Range 1 Standalone*, document 3GPP TS 38.101, v.16.1.0, Release 16, 3GPP, Dec. 2019.

[22] *User Equipment (UE) Radio Transmission and Reception Part 2: Range 2 Standalone*, document 3GPP TS 38.101-2, v.16.1.0, Release 16, 3GPP, Dec. 2019.

[23] E. R. S. Castro, M. S. Alencar, and I. E. Fonseca, “Probability density functions of the packet length for computer networks with bimodal traffic,” *Int. J. Comput. Netw. Commun.*, vol. 5, no. 3, pp. 17–31, May 2013.

**HYEONDEOK JANG** received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2019, where he is currently pursuing the M.S. degree in electrical and electronic engineering. He is also a Researcher at the Communications and Networking Laboratory (CNL), Yonsei University. His current research interests include 5G communication, cellular vehicular to everything communication (C-V2X), and deep learning.

**JUNSUNG KIM** received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2015, where he is currently pursuing the combined M.S. and Ph.D. degree in electrical and electronic engineering. He is also a Researcher at the Communications and Networking Laboratory (CNL), Yonsei University. His current research interests include the Internet of Things (IoT), 5G, and hybrid automatic repeat request (HARQ).

**WONSUK YOO** received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2016, where he is currently pursuing the combined M.S. and Ph.D. degree in electrical and electronic engineering. He is also a Researcher at the Communications and Networking Laboratory (CNL), Yonsei University. His current research interests include 5G communication, cellular vehicular to everything communication (C-V2X), and dedicated short-range communications (DSRC).

**JONG-MOON CHUNG** (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Yonsei University and the Ph.D. degree in electrical engineering from Pennsylvania State University. Since 2005, he has been a Professor with the School of Electrical and Electronic Engineering, Yonsei University, where he is currently the Associate Dean of the College of Engineering and a Professor of emergency medicine at the College of Medicine, Yonsei University. From 1997 to 1999, he was an Assistant Professor and Instructor at Pennsylvania State University. From 2000 to 2005, he was with Oklahoma State University (OSU), as a tenured Associate Professor. His 12 Coursera courses focused on deep learning, big data, cloud computing, 5G and 4G mobile communications, Wi-Fi, Bluetooth, augmented reality (AR), the Internet of Things (IoT), Skype, YouTube, TCP/IP, as well as hardware and software of smartphones and smartwatches. In 2019, he received the Minister Award from the Ministry of the Interior and Safety and the Appreciation Award from the National Police Agency of the South Korea. In 2008, 2019, and 2018, he received the Outstanding Accomplishment Faculty Awards, and in 2009, 2014, and 2019, he received the Outstanding Teaching Awards from Yonsei University. In 2012, he received the Republic of Korea Government’s Defense Acquisition Program Administration (DAPA) Award. As a tenured Associate Professor at OSU, in 2005, he received the Regents Distinguished Research Award and the Halliburton Outstanding Young Faculty Award. In 2003 and 2004, he received the Technology Innovation Award and the Distinguished Faculty Award from OSU, respectively, and the First Place Outstanding Paper Award at the IEEE EIT 2000 Conference, Chicago, USA, in 2000. He is also a Pledge Book Award Winning Member of the Eta Kappa Nu (HKN). He is currently serving as the General Chair of the IEEE ICCE-Asia 2020 and the IEEE ICCE 2022, and was the General Chair of several conferences, including the IEEE MWSCAS 2011. He is also serving as the Vice President of the IEEE Consumer Electronics Society, an Editor of the IEEE Transactions on Vehicular Technology, an Associate Editor of the IEEE Transactions on Consumer Electronics, a Section Editor of the EURASIP Journal (Wiley), and the Co-Editor-in-Chief of the KSII Transactions on Internet and Information Systems.