Exploiting Multi-user Diversity for Ultra-reliable and Low-latency Communications

Chengjian Sun, Changyang She and Chenyang Yang
School of Electronics and Information Engineering,
Beihang University, Beijing, China
Email: {sunchengjian,cyshe,cyyang}@buaa.edu.cn

***Abstract*—In this paper, we study how to exploit multi-user diversity for ultra-reliable and low-latency communications. The basic idea is that the users with good channel conditions share resources with the users with bad channel conditions. We propose a method to optimize resource allocation among multiple users under the quality-of-service (QoS) constraints, including transmission delay, transmission error probability, queueing delay bound and queueing delay violation probability. If the minimal transmit power required to satisfy these constraints is less than the total transmit power of the base station, then the global optimal resource allocation policy can be obtained. Otherwise, some packets are dropped proactively. Simulation results show that compared with an existing policy that does not exploit multi-user diversity, the proposed policies can double the number of users with QoS guarantee or reduce proactive packet dropping probability remarkably with given total bandwidth and transmit power. This indicates that with multi-user diversity, a better tradeoff between reliability and throughput (or spectral efficiency) can be achieved.**

***Index Terms*—Ultra-reliable and low-latency communications, resource allocation, multi-user diversity, short blocklength**

I. Introduction

Together with enhanced mobile broadband and massive machine-type communications, ultra-reliable and low-latency communications (URLLC) has become one kind of the new application scenarios in the fifth generation (5G) cellular networks [1]. With stringent requirements on reliability and latency, URLLC is crucial to enable future mission critical applications in tactile internet, autonomous vehicle networks, and smart factories [2].

To achieve ultra-high reliability, different kinds of diversities are essential [3–7]. As shown in [3], both macro- and micro-diversity are helpful for reducing outage probability (i.e., the probability that the signal-to-noise ratio (SNR) is lower than a required threshold). Frequency diversity was studied in [4], where the number of Rayleigh-fading links is optimized to minimize the required transmit power under the constraint on reliability. Studies in [5] show that by serving one user with multiple base stations (BSs), the outage probability can be reduced. To exploit spatial diversity, multiple antenna systems are considered in [5]. The reliability with cooperative Automatic Repeat reQuest (ARQ) was studied in machine-to-machine communications [6]. When a direct transmission between two nodes fails, a BS will retransmit the packet. The results in [6] indicate that benefitting from macro-diversity, the reliability can be improved with cooperative ARQ. More recently, how to trade-off bandwidth usage and reliability with path diversity was studied in [7], where different paths use different communication interfaces (e.g., wired communications, cellular links, and machine-to-machine communications).

In order to achieve ultra-high reliability with low-latency, more resources are required [3–7]. Studies in [8] show that there is a tradeoff among reliability and throughput given latency requirement. It is nature to raise the following question: can we improve the tradeoff between reliability and throughput with given latency requirement? As shown in [9], with multi-user diversity, throughput can be improved without extra resources. This offers us a chance to trade throughput for reliability. Since studies in [3–8] focus on single-user case, how to exploit multi-user diversity for URLLC remains unclear.

In URLLC, packet size could be short (e.g., 20 bytes) [1]. To transmit a short packet within a small transmission duration, the blocklength of channel codes is short, and Shannon's capacity cannot be used to characterize the achievable rate of URLLC [10]. To achieve a given block error probability, an accurate approximation of the achievable rate in short blocklength regime has been derived in [11]. Different from Shannon's capacity, the achievable rate in finite blocklength regime is not jointly concave in bandwidth and transmit power. Therefore, applying the achievable rate in finite blocklength regime for resource management for URLLC is very challenging [12].

In this paper, we study how to exploit multi-user diversity for URLLC. The basic idea is that the users with good channel conditions share resources with the users with bad channel conditions to ensure the reliability and latency of all users. To this end, bandwidth and transmit power are jointly allocated to multiple users according to their channel gains. When the constraints on transmission delay, transmission error probability, queueing delay bound and queueing delay violation probability can be satisfied, the global optimal resource allocation policy is obtained. When these constraints cannot be satisfied, to reduce packet dropping probability, some packets are dropped proactively. Simulation results show that by exploiting multi-user diversity, we can improve the tradeoff between reliability and throughput (i.e., the maximal number of users with quality-of-service (QoS) guarantee) with given latency requirement, or we can save bandwidth given the number of users and the requirements on latency and reliability. This indicates that

a better tradeoff between reliability and spectral efficiency (defined as the ratio of throughput to bandwidth) can be achieved with multi-user diversity.

II. System Models

A. System and Traffic Models

Consider a downlink cellular network, where each BS with $N_{\rm t}$ antennas serves K single-antenna users with bandwidth $W_{\max}$. The maximal transmit power of each BS is $P_{\max}$. Frequency division multiple access is applied to avoid multi-user interference, and different frequency bands are used in adjacent cells to avoid strong inter-cell interference.

Time is discretized into frames. The duration of each frame is $T_{\rm f}$. Time division duplex is adopted [13]. The duration for downlink transmission in one frame is τ.

Packets for every user arrive at the buffer of the BS randomly, and the inter-arrival time between packets could be shorter than the service time (i.e., transmission duration) of each packet. Hence, the queueing delay cannot be ignored, and the probability that the queueing delay exceeds a required bound should be controlled. We consider a queueing model that the packets for different users are waiting in different queues. Denote the required queueing delay bound for guaranteeing end to end (E2E) delay as $D^{\rm q}$, and the probability that the queueing delay violates the delay bound as $\varepsilon^{\rm q}$.

B. Achievable Rate with Short Blocklength

In URLLC, the blocklength of channel coding is short due to the required low-latency. Hence, the impact of transmission error on reliability cannot be ignored. The Shannon's Capacity cannot be applied to characterize the probability of transmission error $\varepsilon^{\rm c}$ [14]. 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 (e.g., 20 bytes [1]), 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 rate of the kth user can be accurately approximated by [11],

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

where 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 [11],

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

Although the achievable rate is in closed-form, it is still too complicated to obtain graceful results. In this work, we focus on high SNR regime, which is required to ensure ultra-high reliability and ultra-low latency. As shown in [14], 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 r_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).

C. Quality-of-service

The QoS of URLLC can be characterized by the E2E delay requirement $D_{\max}$ and overall packet loss probability in downlink transmission $\varepsilon^{\rm D}_{\max}$.

Assume that uplink and downlink transmissions can be finished in one frame, and the backhaul latency is $T_{\rm f}$. Then, the E2E delay requirement can be satisfied when

$$D^{\rm q} = D_{\max} - 2T_{\rm f}. \quad (3)$$

When the queueing delay is shorter than the channel coherence time, the service rate is constant within the queueing delay bound of each packet. To ensure the queueing delay requirement $(D^{\rm q}, \varepsilon^{\rm q})$, the constant service rate should be equal to or higher than the effective bandwidth of the arrival process [15]. For a Poisson process with arrival packet rate λ, the effective bandwidth can be derived as [16],

$$E^{\rm B} = \frac{uT_{\rm f}\ln\left(1/\varepsilon^{\rm q}\right)}{D^{\rm q}\ln\left(1+\frac{T_{\rm f}\ln\left(1/\varepsilon^{\rm q}\right)}{\lambda D^{\rm q}}\right)}, \quad (4)$$

where u is the number of bits contained in each packet. By substituting (1) and (4) into $r_k \geq E^{\rm B}$, we can obtain the required transmit power to ensure $\varepsilon^{\rm c}$ and $(D^{\rm q}, \varepsilon^{\rm q})$, i.e.,

$$P_k \geq \frac{N_0 W_k}{\alpha_k g_k}\left\{\exp\left[\frac{E^{\rm B}\ln 2}{\tau W_k}+\frac{Q_{\rm G}^{-1}(\varepsilon^{\rm c})}{\sqrt{\tau W_k}}\right]-1\right\} \triangleq y_k(W_k, E^{\rm B}). \quad (5)$$

As shown in [16], for Rayleigh fading channel or Nakagami-m fading channel, P_k is unbounded when the small-scale channel gain is close to zero. This is because $y_k(W_k, E^{\rm B})$ is proportional to the inverse of the channel gain. Thus, the total transmit power required to ensure $\varepsilon^{\rm c}$ and $(D^{\rm q}, \varepsilon^{\rm q})$ may exceed the total transmit power of the BS $P_{\max}$. To deal with this issue, a proactive packet dropping mechanism was proposed in [16]. The proactive packet dropping probability is denoted as $\varepsilon^{\rm p}$. To ensure the overall packet loss probability, the three packet loss components should satisfy

$$\varepsilon^{\rm c} + \varepsilon^{\rm q} + \varepsilon^{\rm p} \leq \varepsilon^{\rm D}_{\max}. \quad (6)$$

III. Existing Resource Allocation Policy

In this section, we briefly summarize the policy in [16], which was proposed in single-user case and extended to multi-user case without exploiting multi-user diversity.

To satisfy the total transmit power constraint, the policy in [16] divides $P_{\max}$ into K parts that satisfy $\sum_{k=1}^{K} P_k^{\rm th} \leq P_{\max}$. The transmit power allocated to the kth user cannot exceed the threshold $P_k^{\rm th}$. By optimizing $P_k^{\rm th}$ and W_k according to channel distribution and QoS requirements, $P_k^{\rm th}$ and W_k are

reserved for the kth user, and cannot be shared to other users in deep fading. As a result, multi-user diversity is not exploited in this policy.

To ensure the QoS requirement of each user, the transmit power and the number of packets that are dropped proactively were optimized according to channel fading in [16]. Denote the number of packets for the kth user that are proactively dropped in a frame as d_k. To ensure $(D^{\mathrm{q}}, \varepsilon^{\mathrm{q}})$, the required service rate is E^{B}. Since the data dropping rate is $d_k u$, the data rate transmitted to the user is $E^{\mathrm{B}} - d_k u$. To ensure the transmission error probability ε^{c}, the required transmit power of the kth user is

$$P_k \geq y_k(W_k, E^{\mathrm{B}} - d_k u). \tag{7}$$

It is not hard to see that $y_k(W_k, E^{\mathrm{B}} - d_k u)$ decreases with d_k and g_k. If $y_k(W_k, E^{\mathrm{B}}) \leq P_k^{\mathrm{th}}$, then all the packets for the kth user in the current frame can be transmitted with $P_k \leq P_k^{\mathrm{th}}$. Otherwise, the BS needs to increase d_k until $y_k(W_k, E^{\mathrm{B}} - d_k u) \leq P_k^{\mathrm{th}}$.

$P_k > P_k^{\mathrm{th}}$ does not equivalent to $\sum_{k=1}^{K} P_k > P_{\max}$. This is because some other users with good channel conditions may only need little transmit power. As a result, $\sum_{k=1}^{K} P_k$ could be less than the total transmit power of the BS. If the users with good channels can share resources to users in deep fading, the proactive packet dropping probability can be reduced.

IV. Resource Allocation Policies with Multiuser Diversity

In this section, we propose two different resource allocation policies with multiuser diversity.

A. Policy A

Policy A includes resource allocation and proactive packet dropping. In order to exploit the multi-user diversity, resource allocation should be optimized based on the channel gains of all users. Packets are dropped proactively when $\sum_{k=1}^{K} P_k > P_{\max}$. In what follows, we design a policy to minimize the probability that proactive packet dropping happens, i.e., $\Pr\{\sum_{k=1}^{K} P_k > P_{\max}\}$. The required total transmit power depends on the channel gains. To minimize $\Pr\{\sum_{k=1}^{K} P_k > P_{\max}\}$, we need to minimize $\sum_{k=1}^{K} P_k$ for any given α_k and g_k, $k = 1, \cdots, K$.

To this end, we optimize W_k and P_k to minimize the total transmit power required to ensure ε^{c} and $(D^{\mathrm{q}}, \varepsilon^{\mathrm{q}})$, i.e.,

$$\min_{P_k, W_k} \quad \sum_{k=1}^{K} P_k \tag{8}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} W_k \leq W_{\max}, \tag{9}$$

$$(5),\ P_k > 0,\ W_k > 0,\ k = 1, 2, \cdots, K,$$

where the constraint on total transmit power is removed.

It should be noticed that, $y_k(W_k, E^{\mathrm{B}})$ in constraint (5) is non-convex in W_k, which results from the fact that the achievable rate in (1) is non-concave. Thus, problem (8) is non-convex, and finding the global optimal solution is very challenging. To overcome this difficulty, we first show a property of $y_k(W_k, E^{\mathrm{B}})$ [17].

Property 1. There is a unique solution W_k^{th} that minimizes $y_k(W_k, E^{\mathrm{B}})$. Moreover, $y_k(W_k, E^{\mathrm{B}})$ is convex in W_k when $0 < W_k \leq W_k^{\mathrm{th}}$.

Based on Property 1, we have the following Property (See proof in Appendix A),

Property 2. The global optimal solution of problem (8) satisfies

$$W_k \leq W_k^{\mathrm{th}},\ k = 1, 2, \cdots, K. \tag{10}$$

According to Property 2, problem (8) is equivalent to the following problem,

$$\min_{P_k, W_k} \quad \sum_{k=1}^{K} P_k \tag{11}$$

$$\text{s.t.} \quad (5), (9), (10),\ P_k > 0,\ W_k > 0,\ k = 1, 2, \cdots, K,$$

which is convex according to Property 1, and can be solved by interior-point method [18].

Denote the required minimal total transmit power as $P_{\mathrm{A}}^*(\varepsilon^{\mathrm{c}}, \varepsilon^{\mathrm{q}})$. When $P_{\mathrm{A}}^*(\varepsilon^{\mathrm{c}}, \varepsilon^{\mathrm{q}}) \leq P_{\max}$, the global optimal bandwidth and transmit power allocation is obtained by solving problem (8). When $P_{\mathrm{A}}^*(\varepsilon^{\mathrm{c}}, \varepsilon^{\mathrm{q}}) > P_{\max}$, some packets are dropped proactively according to the following policy.

Dropping packets to the users with the worst channel gains can help reduce proactive packet dropping probability. However, the fairness among users is not considered. To avoid that the user with the worst large-scale channel gain has the highest packet dropping probability, Policy A drops packets based on the small-scale channel gains of the users. Thus, Policy A can improve fairness by sacrificing proactive packet dropping probability. Without loss of generality, we assume $g_k < g_{k+1}$ for all $k = 1, \cdots, K-1$, i.e., the small-scale channel gains of the users increase with the indices. When $P_{\mathrm{A}}^*(\varepsilon^{\mathrm{c}}, \varepsilon^{\mathrm{q}}) > P_{\max}$, the packet dropping policy can be obtained from Algorithm 1, where the packets to the users with the worst small-scale channel gains are discarded sequentially until $\sum_{k=1}^{K} y_k(W_k, E^{\mathrm{B}} - d_k u) \leq P_{\max}$. Given a packet dropping policy $\{d_k, k = 1, \cdots, K\}$, the resource allocation $\{P_k, W_k, k = 1, \cdots, K\}$ can be obtained by solving problem (11), where constraint (5) is replaced by (7).

B. Policy B

Policy A needs to optimize bandwidth and transmit power allocation when the small-scale channel gains change. Thus, it requires high computing resource and may lead to extra computing delay. To reduce the required computing resource, we propose another resource allocation and packet dropping policy (i.e., Policy B) that only needs to optimize bandwidth allocation when large-scale channel gains change. To satisfy ε^{c} and $(D^{\mathrm{q}}, \varepsilon^{\mathrm{q}})$, more bandwidth will be allocated to the users with smaller α_k. With Policy B, the transmit power is set

Algorithm 1 Proactive packet dropping with Policy A.

Input: α_k, g_k, W_k^{th}, $k=1,\cdots,K$, u, N_0, τ, $W_{\max}$, ε^{c}, E^{B} with λ, $D^{\mathrm{q}}=D_{\max}-2T_{\mathrm{f}}$ and ε^{q};

Output: Number of discarded packets $d_k, k=1,\cdots,K$ and resource allocation W_k, P_k, $k=1,\cdots,K$;

$d_k \leftarrow 0, k=1,\cdots,K$, $n \leftarrow 1$;
while $n \le K$ **do**
  while $d_n u < E^{\mathrm{B}}$ **do**
    Get W_k, P_k, $k=n,\cdots,K$ by solving problem (11) with interior-point method, where constraint (5) is replaced by (7);
    if $\sum_{k=n}^{K} P_k > P_{\max}$ **then**
      $d_n \leftarrow d_n + 1$;
    else
      return $d_k, k=1,\cdots,K$;
    end if
  end while
  $W_n \leftarrow 0$, $P_n \leftarrow 0$;
  $n \leftarrow n+1$;
end while

as $P_k = y_k(W_k, E^{\mathrm{B}})$, and W_k is optimized to minimize the expectation of total transmit power,

$$\min_{W_k} \quad \sum_{k=1}^{K} \mathbb{E}_g\left\{y_k(W_k, E^{\mathrm{B}})\right\} \tag{12}$$
$$\text{s.t.} \quad (9),\ W_k > 0,\ k=1,2,\cdots,K,$$

where the average is taken over the small-scale channel gains. Since $y_k(W_k, E^{\mathrm{B}})$ is proportional to $1/g_k$, $\mathbb{E}_g\left\{y_k(W_k, E^{\mathrm{B}})\right\}$ is proportional to $\mathbb{E}_g\{1/g_k\}$, which is a constant with given channel distribution. As a result, Property 1 is also applicable to $\mathbb{E}_g\left\{y_k(W_k, E^{\mathrm{B}})\right\}$, and Property 2 can also be applied in solving problem (12). Then, the optimal bandwidth allocation W_k^* can be obtained by solving the following convex problem,

$$\min_{W_k} \quad \sum_{k=1}^{K} \mathbb{E}_g\left\{y_k(W_k, E^{\mathrm{B}})\right\} \tag{13}$$
$$\text{s.t.} \quad (9), (10),\ W_k > 0,\ k=1,2,\cdots,K.$$

Although the bandwidth allocation of Policy B does not depend on small-scale channel gains, the total transmit power can be shared among all users. Hence, multi-user diversity is also exploited in Policy B. The required minimal total transmit power of Policy B is denoted as $P_{\mathrm{B}}^*(\varepsilon^{\mathrm{c}}, \varepsilon^{\mathrm{q}})$. Similar to Policy A, if $P_{\mathrm{B}}^*(\varepsilon^{\mathrm{c}}, \varepsilon^{\mathrm{q}}) > P_{\max}$, then packets to the users with worst small-scale channel gains are discarded sequentially until $\sum_{k=1}^{K} y_k(W_k^*, E^{\mathrm{B}} - d_k u) \le P_{\max}$.

V. Simulation and Numerical Results

In this section we first evaluate packet dropping probabilities of different policies. Then, we show the required total bandwidth with different number of users.

Proactive Packet Dropping Policies: The existing proactive packet dropping policy proposed in [16] is referred to as Policy C here. The proactive packet dropping probabilities of Policies A, B and C are denoted as $\varepsilon_{\mathrm{A}}^{\mathrm{p}}$, $\varepsilon_{\mathrm{B}}^{\mathrm{p}}$ and $\varepsilon_{\mathrm{C}}^{\mathrm{p}}$, respectively.

Reactive Packet Dropping Policies: In URLLC, it may sound counterintuitive to proactively drop packets since the required packet loss probability is very low. To understand how these proactive packet dropping policies behave, we compare them with three corresponding reactive packet dropping policies in our simulation, which are referred to as Policies $\tilde{\mathrm{A}}$, $\tilde{\mathrm{B}}$ and $\tilde{\mathrm{C}}$. With Policy $\tilde{\mathrm{A}}$ (or Policy $\tilde{\mathrm{B}}$), all packets transmitted in the current frame will be dropped if the required minimal total power of Policy A (or Policy B) exceeds $P_{\max}$. With Policy $\tilde{\mathrm{C}}$, all packets to the kth user will be dropped if $P_k > P_k^{\mathrm{th}}$. The difference between Policy $\tilde{\mathrm{A}}$ (or Policies $\tilde{\mathrm{B}}$ and $\tilde{\mathrm{C}}$) and Policy A (or Policies B and C) lie in the number of packets that are dropped when the required total transmit power exceeds $P_{\max}$. The packet dropping probabilities of Policies $\tilde{\mathrm{A}}$, $\tilde{\mathrm{B}}$ and $\tilde{\mathrm{C}}$ are denoted as $\varepsilon_{\tilde{\mathrm{A}}}^{\mathrm{p}}$, $\varepsilon_{\tilde{\mathrm{B}}}^{\mathrm{p}}$ and $\varepsilon_{\tilde{\mathrm{C}}}^{\mathrm{p}}$, respectively.

We consider a single cell scenario, where all users are located at the edge of the cell (i.e., the worst case). The user-BS distance is 250 m. Rayleigh fading channel is considered. Other parameters are listed in Table I, unless otherwise specified.

TABLE I
Simulation Parameters

Overall packet loss probability $\varepsilon_{\max}^{\mathrm{D}}$	10^{-5}
Queueing delay requirement D^{q}	0.8 ms
Duration of each frame T_{f}	0.1 ms
Duration of DL transmission τ	0.05 ms
Maximal transmit power of BS $P_{\max}$	43 dBm
Single-sided noise spectral density N_0	-173 dBm/Hz
Packet size u	20 bytes (160 bits) [1]
Arrival packet rate λ	0.2 packets/frame
Path loss model $10\lg(\alpha_k)$	$35.3 + 37.6\lg(d_k)$

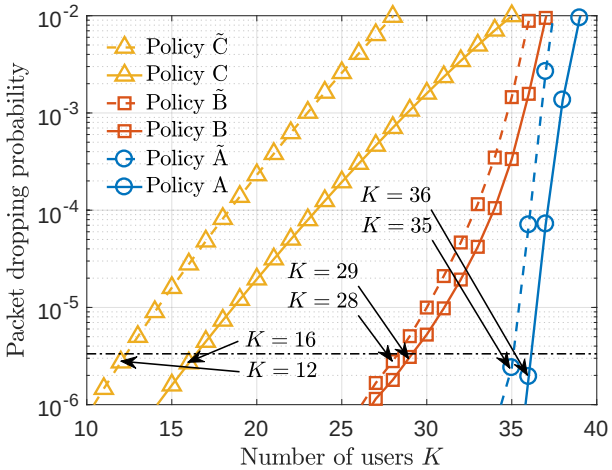

Fig. 1. Packet dropping probabilities vs. the number of users, $N_{\mathrm{t}} = 4$ and $W_{\max} = 20$ MHz.

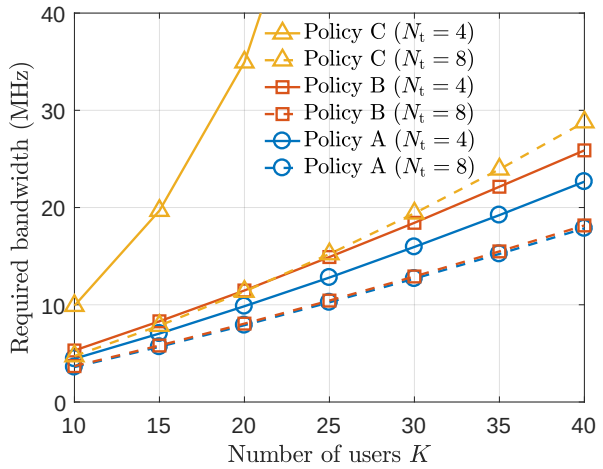

Fig. 2. The bandwidth required to ensure QoS vs. the number of users.

The packet dropping probabilities versus number of users are shown in Fig. 1. $\varepsilon_{\mathrm{A}}^{\mathrm{p}}$, $\varepsilon_{\mathrm{B}}^{\mathrm{p}}$, $\varepsilon_{\tilde{\mathrm{A}}}^{\mathrm{p}}$ and $\varepsilon_{\tilde{\mathrm{B}}}^{\mathrm{p}}$ are obtained by simulation; $\varepsilon_{\mathrm{C}}^{\mathrm{p}}$ and $\varepsilon_{\tilde{\mathrm{C}}}^{\mathrm{p}}$ are computed numerically according to the results in [16]. To ensure the overall reliability requirement in (6), we set $\varepsilon^{\mathrm{c}} = \varepsilon^{\mathrm{q}} = \varepsilon_{\max}^{\mathrm{D}}/3$, and study the maximal number of users that can be served under constraints that $\varepsilon_{\mathrm{A}}^{\mathrm{p}}$, $\varepsilon_{\mathrm{B}}^{\mathrm{p}}$, $\varepsilon_{\mathrm{C}}^{\mathrm{p}}$, $\varepsilon_{\tilde{\mathrm{A}}}^{\mathrm{p}}$, $\varepsilon_{\tilde{\mathrm{B}}}^{\mathrm{p}}$ and $\varepsilon_{\tilde{\mathrm{C}}}^{\mathrm{p}}$ are less than $\varepsilon_{\max}^{\mathrm{D}}/3$.

It is shown that given the number of users, $\varepsilon_{\mathrm{A}}^{\mathrm{p}} \ll \varepsilon_{\mathrm{B}}^{\mathrm{p}} \ll \varepsilon_{\mathrm{C}}^{\mathrm{p}}$ and $\varepsilon_{\tilde{\mathrm{A}}}^{\mathrm{p}} \ll \varepsilon_{\tilde{\mathrm{B}}}^{\mathrm{p}} \ll \varepsilon_{\tilde{\mathrm{C}}}^{\mathrm{p}}$. The maximum numbers of users that can be served with QoS guarantee are pointed out by arrows. These results indicate that by exploiting multi-user diversity, the number of users with QoS guarantee can be doubled or the packet dropping can be reduced remarkably with given total bandwidth and transmit power.

For Policies A, B and C, given the same number of users K, the proactive packet dropping probabilities are much lower than the reactive packet dropping probabilities with Policies $\tilde{\mathrm{A}}$, B and $\tilde{\mathrm{C}}$, respectively. This is because when the required minimal total transmit power exceeds the maximum transmit power of the BS, only a small portion of packets are discarded with proactive packet dropping policies.

The total bandwidth requirements with QoS guarantee versus number of users are shown in Fig. 2. It is shown that Policies A and B can save a large amount of bandwidth compared with Policy C, benefiting from the multi-user diversity. Given the total bandwidth, the number of users able to be served with Policy B is $10 \sim 20\%$ less than that with Policy A for the case of $N_{\mathrm{t}} = 4$. However, the gap between Policy A and Policy B shrinks when $N_{\mathrm{t}} \geq 8$. This indicates that there is no need to adjust bandwidth according to small-scale channel gains when the number of antennas is large.

From Figs. 1 and 2, we can see that proactive packet dropping probability is very sensitive to the number of users, and the required bandwidth grows almost linearly with number of users. As a result, the proactive packet dropping probability decreases very fast as the total bandwidth increases. For a given number of users, we can improve reliability significantly by increasing a small amount of bandwidth. Compared with Policy C, Policy A and Policy B can achieve a better tradeoff between reliability and throughput (or spectral efficiency, defined as the ratio of throughput to bandwidth) given latency requirement.

VI. Conclusion

In this paper, we studied how to exploit multi-user diversity for URLLC. The optimal bandwidth and transmit power allocation policy was obtained, which can minimize the probability that the required minimal transmit power exceeds the total transmit power of the BS. Simulation results showed that by exploiting multi-user diversity, the number of users with QoS guarantee can be increased significantly given total bandwidth and transmit power, compared with an existing policy. We also compared the proactive packet dropping with the reactive packet dropping. The results show that serving the same number of users with the given resources, proactive packet dropping can achieve a much lower packet dropping probability than reactive packet dropping. Moreover, the proactive packet dropping probability can be reduced remarkably by increasing total bandwidth slightly. This means that with multi-user diversity, a better tradeoff between reliability and throughput (or spectral efficiency) can be achieved with given latency requirement.

Appendix A
Proof of Property 2

Proof. Denote the feasible solutions of problem (8) that do not satisfy condition (10) as,

$$\widetilde{\boldsymbol{x}} \triangleq (\widetilde{W}_1, \cdots, \widetilde{W}_K, \widetilde{P}_1, \cdots, \widetilde{P}_K),$$
$$\text{where } \widetilde{W}_j > W_j^{\mathrm{th}} \text{ for } j \in \mathcal{J}.$$

To prove the property, for any $\widetilde{\boldsymbol{x}}$, we construct another feasible solution of problem (8) that satisfies condition (10) and requires less transmit power than $\widetilde{\boldsymbol{x}}$, denoted as $\widetilde{\boldsymbol{x}}^a \triangleq (\widetilde{W}_1^a, \cdots, \widetilde{W}_K^a, \widetilde{P}_1^a, \cdots, \widetilde{P}_K^a)$.

In the following we show that $\widetilde{\boldsymbol{x}}^a$ can be obtained by replacing $\widetilde{W}_j$ and $\widetilde{P}_j$, $j \in \mathcal{J}$ in $\widetilde{\boldsymbol{x}}$ with W_j^{th} and $y_j(W_j^{\mathrm{th}})$, respectively.

Since $\widetilde{\boldsymbol{x}}$ is a feasible solution, we have $\sum_{k=1}^K \widetilde{W}_k \leq W_{\max}$. Then, from $\widetilde{W}_j^a \triangleq W_j^{\mathrm{th}} < \widetilde{W}_j$, we know that $\sum_{k=1}^K \widetilde{W}_k^a < \sum_{k=1}^K \widetilde{W}_k \leq W_{\max}$. Therefore, constraint (9) is satisfied by $\widetilde{\boldsymbol{x}}^a$. It is not hard to find that the constraints in (5) are also satisfied by $\widetilde{\boldsymbol{x}}^a$. Hence, $\widetilde{\boldsymbol{x}}^a$ is indeed a feasible solution of problem (8).

Since $\widetilde{\boldsymbol{x}}$ is a feasible solution, we have $\widetilde{P}_j \geq y_j(W_j; E^{\mathrm{B}})$. According to Property 1, $y_j(W_j; E^{\mathrm{B}})$ is minimized when $W_j = W_j^{\mathrm{th}}$. As a result, $\widetilde{P}_j^a \triangleq y_j(W_j^{\mathrm{th}}; E^{\mathrm{B}}) < y_j(W_j; E^{\mathrm{B}}) \leq \widetilde{P}_j^a$. Thus, $\sum_{k=1}^K \widetilde{P}_k^a < \sum_{k=1}^K \widetilde{P}_k$.

Now, we have shown that the feasible solution $\widetilde{\boldsymbol{x}}^a$ satisfying condition (10) requires less transmit power than $\widetilde{\boldsymbol{x}}$. This suggests that all solutions that do not satisfy condition (10) are not optimal, i.e., the optimal solution must satisfy condition (10). This completes the proof. □

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] G. Pocovi, B. Soret, M. Lauridsen, *et al.*, "Signal quality outage analysis for ultra-reliable communications in cellular networks," in *IEEE Globecom Workshops*, 2015.

[4] D. Ohmann, M. Simsek, and G. P. Fettweis, "Achieving high availability in wireless networks by an optimal number of Rayleigh-fading links," in *IEEE Globecom Workshop*, 2014.

[5] F. Kirsten, D. Ohmann, M. Simsek, and G. P. Fettweis, "On the utility of macro- and microdiversity for achieving high availability in wireless networks," in *Proc. IEEE PIMRC*, 2015.

[6] 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.

[7] J. J. Nielsen and P. Popovski, "Latency analysis of systems with multiple interfaces for ultra-reliable M2M communication," in *Proc. IEEE SPAWC*, 2016.

[8] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *IEEE Globecom Workshop*, Dec. 2014.

[9] D. Tse, *Fundamentals of Wireless Communication*. Cambridge Univ. Press, 2005.

[10] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Aug. 2016.

[11] 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.

[12] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

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

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

[15] 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.

[16] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun., revised*, https://arxiv.org/pdf/1703.09575.pdf.

[17] C. Sun, C. She, and C. Yang, "Energy-Efficient resource allocation for ultra-reliable and low-latency communications," in *IEEE Globecom*, 2017.

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