Retransmission Policy with Frequency Hopping for Ultra-reliable and Low-latency Communications

Chengjian Sun, Changyang She and Chenyang Yang

***Abstract*—In this work, we study the benefit of retransmission to ultra-reliable and low-latency communications (URLLC). We consider a retransmission policy that employs frequency hopping to improve the retransmission success probability for the packets suffering deep fading. We investigate how to satisfy the quality-of-service (QoS) with retransmission by taking downlink transmission as an example. End-to-end (E2E) delay components, including transmission delay, queueing delay and backhaul latency, and packet loss components, including transmission error probability and E2E delay violation probability are considered. Resource allocation for the retransmission policy is optimized. Numerical results show that the requirement on transmission error probability can be significantly relaxed with retransmission, which does not compromise the E2E delay requirement owing to the short packets in URLLC. Simulation results indicate that more users can be supported with QoS guarantee compared to a counterpart system without retransmission, and the performance gain increases with the times of retransmissions.**

***Index Terms*—Ultra-reliable and low-latency communications, retransmission, quality-of-service, short blocklength**

I. Introduction

Ultra-reliable and low-latency communications (URLLC) is critical to support emerging mission critical applications in tactile internet, autonomous vehicle networks, and smart factories of the fifth generation (5G) cellular networks [1, 2].

Retransmission is a natural way to achieve ultra-high reliability (e.g., $99.999\% \sim 99.99999\%$). However, due to the ultra-low latency requirement in URLLC, the end-to-end (E2E) delay (e.g., 1 ms) is typically shorter than the channel coherence time. Thus, successive retransmissions to a user suffering deep fading are more likely also unreliable, and the reliability can hardly benefit from standard retransmission techniques such as automatic repeat request (ARQ). A method to deal with such problem is to combine retransmission with various diversities. A cooperative ARQ was studied in machine-to-machine communications [3], where macro-diversity was combined with retransmission. The results in [3] show that the cooperative ARQ can increase the reliability by several orders of magnitude.

In [3], Shannon's capacity was employed, which is accurate to approximate the maximal achievable rate when the blocklength of channel code is large. However, due to the small packet size (e.g., 20 bytes) [1] and short transmission delay, the blocklength is very short in URLLC. More importantly, to ensure the stringent quality-of-service (QoS) requirements imposed on both delay and reliability, the data rate and transmission error probability should be jointly considered in URLLC, in contrast to traditional communication systems. While Shannon's capacity is widely used to optimize resource allocation for traditional systems, it cannot characterize the connection between the achievable rate and the transmission error probability, whose impact on the reliability of URLLC can never be ignored.

An accurate approximation of the achievable rate in short blocklength regime has been derived in [4] and was employed in [5–9] to study how to ensure QoS in URLLC. In [5], bandwidth allocation was optimized to ensure the reliable communication between two devices in a cellular network with limited transmission attempts, where uplink (UL) and downlink (DL) transmission delay were jointly considered. Optimal power allocation was derived in closed-form for ARQ in [6], achieving low outage probability. However, the impact of queueing delay was not taken into account in [5,6], which was shown non-negligible in [7] and was addressed in [8,9]. Nonetheless, retransmission was not considered in [7–9]. When retransmission is allowed, the mechanism to satisfy both the E2E delay and the overall packet loss probability by adjusting among delay and reliability components will change. How to model and ensure the QoS of URLLC with retransmission is not well-understood.

In this paper, we investigate the benefit of using retransmission in URLLC. We first present a retransmission policy, where the packets failed in previous transmission are allowed to be retransmitted. An individual subchannel is assigned to each retransmission of a packet, such that the packets failed to be transmitted in previous frame can be retransmitted at the same time with the first-time transmission of the newly arrived packets. Frequency hopping is employed for retransmission to exploit frequency diversity. We proceed to show how to ensure the QoS with such a retransmission policy, and optimize the resources allocated to each transmission of a packet for the policy. Transmission delay, queueing delay and backhaul latency are considered in the E2E delay, and transmission error probability and E2E delay violation probability are controlled to satisfy the reliability requirement. Numerical and simulation results show that, with more retransmissions the requirement on transmission error probability can be relaxed, as expected, which can simplify the coding design for URLLC. Moreover, higher spectral efficiency can be achieved, owing to the short

C. Sun and C. Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China (e-mail:{sunchengjian,cyyang}@buaa.edu.cn).

C. She is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372 (e-mail: shechangyang@gmail.com).

This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61429101.

packets, one of the unique traffic feature of URLLC.

II. System Model

Consider a cellular network, where each base station (BS) with $N_{\rm t}$ antennas serves K single-antenna users with bandwidth $W_{\max}$. The maximal transmit power of each BS is $P_{\max}$. Since interference is detrimental to reliability, we consider to avoid multi-user interference by frequency division multiple access (say by orthogonal frequency division multiple access), and avoid strong inter-cell interference by frequency reuse among adjacent cells.

Time is discretized into frames. The duration of each frame is $T_{\rm f}$. The duration for UL or DL transmission in one frame is $\tau < T_{\rm f}$. Time division duplex is used as an example [10], while the analysis is also applicable to frequency division duplex.

A. Achievable Rate with Short Blocklength

In URLLC, the blocklength of channel coding is short due to the required low E2E latency. Further considering the high reliability requirement, the packet loss caused by transmission errors cannot be ignored. The Shannon's Capacity cannot characterize the transmission error probability, denoted as $\varepsilon^{\rm c}$. In typical application scenarios of URLLC, the channel coherence time is longer than the E2E delay, and thereby the channel is quasi-static. Besides, the packet size in URLLC is small, hence it is reasonable to assume that the bandwidth allocated for transmitting each packet is less than the channel coherence bandwidth. In quasi-static flat fading channel, when channel state information is available at the transmitter and receiver, the maximal achievable service rate of the kth user can be accurately approximated as [4],

$$s_k \approx \frac{\tau W_k}{u T_{\rm f} \ln 2}\left[\ln\left(1+\frac{\alpha_k g_k P_k}{N_0 W_k}\right) - \sqrt{\frac{V_k}{\tau W_k}} Q_{\rm G}^{-1}(\varepsilon^{\rm c})\right], \quad (1)$$

where u is the number of bits contained in each packet, W_k and P_k are the bandwidth and transmit power allocated to the kth user, respectively, α_k and g_k are the large-scale channel gain and small-scale channel gain of the kth user, respectively, N_0 is the single-side noise spectral density, $Q_{\rm G}^{-1}(x)$ is the inverse of the Gaussian Q-function, and V_k is the channel dispersion given by [4]

$$V_k = 1 - \frac{1}{\left(1+\frac{\alpha_k g_k P_k}{N_0 W_k}\right)^2}. \quad (2)$$

Although the achievable rate is in closed-form, it is still too complicated to be tractable for optimizing resource allocation. Fortunately, to ensure ultra-low latency and ultra-high reliability, the system should operate in high SNR level. As shown in [11], if SNR is higher than 10 dB, $V_k \approx 1$ is accurate. Even when the SNR is not high, we can obtain a lower bound of the achievable rate by substituting $V_k \approx 1$ into s_k. If the required $\varepsilon^{\rm c}$ can be satisfied with the lower bound, it can also be satisfied with the achievable rate in (1).

B. QoS Requirement

The QoS requirement of the users supported by URLLC can be characterized by ensuring both E2E delay $D_{\max}$ and overall packet loss probability $\varepsilon_{\max}$.

We consider a local communication scenario, e.g., autonomous vehicle communications and smart factory [2], where the communication distance is within the coverage of adjacent cells. Then, propagation delay can be ignored, and the delay components includes backhaul latency, queueing delay, and transmission delay in the UL and DL. The reliability components highly depend on whether or not we allow retransmission, as detailed in the sequel.

III. QoS Requirement without Retransmission

To help understand how retransmission changes the way in ensuring the QoS, we first state the QoS components for a system without retransmission in this section.

With one-hop backhaul link, the backhaul latency can be regarded as deterministic and identical for different users, denoted as $D^{\rm b}$. Since the packets target to each user arrive at the buffer of BS randomly, and the inter-arrival time between packets could be shorter than the service time (i.e., transmission duration) of each packet, queueing delay cannot be ignored. We consider a queueing model that the packets for different users wait in different queues. Denote the queueing delay of a packet for the kth user as $D_k^{\rm q}$, its UL and DL transmission delay as $D_k^{\rm u}$ and $D_k^{\rm d}$, respectively. Since $D_k^{\rm q}$ is random, the E2E delay requirement can be expressed as

$$\Pr\left\{D_k^{\rm u} + D_k^{\rm d} + D_k^{\rm q} + D^{\rm b} \geq D_{\max}\right\} \leq \varepsilon^{\rm e}, \quad (3)$$

where $\varepsilon^{\rm e}$ is the required E2E delay violation probability.

As shown in [7], for Rayleigh fading channel or Nakagami-m fading channels, the transmit power required to ensure transmission error probability and $(D_{\max}, \varepsilon^{\rm e})$ is unbounded when the small-scale channel gain is close to zero, which exceeds $P_{\max}$. To ensure the QoS with finite transmit power, a proactive packet dropping mechanism was proposed in [7]. Denote the proactive packet dropping probability in UL and DL transmission as $\varepsilon^{\rm pu}$ and $\varepsilon^{\rm pd}$, respectively. Then, to ensure the reliability of URLLC without retransmission, the overall packet loss probability requirement can be expressed as

$$\begin{aligned} &1-(1-\varepsilon^{\rm cu})(1-\varepsilon^{\rm cd})(1-\varepsilon^{\rm pu})(1-\varepsilon^{\rm pd})(1-\varepsilon^{\rm e}) \\ \approx &\varepsilon^{\rm cu}+\varepsilon^{\rm cd}+\varepsilon^{\rm pu}+\varepsilon^{\rm pd}+\varepsilon^{\rm e} \leq \varepsilon_{\max}, \end{aligned} \quad (4)$$

where $\varepsilon^{\rm cu}$ and $\varepsilon^{\rm cd}$ are transmission error probabilities required for UL and DL. Such approximation is accurate because each probability is small.

To satisfy the reliability requirement in (4), the probability of each packet loss component should be extremely small. Very low $\varepsilon^{\rm cu}$ and $\varepsilon^{\rm cd}$ make coding design challenging. Very small value of $\varepsilon^{\rm pu}$ and $\varepsilon^{\rm pd}$ require very high transmit power [7]. One way to relax the requirement on $\varepsilon^{\rm cu}$ and $\varepsilon^{\rm cd}$ is to introduce retransmission mechanism. With retransmission, proactive packet dropping becomes unnecessary.

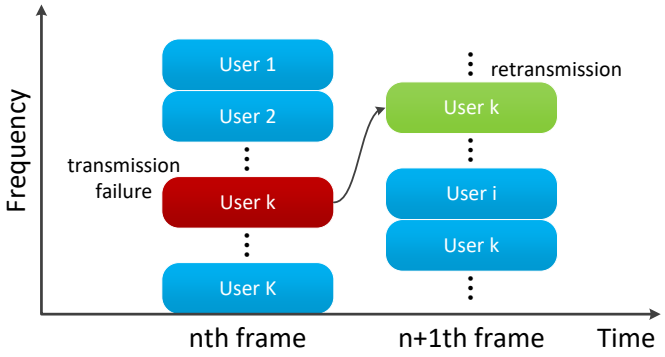

Fig. 1. A retransmission policy with frequency hopping.

IV. A Retransmission Policy

In this section, we propose a retransmission policy and derive the required radio resources to support the QoS. For easy exposition, we take DL transmission as an example.

A. Retransmission with Frequency Hopping

When a packet is not transmitted successfully in a frame, the packet will be retransmitted at the same service rate in the next frame, as shown in Fig. 1. Since the E2E delay is typically shorter than the channel coherence time, a user suffering from deep fading in a frame may still experience deep fading in the subsequent frames. In such scenario, frequency hopping can be employed for retransmission. The separations between the frequency bands assigned to the multiple transmissions for a packet should be larger than the channel coherence bandwidth, such that the small-scale channel gains in the transmissions are independent. Note that sometimes a user (say the kth user in Fig. 1) may have some new packets to receive in each frame. For instance, when some packets that are not received successfully in the nth frame are retransmitted in the $n+1$th frame, some new packets may also need to transmit to the user in the $n+1$th frame. In order not to affect the transmission for the newly arrived packets, different subchannels are used for retransmissions, i.e., the failed packets are retransmitted in the same frame as the transmission of the new packets.

Denote $N_{\mathrm{R}} \geq 0$ as the maximum number of times allowed to retransmit each packet. $N_{\mathrm{R}} = 0$ means that no retransmission is allowed, which corresponds to the non-retransmission policies [7–9]. If the number of times to retransmit a packet has not reached N_{R}, the failed packet will be retransmitted in the next frame. Otherwise, the packet is lost.

B. Ensuring the QoS with Retransmission

Since the packets to be retransmitted on each subchannel are the packets failed to be transmitted in previous transmission, the "arrival rate" of these packets at the BS is not higher than the service rate of previous transmission. By setting the service rate of each retransmission for a packet equal to the service rate of the first-time transmission for the packet, the retransmitted packets will not accumulate into a queue. Then, there is no queueing delay for retransmissions, and only the queueing delay for the first-time transmission needs to be considered to ensure the E2E delay.

Since ensuring E2E delay is meaningless for a packet that is not successfully received by a user, we only consider the E2E delay of the packets that can be successfully transmitted within N_{R} retransmissions. Let M denote the number of times for retransmitting a packet that ensures the packet to be transmitted successfully. If $M > N_{\mathrm{R}}$, then the packet is lost due to the exhaustion of retransmissions.

With the retransmission policy, the total delay of a packet caused by the UL and DL transmissions for the kth user, i.e., $D_k^{\mathrm{u}} + D_k^{\mathrm{d}}$, is a random variable, because M is random. For the case that a packet is successfully received by the kth user in the mth retransmission, $D_k^{\mathrm{u}} + D_k^{\mathrm{d}} = (m+1)T_{\mathrm{f}}$ for a time division duplex system. The E2E delay violation probability in this case can be expressed as

$$\Pr\{D_k^{\mathrm{u}} + D_k^{\mathrm{d}} + D_k^{\mathrm{q}} + D^{\mathrm{b}} \geq D_{\max} | M = m\} = \Pr\{D_k^{\mathrm{q}} \geq D_{\max,m}^{\mathrm{q}}\}, \tag{5}$$

where $D_{\max,m}^{\mathrm{q}} \triangleq D_{\max} - D^{\mathrm{b}} - (m+1)T_{\mathrm{f}}$. Then, the E2E delay requirement in (3) can be rewritten as

$$\sum_{m=0}^{N_{\mathrm{R}}} \Pr\{M = m \mid M \leq N_{\mathrm{R}}\} \Pr\{D_k^{\mathrm{q}} \geq D_{\max,m}^{\mathrm{q}}\} \leq \varepsilon^{\mathrm{e}}. \tag{6}$$

Since we consider DL as an example, in the sequel we denote the DL transmission error probability in the mth retransmission as $\varepsilon_m^{\mathrm{c}}$, $0 \leq m \leq N_{\mathrm{R}}$, where $m = 0$ indicates the first-time transmission. If the service rate required to ensure $\varepsilon_m^{\mathrm{c}}$ and $(D_{\max}, \varepsilon^{\mathrm{e}})$ cannot be achieved with finite transmit power due to channel fading, instead of being proactively dropped, some packets will wait for retransmission and not be served in the current transmission.

Denote $\varepsilon_m^{\mathrm{h}}$ as the probability that a packet has to wait for next retransmission in the mth retransmission. Then, the retransmission probability of the packet can be expressed as $1-(1-\varepsilon_m^{\mathrm{c}})(1-\varepsilon_m^{\mathrm{h}}) \approx \varepsilon_m^{\mathrm{c}} + \varepsilon_m^{\mathrm{h}}$, and the probability that the packet failed to be transmitted in all the m retransmissions can be accurately approximated as

$$\Pr\{M > m\} \approx \prod_{i=0}^{m} (\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}}). \tag{7}$$

Hence, we have

$$\begin{aligned} &\Pr\{M = m \mid M \leq N_{\mathrm{R}}\} \\ \approx &\frac{\Pr\{M > m-1\} - \Pr\{M > m\}}{1 - \Pr\{M > N_{\mathrm{R}}\}} \\ = &\frac{\prod_{i=0}^{m-1} (\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}}) - \prod_{i=0}^{m} (\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}})}{1 - \prod_{i=0}^{N_{\mathrm{R}}} (\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}})} \\ = &\frac{(1 - \varepsilon_i^{\mathrm{c}} - \varepsilon_i^{\mathrm{h}}) \prod_{i=0}^{m-1} (\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}})}{1 - \prod_{i=0}^{N_{\mathrm{R}}} (\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}})} \triangleq p_m. \end{aligned} \tag{8}$$

Since a packet will be lost if it cannot be successfully transmitted within N_{R} retransmissions or will be useless if its' E2E delay exceeds $D_{\max}$, the reliability requirement with the considered retransmission policy can be expressed as

$$1 - (1 - \Pr\{M > N_{\mathrm{R}}\})(1 - \varepsilon^{\mathrm{e}})$$

$$\approx \Pr\{M > N_{\mathrm{R}}\} + \varepsilon^{\mathrm{e}}$$
$$\approx \prod_{i=0}^{N_{\mathrm{R}}} \left(\varepsilon_i^{\mathrm{c}} + \varepsilon_i^{\mathrm{h}}\right) + \varepsilon^{\mathrm{e}} \leq \varepsilon_{\max}^{\mathrm{D}}, \quad (9)$$

where $\varepsilon_{\max}^{\mathrm{D}}$ is the maximum packet loss probability allowed for DL transmission in order to satisfy the overall reliability. Again, the approximations are accurate since the probabilities are very small. From (9) we can see that, given ε^{e}, the probabilities $\varepsilon_m^{\mathrm{c}}$ and $\varepsilon_m^{\mathrm{h}}$ increase with N_{R}. This indicates that the reliability of each transmission can be relaxed with more retransmissions.

According to the analysis in [7], when queueing delay is shorter than coherence time, the service rate is constant within the queueing delay bound of each packet, and then effective bandwidth can be used to control the queueing delay and queueing violation probability. An upper bound of the queueing violation probability is given in [7] as follows,

$$\Pr\{D_k^{\mathrm{q}} \geq D_{\max,m}^{\mathrm{q}}\} \leq \exp\left[-\theta_k E_k^{\mathrm{B}}(\theta_k) D_{\max,m}^{\mathrm{q}}\right], \quad (10)$$

where $E_k^{\mathrm{B}}(\theta_k)$ is the effective bandwidth for the kth user, and θ_k is the QoS exponent reflecting the performance in terms of queue length [12].

Then, by substituting (10) into (6), a conservative constraint on the E2E delay can be obtained as,

$$\sum_{m=0}^{N_{\mathrm{R}}} p_m \exp\left[-\theta_k E_k^{\mathrm{B}}(\theta_k) D_{\max,m}^{\mathrm{q}}\right] = \varepsilon^{\mathrm{e}}, \quad (11)$$

with which the constraint in (6) can be satisfied.

For a Poisson process, which is a typical arrival process in vehicle communication scenarios and other machine type communication scenarios [13,14], the effective bandwidth can be derived as [7]

$$E_k^{\mathrm{B}}(\theta_k) = \frac{\lambda_k}{T_{\mathrm{f}}\theta_k}\left(e^{\theta_k} - 1\right), \quad (12)$$

where λ_k is average arrival rate of the packets desired by the kth user.

Substituting (12) into (11), θ_k can be solved numerically, and then the effective bandwidth required to ensure the E2E delay can be obtained by substituting θ_k into (12). For notational simplicity, (θ_k) will be omitted from $E_k^{\mathrm{B}}(\theta_k)$ in the rest of the paper.

When there are packets to be transmitted to the kth user, $(D_{\max}, \varepsilon^{\mathrm{e}})$ can be ensured when the service rate is equal to or higher than the effective bandwidth. Substituting (1) and $V_k \approx 1$ into $s_k \geq E_k^{\mathrm{B}}$, we can obtain the transmit power required to ensure $\varepsilon_m^{\mathrm{c}}$ and $(D_{\max}, \varepsilon^{\mathrm{e}})$, i.e.,

$$P_{k,m} \geq \frac{N_0 W_{k,m}}{\alpha_k g_{k,m}}\left\{\exp\left[\frac{uT_{\mathrm{f}}E_k^{\mathrm{B}}\ln 2}{\tau W_{k,m}} + \frac{Q_{\mathrm{G}}^{-1}(\varepsilon_m^{\mathrm{c}})}{\sqrt{\tau W_{k,m}}}\right] - 1\right\}$$
$$\triangleq y_{k,m}(W_{k,m}, E_k^{\mathrm{B}}), \quad (13)$$

where $P_{k,m}$ and $W_{k,m}$ are the transmit power and bandwidth allocated to the kth user for the mth retransmission, respectively, and $g_{k,m}$ is the small-scale channel gain of the kth user in the mth retransmission. Here, $W_{k,m}$ is in fact the bandwidth of the subchannel for the kth user in the mth retransmission.

It is worthy to notice that when N_{R} increases, the transmission delay increases, and then the queueing delay requirement becomes more stringent in order to ensure the E2E delay. As a result, a larger effective bandwidth is required. Since $y_{k,m}(W_{k,m}, E_k^{\mathrm{B}})$ increases with E_k^{B}, the required transmit power will increase with N_{R}. On the other hand, however, more transmit power can be saved by relaxing the reliability requirement for each transmission when N_{R} grows. As to be shown by simulation in Section VI, more users can be supported by more retransmissions with QoS guarantee for given system resources.

V. Resource Allocation Optimization for Retransmission Policy

To evaluate the potential of the proposed retransmission policy, we optimize resource allocation for the system with retransmission in this section.

If a user has no packet to receive in a frame, then the BS will not allocate power and bandwidth to the user in the frame. The bandwidth and power allocated to the users that have packets to receive are optimized to minimize the total transmit power required to ensure $\varepsilon_m^{\mathrm{c}}$, $(D_{\max}, \varepsilon^{\mathrm{e}})$ according to the large-scale and small-scale channel gains.

The resource allocation optimization problem can be formulated as

$$\min_{P_{k,m}, W_{k,m}} \quad \sum_{k=1}^{K}\sum_{m=0}^{N_{\mathrm{R}}} P_{k,m} \quad (14)$$
$$\text{s.t.} \quad \sum_{k=1}^{K}\sum_{m=0}^{N_{\mathrm{R}}} W_{k,m} \leq W_{\max}, \quad (15)$$
$$(13),\ P_{k,m} > 0,\ W_{k,m} > 0,$$
$$k = 1, 2, \cdots, K,\ m = 0, 1, \cdots, N_{\mathrm{R}}.$$

The total transmit power $\sum_{k=1}^{K}\sum_{m=0}^{N_{\mathrm{R}}} P_{k,m}$ is a random variable depending on the small-scale channel gains. Some packets need to wait for next time retransmission when $\sum_{k=1}^{K}\sum_{m=0}^{N_{\mathrm{R}}} P_{k,m} > P_{\max}$. When the total transmit power is minimized for any given small-scale channel gains, the optimization is equivalent to minimizing the probability that the packets have to wait for retransmission, i.e., $\Pr\{\sum_{k=1}^{K}\sum_{m=0}^{N_{\mathrm{R}}} P_{k,m} > P_{\max}\}$.

Although problem (14) is non-convex due to the non-convex function $y_{k,m}(W_{k,m}, E_k^{\mathrm{B}})$ in constraint (13), it has been proved as equivalent to the following convex problem in [8,9],

$$\min_{P_{k,m}, W_{k,m}} \quad \sum_{k=1}^{K}\sum_{m=0}^{N_{\mathrm{R}}} P_{k,m} \quad (16)$$
$$\text{s.t.} \quad (13),\ (15),\ W_{k,m} \leq W_{k,m}^{\mathrm{th}},\ P_{k,m} > 0,\ W_{k,m} > 0,$$
$$k = 1, 2, \cdots, K,\ m = 0, 1, \cdots, N_{\mathrm{R}},$$

where $W_{k,m}^{\mathrm{th}}$ is the minimum point of $y_{k,m}(W_{k,m}, E_k^{\mathrm{B}})$ for a given E_k^{B}. Then, problem (16) can be solved [15].

In what follows, we show how to implement the proposed retransmission policy with the optimized resource allocation by a BS with maximal transmit power $P_{\max}$.

Denote the minimized total transmit power in problem (16) as P^*_{tot}. When $P^*_{\text{tot}} \le P_{\max}$, the global optimal transmit power and bandwidth allocation can be obtained by solving problem (16). When $P^*_{\text{tot}} > P_{\max}$, some packets have to wait for retransmission according to the following procedure.

For easy exposition, we use index (k,m) to denote the subchannel assigned to the kth user for the mth retransmission. The number of packets waiting for retransmission but should have been transmitted on subchannel (k,m) is denoted as $d_{k,m}$, whose initial value is zero.

When $P^*_{\text{tot}} > P_{\max}$, a packet to be transmitted on subchannel (k',m') with the lowest channel gain will be removed from the service list of current frame, waiting for the next time retransmission, and $d_{k',m'}$ increases by 1. If there is no packet to be transmitted on subchannel (k',m') with the updated service list, $P_{k',m'}=0$, $W_{k',m'}=0$. Otherwise, since some packets are removed from the service list for subchannel (k',m') with rate $d_{k',m'}/T_{\text{f}}$, the requirement $(D_{\max},\varepsilon^{\text{e}})$ can be ensured when the service rate is $E^{\text{B}}_{k'}-d_{k',m'}/T_{\text{f}}$. Hence, the transmit power requirement for subchannel (k',m') in problem (16) can be relaxed to $P_{k',m'} \ge y_{k',m'}(W_{k',m'}, E^{\text{B}}_{k'}-d_{k',m'}/T_{\text{f}})$. Then, the resources are reallocated by solving problem (16). If $P^*_{\text{tot}} > P_{\max}$ still holds with the newly allocated resources, more packets need to wait for retransmission until $P^*_{\text{tot}} \le P_{\max}$.

VI. Simulation and Numerical Results

In this section, we first show the impact of the maximum number of retransmission times on the required transmission error probability as well as the tradeoff between transmission delay and queueing delay via numerical results. Then, we evaluate the packet loss probabilities in DL transmission with and without retransmission, and the bandwidth required to ensure the QoS via simulations.

We consider a single cell scenario, where all users are located at the edge of the cell (i.e., the worst case), and the user-BS distance is 250 m. Rayleigh fading channel is considered. Other parameters are listed in Table I, unless otherwise specified. According to the parameters, the delay for a packet with eight retransmissions will exceed $D_{\max}$, hence in what follows we consider $N_{\text{R}}=0\sim 7$.

Optimizing the combination of the reliability components can enhance the system performance, which however leads to intractable optimization. According to [7], the optimal solutions of the reliability components in the non-retransmission policy are in same order of magnitude. To make the policies with and without retransmission comparable, for the non-retransmission policy we simply set $\varepsilon^{\text{c}}=\varepsilon^{\text{p}}=\varepsilon^{\text{e}}=\varepsilon^{\text{D}}_{\max}/3$, and for the retransmission police we set $\varepsilon^{\text{e}}=\varepsilon^{\text{D}}_{\max}/3$, and set $\varepsilon^{\text{c}}_m=\varepsilon^{\text{h}}_m=\sqrt[N_{\text{R}}+1]{2\varepsilon^{\text{D}}_{\max}/3}/2$, $m=0,1,\cdots,N_{\text{R}}$ such that the equality in constraint (9) is satisfied.

According to the above setup, the numerical results of the required transmission error probability to ensure the E2E delay and overall reliability with different values of N_{R} are shown in Table II. The results show that ε^{c}_m is significantly increased by using retransmission. Only by retransmitting once, ε^{c}_m is in the level of 10^{-3}.

TABLE I
Simulation Parameters

Max. DL packet loss probability $\varepsilon^{\text{D}}_{\max}$	10^{-5}
E2E delay requirement $D_{\max}$	1 ms
Duration of each frame T_{f}	0.1 ms
Duration of DL transmission τ	0.05 ms
Backhaul latency D^{b}	0.1 ms [16]
Max. transmit power of BS $P_{\max}$	43 dBm
Available bandwidth for DL $W_{\max}$	20 MHz
Number of transmit antennas of BS N_{t}	4
Single-sided noise spectral density N_0	-173 dBm/Hz
Packet size u	20 bytes [1]
Packet arrival rate λ_k	2000 packets/s
Path loss model $10\lg(\alpha_k)$	$35.3+37.6\lg(d_k)$

TABLE II
Transmission error probabilities with different N_{R}.

N_{R}	0	1	2	3
ε^{c}_m	3.33×10^{-6}	1.29×10^{-3}	9.41×10^{-3}	2.54×10^{-2}
N_{R}	4	5	6	7
ε^{c}_m	4.61×10^{-2}	6.86×10^{-2}	9.11×10^{-2}	1.13×10^{-1}

To understand why so many times of retransmissions do not make the E2E delay harder to ensure, we provide the complementary cumulative distribution functions (CCDFs) of transmission delay and queueing delay in Fig. 2, which are numerically computed with (7) and the right-hand side of (10), respectively. We can observe a tradeoff between transmission delay and queueing delay as N_{R} increases, and the queueing delay does not increase significantly until $N_{\text{R}}=7$. Essentially, this is because the packets are with small size in URLLC, which can be transmitted in a short time compared with the average inter-arrival time between packets. Further considering the randomness of packet arrival and wireless channel, the

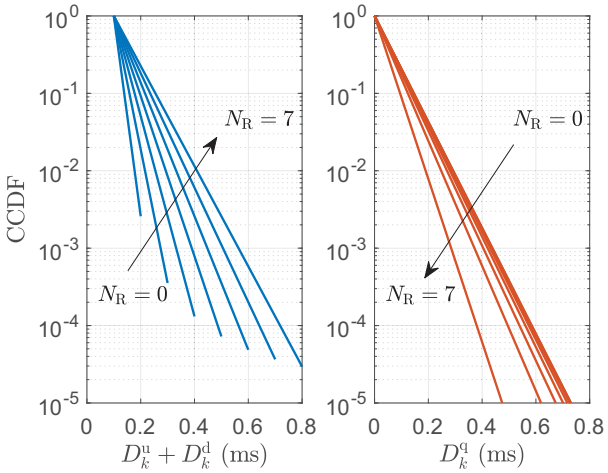

Fig. 2. CCDF of transmission delay (left) and queueing delay (right).

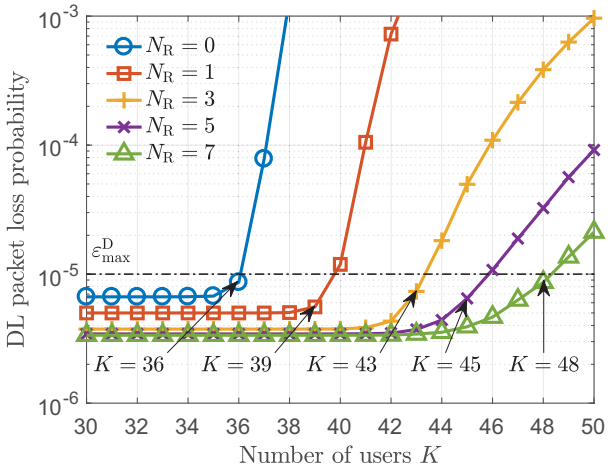

Fig. 3. DL packet loss probability vs. number of users.

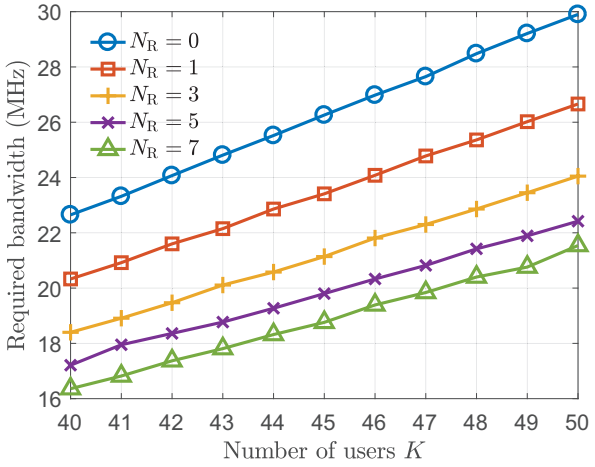

Fig. 4. Total bandwidth required to ensure QoS vs. number of users.

probability that a packet both undergoes deep fading during transmission and waits in long queue is low. When a packet needs retransmission, the queue for the packet is more likely empty, hence more transmission delay is allowed without queueing delay. This suggests that pre-assigning a queueing delay requirement according to the E2E delay requirement is too conservative in URLLC.

Fig. 3 shows the DL packet loss probabilities with different number of users. The results of $N_R = 2, 4, 6$ are similar, hence are omitted. The results are obtained via simulations over 10^7 Rayleigh fading channel realizations. The maximum numbers of users can be supported with QoS guarantee, which reflects the system capacity, are pointed out by arrows. It can be found that with given system resources, the capacity increases with N_R. Compared with the non-retransmission policy ($N_R = 0$), the capacity gain is 30% with the seven-times retransmission policy ($N_R = 7$).

Fig. 4 shows the total bandwidth required to ensure the QoS. The results show that less bandwidth is required to serve a given number of users with more retransmissions. Around 30% of bandwidth can be saved with the seven-times retransmission policy compared with the non-retransmission policy.

VII. Conclusion

In this paper we investigated how to ensure the stringent QoS of URLLC by a retransmission policy with frequency hopping, and optimized the resource allocation for the policy. Numerical results showed that with the optimized retransmission policy, the requirement for transmission error probability can be relaxed significantly, and the transmission delay and queueing delay can be traded off by retransmission. Simulation results showed that the performance gain of retransmission increases with the times of retransmission. Compared with the non-retransmission policy, a 30% gain on the maximal number of users with QoS guarantee can be achieved with seven-times retransmission policy when the resources are given, and around 30% of bandwidth can be saved when the number of users is given. These results indicate that a proper designed retransmission policy has large potential for URLLC, thanks to its traffic with short packets.

References

[1] 3GPP, *Study on Scenarios and Requirements for Next Generation Access Technologies*. Technical Specification Group Radio Access Network, Technical Report 38.913, Release 14, Oct. 2016.

[2] M. Simsek, A. Aijaz, M. Dohler, *et al.*, "5G-enabled tactile internet," *IEEE J. Select. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[3] M. Serror, C. Dombrowski, K. Wehrle, and J. Gross, "Channel coding versus cooperative ARQ: Reducing outage probability in ultra-low latency wireless communications," in *IEEE Globecom Workshop*, 2015.

[4] W. Yang, G. Durisi, T. Koch, *et al.* , "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4264, Jul. 2014.

[5] H. Shariatmadari, S. Iraji, Z. Li, M. A. Uusitalo, and R. Jantti, "Optimized transmission and resource allocation strategies for ultra-reliable communications," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016.

[6] E. Dosti, U. L. Wijewardhana, H. Alves, and M. Latva-aho, "Ultra reliable communication via optimum power allocation for type-i arq in finite block-length," in *2017 IEEE International Conference on Communications (ICC)*, 2017.

[7] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, Jan 2018.

[8] C. Sun, C. She, and C. Yang, "Exploiting muli-user diversity for ultra-reliable and low-latency communications," in *Proc. IEEE Globecom Workshop*, 2017.

[9] ——, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *Proc. IEEE Global Commun. Conf. (Globecom)*, 2017, https://arxiv.org/pdf/1707.09720.pdf.

[10] P. Kela, J. Turkka, *et al.*, "A novel radio frame structure for 5G dense outdoor radio access networks," in *Proc. IEEE VTC Spring*, 2015.

[11] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. ACM MSWiM*, 2015.

[12] C. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[13] M. Khabazian, S. Aissa, and M. Mehmet-Ali, "Performance modeling of safety messages broadcast in vehicular ad hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 380–387, Mar. 2013.

[14] G. R1-120056, "Analysis on traffic model and characteristics for MTC and text proposal." Technical Report, TSG-RAN Meeting WG1#68, Dresden, Germany, 2012. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R1-68--29000.htm

[15] S. Boyd and L. Vandanberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[16] G. Zhang, T. Q. S. Quek, M. Kountouris, A. Huang, and H. Shan, "Fundamentals of heterogeneous backhaul designłanalysis and optimization," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 876–889, Feb 2016.