# Optimizing Resource Allocation in the Short Blocklength Regime for Ultra-Reliable and Low-Latency Communications

Chengjian Sun, *Student Member, IEEE*, Changyang She, *Member, IEEE*, Chenyang Yang, *Senior Member, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, Yonghui Li, *Senior Member, IEEE*, and Branka Vucetic, *Fellow, IEEE*

***Abstract*—In this paper, we aim to find the global optimal resource allocation for ultra-reliable and low-latency communications (URLLC), where the blocklength of channel codes is short. The achievable rate in the short blocklength regime is neither convex nor concave in bandwidth and transmit power. Thus, a non-convex constraint is inevitable in optimizing resource allocation for URLLC. We first consider a general resource allocation problem with constraints on the transmission delay and decoding error probability, and prove that a global optimal solution can be found in a convex subset of the original feasible region. Then, we illustrate how to find the global optimal solution for an example problem, where energy efficiency (EE) is maximized by optimizing antenna configuration, bandwidth allocation, and power control under the latency and reliability constraints. To improve the battery life of devices and EE of communication systems, both uplink and downlink resources are optimized. Simulation and numerical results validate the analysis and show that the circuit power is dominated by the total power consumption when the average inter-arrival time between packets is much larger than the required delay bound. Therefore, optimizing antenna configuration and bandwidth allocation without power control leads to minor EE loss.**

***Index Terms*—Ultra-reliable and low-latency communications, resource allocation, energy efficiency.**

## I. Introduction

Ultra-reliable and low-latency communications (URLLC) has been considered as one of the new application scenarios in the fifth generation (5G) cellular networks [2]. It is crucial for enabling mission-critical applications such as autonomous vehicle communications, factory automation and haptic communications [3, 4]. To ensure the stringent end-to-end (E2E) delay including transmission delay, coding and processing delay, queueing delay, and routing delay in backhaul and core networks, it is necessary to develop new enabling transmission algorithms, network architecture and protocols [5].

In the Long Term Evolution systems, the minimal time unit for resource allocation is 1 ms [6], and hence cannot satisfy the E2E delay requirement of URLLC. To reduce latency, one of the key major reforms will be to employ a short frame structure [7]. With the short frame structure and short packets in URLLC, the blocklength of channel codes is short. The way to characterize the relationship among achievable rate, decoding error probability, and transmission delay in the short blocklength regime is fundamentally different from that in the long blocklength regime, where Shannon's capacity is applied [8]. If Shannon's capacity is used in designing resource allocation or analyzing the performance of URLLC, the reliability and latency will be underestimated [9], and hence the quality-of-service (QoS) cannot be satisfied.

### A. Related Works

Transmission scheme and resource allocation for URLLC have been studied in existing literatures [10–12]. Outage probability was utilized in these works to characterize the relationship among data rate, reliability and delay. The basic assumption is that if the signal-to-noise-ratio (SNR) is higher than a threshold such that Shannon's capacity is higher than the required data rate, then the packets are transmitted successfully. Otherwise, the transmission fails, and the packets are lost. Since the blocklength is short in URLLC, the decoding error probability is non-zero for arbitrarily high SNR. As a result, the existing outage probability based on Shannon's capacity cannot be applied in the short blocklength regime. The decoding error probability in the short blocklength regime should be taken into account when formulating a constraint on the reliability.

In a seminal work in [13], the authors derived an accurate approximation on the maximal achievable rate in the short blocklength region over AWGN channel. The result was extended to multiple antenna systems over the quasi-static channel in [14]. The maximal achievable rates obtained in [13,14] are inevitable in portraying decoding error probability, and have been applied in spectrum sharing networks [15], relay systems [16], energy-efficient packet scheduling [17], and cross-layer resource allocation [18]. However, the expression of the achievable rate is

This paper has been presented in part at the IEEE Global Communications Conference 2017 [1].

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 was with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372. He is now with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia. (e-mail: shechangyang@gmail.com).

T. Q. S. Quek is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

Y. Li and B. Vucetic are with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia (e-mail:{yonghui.li,branka.vucetic}@sydney.edu.au).

The work of C. Sun and C. Yang was supported by National Natural Science Foundation of China (NSFC) under Grant 61731002. The work of C. She and B. Vucetic was supported by ARC Laureate Fellowship under Grant FL160100032. This work of T. Q. S. Quek was supported in part by the MOE ARF Tier 2 under Grant MOE2015-T2-2-104, and the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/01/2016. The work of Y. Li was supported by ARC under Grant DP150104019.

still complex, which is neither convex nor concave with respect to radio resources [17]. As a result, how to obtain the optimal solution when characterizing the QoS requirements of URLLC with such an achievable rate remains as an open challenge [19].

Resource allocation is usually optimized towards improving spectrum efficiency or energy efficiency (EE), which are important metrics in 5G communications [20]. Energy efficient resource allocation has been extensively studied for traditional best effort or real-time services. In the typical URLLC scenarios, the required E2E delay (around 1 ms) is shorter than the channel coherence time. According to the studies in [21], when the delay bound approaches to the channel coherence time, the transmit power required to ensure the QoS requirements becomes unbounded. As a result, EE of URLLC is much lower than traditional real-time services with longer delay requirement. On the one hand, most of the power of cellular networks is consumed by BSs. To improve energy efficiency, resource optimization with short blocklength channel codes was studied for downlink (DL) data transmission in orthogonal and non-orthogonal multiple access systems [17, 22]. On the other hand, mobile devices have limited battery capacity in many mission-critical internet-of-thing applications [5]. To prolong the battery life, wireless energy transfer and transmit power optimization were studied for uplink (UL) short packet transmission [23, 24]. However, to ensure the E2E delay and overall reliability of URLLC, UL and DL resources should be jointly allocated [25]. This calls for new solutions for joint UL and DL resource allocation to improve the EE of URLLC.

### B. Our Contributions

In this paper, we aim to find the global optimal resource allocation for URLLC, where the achievable rate in the short blocklength regime is applied. We first study a general resource allocation problem, and show how to find the global optimal solution. Then, we illustrate our method in a concrete EE maximizing problem, where the number of active antennas, bandwidth allocation, and power control are jointly optimized. The decoding error probability, queueing delay violation probability and proactive packet dropping probability are considered in the overall reliability. The UL and DL transmission delays, queueing delay and backhaul delay are taken into account in the E2E delay. The major contributions of this paper are summarized as follows:

- We show how to find the global optimal solutions of resource allocation problems for URLLC, where the QoS constraint is non-convex due to using the maximal achievable rate in the short blocklength regime. We analyze two properties of the non-convex constraint in medium and high SNR regime, i.e., higher than 5 dB, which is relevant in URLLC for ensuring the stringent QoS requirement. With the properties, we show that the global optimal solution lies in a convex subset of the feasible region. By narrowing down the feasible region, the optimization problem becomes convex, whose solution can be found with well-established methods.
- We take energy-efficient resource allocation for URLLC as an example to show how to optimize the number of active antennas, bandwidth allocation, and power control under the non-convex QoS constraint. By applying the properties, we find the global optimal solution. Numerical results demonstrate a remarkable gain of the proposed solution over that with a fixed number of antennas or that with equal bandwidth allocation among different devices. When the average inter-arrival time between packets at each device is much larger than the required delay bound, circuit power is dominated in the total power consumption, and then the EE loss is minor by optimizing antenna configuration and bandwidth allocation without power control.

The rest of this paper is organized as follows. Section II reviews the achievable rate in the short blocklength regime, and derive two properties of the achievable rate which are useful for optimizing resource allocation for URLLC. Section III presents an example system model in URLLC. In Section IV, we formulate an optimization problem that maximizes EE. In Section V, antenna configuration, bandwidth allocation, and power control are optimized. Simulation and numerical results are presented in Section VI. Finally, we conclude our work in Section VII.

## II. Achievable Rate and Properties for Resource Allocation in URLLC

In this section, we first introduce the achievable rate in short blocklength regime obtained in [14], and discuss the basic assumptions when the achievable rate is applied in URLLC. Then, we consider a general resource allocation problem for URLLC. To find a global optimal solution of the problem, we come up with two properties about the transmit power constraint that is derived from the constraint on the achievable rate in the short blocklength regime. With these two properties, the global optimal resource allocation can be found with some well-established methods. Finally, we discuss possible application scenarios of these properties.

### A. Achievable Rate in the Short Blocklength Regime

Shannon's capacity has been widely used to characterize the maximal achievable rate in traditional services (e.g., [26]) where the blocklength can be sufficiently long, which is jointly convex in bandwidth and transmit power. However, data cannot be transmitted without error with limited blocklength. In URLLC, the blocklength of channel coding is short due to low-latency requirement and small packet size (e.g., 20 bytes [2]), and hence the impact of decoding error is more prominent and cannot be ignored. Shannon's capacity cannot characterize the decoding error probability. Besides, the maximal achievable rate in the short blocklength regime is lower than Shannon's capacity. If Shannon's capacity is used in optimizing resource allocation for URLLC, the latency and reliability will be underestimated [27]. Therefore, the achievable rate in the short blocklength regime is inevitable in optimizing resource allocation for URLLC.

For an interference-free single antenna system subject to quasi-static flat fading channel, the maximal achievable rate in

short blocklength regime can be accurately approximated by [14],

$$r \approx \frac{W}{\ln 2}\left[\ln\left(1+\frac{\alpha g P}{\phi N_0 W}\right)-\sqrt{\frac{V}{\tau W}}Q_{\mathrm{G}}^{-1}(\varepsilon^{\mathrm{c}})\right] \text{ (bits/s)}, \quad (1)$$

where $W$ and $P$ are the bandwidth and transmit power, respectively, $\phi > 1$ is the SNR loss due to imperfect CSI at the transmitter [28],[1] $\alpha$ is the large-scale channel gain that depends on path loss and shadowing, $g$ is the small-scale channel gain that results from multi-path effect, $N_0$ is the single-side noise spectral density, $\tau$ is the data transmission duration, $\varepsilon^{\mathrm{c}}$ is the decoding error probability, $Q_{\mathrm{G}}^{-1}(x)$ is the inverse of the Gaussian Q-function, and $V$ is the channel dispersion given by [14],

$$V = 1 - \frac{1}{\left[1+\frac{\alpha g P}{\phi N_0 W}\right]^2} \approx 1, \quad (2)$$

where the approximation in (2) is very accurate when the received SNR is higher than 10 dB [27], which can be easily achieved in cellular networks (especially when supporting URLLC). On the other hand, since $V < 1$ in low SNR regime, by substituting $V = 1$ into (1), we can obtain a lower bound of the achievable rate. If the lower bound is applied in optimizing resource allocation, the reliability and delay requirements can be satisfied.

To apply (1) to URLLC, we need two assumptions:

- Each user experiences quasi-static and flat fading channel, which is reasonable for transmitting a short packet to a user with medium and low velocity. In URLLC, the packet size is small, hence the bandwidth required to transmit a packet does not exceed the coherence bandwidth, which is around $0.5$ MHz [30]. On the other hand, the E2E delay requirement of URLLC is around $1$ ms, which is shorter than the channel coherence time when the velocity of the user is less than $120$ km/h [18]. Thus, this assumption is applicable to some outdoor applications like remote control, autonomous vehicle communications and indoor applications such as smart factory and augmented reality [3–5].
- There is no strong intra/inter-cell interference. Weak interference is treated as noise. To ensure the stringent QoS requirements of URLLC, strong interference should be avoided, which leads to severe degradation of the reliability and latency. To avoid intra-cell interference, orthogonal multiple access technologies can be applied. To avoid strong inter-cell interference, various interference coordination techniques can be employed, say setting a less-than-one frequency reuse factor (e.g., $1/3$).

### *B. Two Properties for Resource Allocation with the Achievable Rate in Short Blocklength Regime*

Let $b_{\mathrm{req}}$ be the number of bits to be transmitted in each block. By substituting (1) into $\tau r \geq b_{\mathrm{req}}$, the required transmit power to transmit the $b_{\mathrm{req}}$ bits within the transmission duration is given by

$$P \geq \frac{\phi N_0 W}{\alpha g}\left\{\exp\left[\frac{b_{\mathrm{req}}\ln 2}{\tau W}+\frac{Q_{\mathrm{G}}^{-1}(\varepsilon^{\mathrm{c}})}{\sqrt{\tau W}}\right]-1\right\} \triangleq y(W), \quad (3)$$

where $V \approx 1$ is applied. $y(W)$ in constraint (3) is non-convex in $W$. This is because the achievable rate in (1) is non-concave with respect to transmit power and bandwidth [17]. Nevertheless, by analyzing (3) we can find the following property of $y(W)$.

**Property 1.** There is a unique solution $W^{\mathrm{th}}$ that minimizes $y(W)$. Moreover, $y(W)$ is strictly convex in $W$ when $0 < W \leq W^{\mathrm{th}}$.

*Proof:* Please refer to Appendix A. □

To illustrate how to apply Property 1, we consider a cost function $f(W,P)$ and optimize the bandwidth and transmit power that minimize $f(W,P)$, i.e.,

$$\min_{W,P} \quad f(W,P) \quad (4)$$

$$\text{s.t.} \quad (3),\ P>0,\ W>0. \quad (4a)$$

Based on Property 1, we can obtain the following property.

**Property 2.** If the cost function increases with bandwidth and transmit power, then the global optimal solution of problem (4) satisfies

$$W \leq W^{\mathrm{th}}. \quad (5)$$

*Proof:* Please refer to Appendix B. □

Hence, the global optimal solution of problem (4) can be obtained from the following problem,

$$\min_{W,P} \quad f(W,P) \quad (6)$$

$$\text{s.t.} \quad (3),\ P>0,\ W>0,\ W \leq W^{\mathrm{th}}.$$

According to Property 1, the feasible region of problem (6) is convex. If $f(W,P)$ is convex with respect to $W$ and $P$, problem (6) is a convex problem.

### *C. Possible Application Scenarios of the Properties*

The two properties can be extended into many multiple antenna systems, although they are obtained in single antenna systems. For multi-input-single-output (MISO) or single-input-multiple-output (SIMO) systems, the small-scale channel gain is $g = \boldsymbol{h}^H \boldsymbol{h}^H$, where $\boldsymbol{h}$ is the channel vector and $(\cdot)^{\mathrm{H}}$ denotes conjugate transpose. The only difference between the MISO (or SIMO) system and single-antenna system lies in the distribution of $g$. Because the properties do not rely on the channel distribution, they are also applicable to MISO and SIMO systems with different pre-coding schemes, such as maximal-ratio-transmission/combining. For multiple-input-and-multiple-output (MIMO) systems, the required transmit power depends on the eigenvalues of the channel matrix, and power control is more complex, hence there is no simple extension of these two properties.

The properties can be extended into multi-user systems with time/frequency division multiple access, orthogonal frequency

[1] With longer pilot, the channel estimation is more accurate. However, the time/frequency resources for data transmission decreases with the length of pilot. How to design channel training in short blocklength regime for URLLC have been studied in [29], and will not be discussed in our work.

division multiple access (OFDMA), or spatial division multiple access with zero-forcing (ZF) pre-coding. We take OFDMA systems as an example, where the bandwidth and transmit power allocated to the $k$th user are denoted as $W_k$ and $P_k$, respectively. Then, the cost function is $f(W_1, ..., W_K, P_1, ..., P_K)$. The proof of Property 1 is the same as in point-to-point communications, and the proof of Property 2 can be found in Appendix B with $K > 1$.

By changing the cost function, we can study different problems in URLLC. For example, to improve the spectrum efficiency, the total bandwidth should be minimized with the given data rate requirement. In this problem, $f(W_1, ..., W_K, P_1, ..., P_K) = \sum_{k=1}^{K} W_k$ is a linear function of $W_k$, and the global optimal solution can be obtained. In the following sections, we use another concrete example to illustrate how to optimize resource allocation.

## III. System Model for Illustrating Resource Allocation Optimization

Improving EE is one of the major goals in 5G [31], and most mobile devices are powered by limited batteries [5]. To address these concerns and show how to use the two properties, we jointly optimize UL and DL resource allocation for improving EE of a URLLC system, where both UL and DL power consumptions are taken into account.

### A. System Model

We consider a local communication scenario, where the communication distance does not exceed the area covered by a few BSs connected with one hop backhaul, as shown in Fig. 1. Each BS has $N_{\mathrm{t}}$ active transmit antennas, and each mobile device has one antenna. There are two kinds of mobile devices in the network: sensors that only upload packets and users that only download packets uploaded by the sensors. A user may desire the packets from multiple sensors. Sensors first upload their packets to the accessed BS. If the packets are received by the BS and requested by users from adjacent cells, they will be forwarded by the BS to adjacent BSs via the backhaul. Then, the packets to each target user wait in a queue at the associated BS, and the BS transmits the packets to the target user. Such a scenario can be found in autonomous vehicle communications, smart factory and some augmented reality applications [3–5], where the propagation delay and latency in fiber backhaul are much shorter than 1 ms [32].

We take a frequency division duplex system as an example, where the total bandwidth $W_{\max}$ is shared by UL and DL transmissions. The results can be easily extended to time division duplex systems.

### B. Traffic Model

Time is discretized into frames [33].[2] Let $T_{\mathrm{f}}$ denote the duration of each frame. In UL transmission, a sensor either has a packet to transmit with probability $\kappa$ (i.e., active sensor) or stays dumb with probability $1-\kappa$ in each frame [34]. In DL transmission, a user requests packets from multiple sensors [35]. Let $\mathcal{A}_k$ denote the set of indices of sensors that need to transmit packets to the $k$th user. We assume that the packet arrival processes at all sensors are independent identically distributed (i.i.d). Then, the aggregation of the packet arrival processes from multiple sensors in set $\mathcal{A}_k$ can be modeled as a Poisson process with average arrival rate $\lambda_k \triangleq |\mathcal{A}_k|\kappa$ packets/frame, where $|\mathcal{A}_k|$ is the cardinality of set $\mathcal{A}_k$ [36].

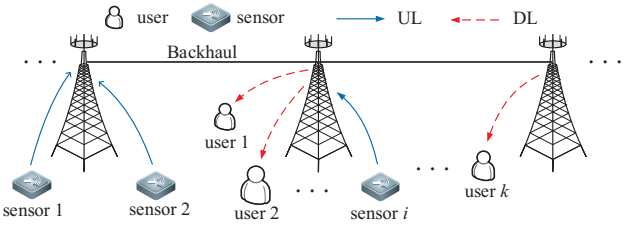



Fig. 1. System model.

### C. Channel Model

Let $\alpha_i^{\mathrm{u}}$ and $\boldsymbol{h}_i^{\mathrm{u}}$ denote the UL large-scale channel gain and channel vector of the $i$th sensor, and let $\alpha_k^{\mathrm{d}}$ and $\boldsymbol{h}_k^{\mathrm{d}}$ denote the DL large-scale channel gain and channel vector of the $k$th user, respectively, $\boldsymbol{h}_i^{\mathrm{u}}, \boldsymbol{h}_k^{\mathrm{d}} \in \mathbb{C}^{N_{\mathrm{t}}\times 1}$. The elements of both UL and DL channel vectors follow independent complex Gaussian distribution with zero mean and unit variance. For SIMO/MISO systems with MRC/MRT,[3] the UL and DL small-scale channel gains of the $i$th sensor and the $k$th user can be expressed as $g_i^{\mathrm{u}} = (\boldsymbol{h}_i^{\mathrm{u}})^{\mathrm{H}}\boldsymbol{h}_i^{\mathrm{u}}$ and $g_k^{\mathrm{d}} = (\boldsymbol{h}_k^{\mathrm{d}})^{\mathrm{H}}\boldsymbol{h}_k^{\mathrm{d}}$, respectively. From (1), the achievable packet service rate (in packets/frame) for the $k$th user and the $i$th sensor can be respectively expressed as follows,

$$r_k^{\mathrm{d}} \approx \frac{\tau W_k^{\mathrm{d}}}{u \ln 2}\left[\ln\left(1+\frac{\alpha_k^{\mathrm{d}} g_k^{\mathrm{d}} P_k^{\mathrm{d}}}{\phi N_0 W_k^{\mathrm{d}}}\right) - \sqrt{\frac{1}{\tau W_k^{\mathrm{d}}}} Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)\right], \quad (7)$$

$$r_i^{\mathrm{u}} \approx \frac{\tau W_i^{\mathrm{u}}}{u \ln 2}\left[\ln\left(1+\frac{\alpha_i^{\mathrm{u}} g_i^{\mathrm{u}} P_i^{\mathrm{u}}}{\phi N_0 W_i^{\mathrm{u}}}\right) - \sqrt{\frac{1}{\tau W_i^{\mathrm{u}}}} Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,u}}\right)\right], \quad (8)$$

where $u$ is the number of bits in each packet, $W_i^{\mathrm{u}}$ and $P_i^{\mathrm{u}}$ are the bandwidth and transmit power of the $i$th sensor, $W_k^{\mathrm{d}}$ and $P_k^{\mathrm{d}}$ are the bandwidth and transmit power of the $k$th user, and $\varepsilon^{\mathrm{c,u}}$ and $\varepsilon^{\mathrm{c,d}}$ are decoding error probabilities in UL and DL, respectively.

### D. Power Model and Energy Efficiency

Considering that the resource allocation at each BS does not depend on those at other BSs, we consider the power consumption and EE in a single-cell scenario. Let $\mathcal{S}$ and $\mathcal{U}$ denote the sets of sensor indices and user indices that access to a BS, respectively.

The average total power consumed by the $i$th sensor is given by

$$\mathbb{E}\left\{P_{\mathrm{tot},i}^{\mathrm{u}}\right\} = \frac{1}{\rho^{\mathrm{u}}}\mathbb{E}\left\{P_i^{\mathrm{u}}\right\} + P^{\mathrm{c,u}}, i \in \mathcal{S}, \quad (9)$$

where $\rho^{\mathrm{u}}$ is the power amplifier (PA) efficiency of each sensor, $P_i^{\mathrm{u}}$ is the transmit power of the $i$th sensor, $P^{\mathrm{c,u}}$ is the circuit

[2] A frame is the minimum scheduling unit in the time domain, which is the mini-slot as proposed in 5G New Radio [7].

[3] The problem with ZF pre-coding is that the distribution of $\boldsymbol{h}^{\mathrm{H}}\boldsymbol{h}$ is unavailable, and we can hardly derive the packet loss probability with it.

power at each sensor, and the average is taken over small-scale channel gains and packet arrival processes (similarly hereinafter).

The average total power consumed by a BS can be modelled as follows [37],

$$\mathbb{E}\left\{P_{\rm tot}^{\rm d}\right\} = \frac{1}{\rho^{\rm d}} \sum_{k\in\mathcal{U}} \mathbb{E}\left\{P_k^{\rm d}\right\} + P^{\rm nt} N_{\rm t} + P^{\rm c,d}, \tag{10}$$

where $\rho^{\rm d}$ is the PA efficiency of the BS, $P_k^{\rm d}$ is the transmit power allocated to the $k$th user, $P^{\rm nt}$ is the circuit power consumed by each antenna at the BS, and $P^{\rm c,d}$ is circuit power component that is independent of the number of active antennas.

The EE is defined as the ratio of the number of bits successfully received by the users to the energy consumed by the sensors and BS. For stationary packet arrival processes and channel fading, the EE is equivalent to the ratio of the average service rate to the average total power consumption, i.e.,

$$\eta \triangleq \frac{\left(u \sum_{k\in\mathcal{U}} \lambda_k / T_{\rm f}\right)\left(1-\varepsilon_{\max}\right)}{P_{\rm tot}}, \tag{11}$$

where the average total power consumption is defined as a weighted sum of UL and DL average total power consumption

$$\begin{aligned} P_{\rm tot} \triangleq & \theta \max_{i\in\mathcal{S}} \mathbb{E}\left\{P_{{\rm tot},i}^{\rm u}\right\} + (1-\theta)\,\mathbb{E}\left\{P_{\rm tot}^{\rm d}\right\} \\ = & \theta \max_{i\in\mathcal{S}} \left\{ \frac{1}{\rho^{\rm u}} \mathbb{E}\left\{P_i^{\rm u}\right\} + P^{\rm c,u} \right\} + \\ & (1-\theta) \left\{ \frac{1}{\rho^{\rm d}} \sum_{k\in\mathcal{U}} \mathbb{E}\left\{P_k^{\rm d}\right\} + P^{\rm nt} N_{\rm t} + P^{\rm c,d} \right\}, \end{aligned} \tag{12}$$

where $\theta \in (0,1)$ is the weight that depends on applications. A large $\theta$ corresponds to the devices that have stringent constraint on power consumption, such as wearable devices.

The numerator of (11) is determined by the packet arrival processes instead of channel capacity or achievable rate of a wireless link. This is because we do not make the full buffer assumption (i.e., there are always enough data in the buffer awaiting for transmission [38]). For a given service, the data arrival rate is determined by the traffic. To guarantee the queueing delay requirement, the average service rate of a system is equal to or higher than the average data arrival rate. In this case, further increasing transmit power does not increase the throughput of the system. Therefore, maximizing EE in (11) is equivalent to minimizing the average total power consumption $P_{\rm tot}$.

## IV. Problem Formulation: Maximizing EE under QoS Constraints

In this section, we formulate a resource allocation problem that maximizes EE under the QoS constraints of URLLC. We first introduce the constraints on E2E delay and overall reliability, as well as how different delay components and packet loss probabilities are ensured by power control and packet dropping policies, and derive the constraints to ensure all the QoS components. Then, we derive the objective function for the resource allocation.

### A. QoS Constraints and Transmission Schemes for Guaranteeing the Constraints

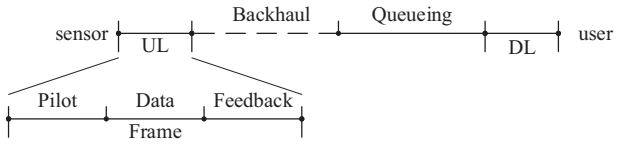



Fig. 2. E2E delay.

*1) E2E Delay and Overall Reliability:* As shown in Fig. 2, the E2E delay $D_{\max}$ consists of UL transmission delay, backhaul latency, queueing delay and DL transmission delay. We assume UL and DL transmissions are finished in two frames (i.e., one frame for each transmission). Each frame consists of three parts: the pilot for channel estimation, data transmission, and CSI feedback. With one-hop fiber backhaul, the backhaul latency is much shorter than $D_{\max}$ [39], which is assumed identical to $T_{\rm f}$ without loss of generality. According to the traffic model, the inter-arrival time between packets at each sensor is equal to or higher than one frame, hence the UL queueing delay is always zero. Since the packets from multiple sensors wait in a separate queue to each user, the DL queueing delay is non-zero. To ensure the E2E delay, the DL queueing delay bound can be obtained as follows,

$$D^{\rm q,d} = D_{\max} - 3T_{\rm f}, \tag{13}$$

where the three frames are occupied by UL/DL transmissions and backhaul delay.

Let $\varepsilon^{\rm q,d}$ be the DL queueing delay violation probability. To satisfy the queueing delay requirement with finite transmit power, proactive packet dropping mechanism can be applied [18] (to be detailed later). Let $\varepsilon^{\rm p,u}$ and $\varepsilon^{\rm p,d}$ denote the UL and DL proactive packet dropping probabilities, respectively. To ensure the overall reliability, the packet loss components should satisfy

$$\begin{aligned} & 1-(1-\varepsilon^{\rm c,u})(1-\varepsilon^{\rm p,u})(1-\varepsilon^{\rm c,d})(1-\varepsilon^{\rm q,d})(1-\varepsilon^{\rm p,d}) \\ \approx & \varepsilon^{\rm c,u}+\varepsilon^{\rm p,u}+\varepsilon^{\rm c,d}+\varepsilon^{\rm q,d}+\varepsilon^{\rm p,d} \le \varepsilon_{\max}. \end{aligned} \tag{14}$$

The approximation is very accurate since each packet loss probability is extremely small.

*2) Ensuring Queueing Delay Requirement and DL Decoding Error Probability:* To derive the constraint for ensuring the queueing delay requirement and DL the decoding error probability, we use the concept of effective bandwidth [40]. According to the analysis in [18], if the delay violation probability is extremely small, which is true in URLLC, the effective bandwidth can be used to analyze the queueing delay at the BS for Poisson arrival processes, interrupted Poisson processes, and switched Poisson processes, where switched Poisson processes are two-state Markovian modulated Poisson processes (MMPP). According to [41], for some other arrival processes, such as discrete-time and continuous-time Markov processes and discrete-time and continuous-time MMPP, effective bandwidth is also applicable.

Since the queueing delay in URLLC is typically shorter than the channel coherence time, the service rate is constant [18].

To ensure the queueing delay requirement $(D^{\mathrm{q,d}}, \varepsilon^{\mathrm{q,d}})$, the constant packet service rate should not be lower than the effective bandwidth [40]. For a Poisson arrival process, effective bandwidth is given by [18],

$$E_k^{\mathrm{B}} = \frac{T_{\mathrm{f}} \ln(1/\varepsilon^{\mathrm{q,d}})}{D^{\mathrm{q,d}} \ln\left[\frac{T_{\mathrm{f}} \ln(1/\varepsilon^{\mathrm{q,d}})}{\lambda_k D^{\mathrm{q,d}}} + 1\right]} \text{ (packets/frame).} \quad (15)$$

The constraint that reflecting the requirements for queueing delay and DL decoding error probability can be expressed as the SNR required to support the average packet arrival rate $\lambda_k$, which can be obtained by substituting (7) into $r_k^{\mathrm{d}} = E_k^{\mathrm{B}}$ as,

$$\gamma_k^{\mathrm{d}} = \exp\left(\frac{E_k^{\mathrm{B}} u \ln 2}{\tau W_k^{\mathrm{d}}} + \frac{Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)}{\sqrt{\tau W_k^{\mathrm{d}}}}\right) - 1. \quad (16)$$

If the received SNR is equal to or higher than $\gamma_k^{\mathrm{d}}$, then $(D_{\max}^{\mathrm{q}}, \varepsilon^{\mathrm{q,d}})$ and $\varepsilon^{\mathrm{c,d}}$ can be satisfied.

*3) Ensuring DL Proactive Packet Dropping Probability:* Fig. 3 (a) shows the fluctuation of the received SNR due to channel fading. When the channel is in deep fading, the received SNR can be lower than the required SNR $\gamma_k^{\mathrm{d}}$ with finite transmit power. In this case, not all the $E_k^{\mathrm{B}}$ packets can be served with the decoding error probability $\varepsilon^{\mathrm{c,d}}$ in the current frame. A simple service policy is to discard all $E_k^{\mathrm{B}}$ packets from the queue when the received SNR is lower than $\gamma_k^{\mathrm{d}}$ to ensure queueing requirement. With such policy, the proactive packet dropping probability equals to the outage probability. However, some packets can still be served under the decoding reliability requirement when the channel is in deep fading. In this paper we adopt the service policy proposed in [18]. Only part of the $E_k^{\mathrm{B}}$ packets are proactively discarded when the channel is in deep fading. In this way, the proactive packet dropping probability can be reduced.[4] By further controlling the proactive packet dropping probability with resource allocation, the overall reliability can be ensured.

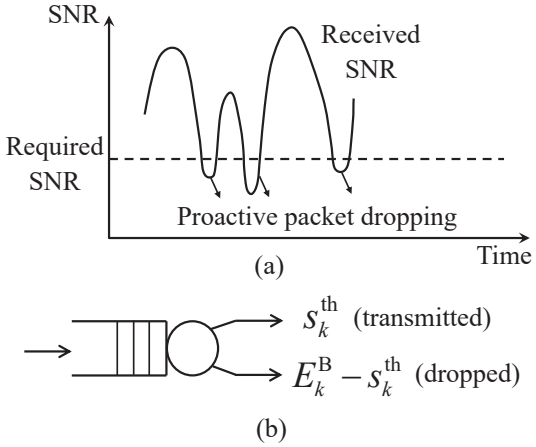



Fig. 3. Intuition on proactive packet dropping mechanism.

Let $P_{\max}^{\mathrm{d}}$ denote the maximal transmit power of the BS. To ensure the power constraint in DL transmission $\sum_{k\in\mathcal{U}} P_k^{\mathrm{d}} \leq P_{\max}^{\mathrm{d}}$, we can introduce a transmit power threshold $P_k^{\mathrm{th,d}}$ to the $k$th user, which satisfies $\sum_{k\in\mathcal{U}} P_k^{\mathrm{th,d}} \leq P_{\max}^{\mathrm{d}}$, and control the power $P_k^{\mathrm{d}}$ not exceeding $P_k^{\mathrm{th,d}}$ [18].

When the channel is in deep fading, i.e., $g_k^{\mathrm{d}} < \frac{\phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} P_k^{\mathrm{th,d}}}$, the transmit power required to ensure the decoding error and queueing delay requirements is higher than $P_k^{\mathrm{th,d}}$. As illustrated in Fig. 3 (b), some packets are transmitted at rate $s_k^{\mathrm{th}}$, which is obtained by substituting $P_k^{\mathrm{th,d}}$ into (7). The rest of the packets are dropped at rate $b_k = \min(E_k^{\mathrm{B}} - s_k^{\mathrm{th}}, 0)$ packets/frame [18]. If there is no packet for the $k$th user in the queue at the BS, no power will be allocated to the user. Otherwise, (16) should be ensured. Then, the power control policy can be expressed as follows when the queue at the BS for the $k$th user is nonempty,

$$P_k^{\mathrm{d}} = \min\left\{P_k^{\mathrm{th,d}}, \frac{\phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} g_k^{\mathrm{d}}}\right\}. \quad (17)$$

As proved in Appendix C, the DL proactive packet dropping probability of the $k$th user can be bounded by

$$\begin{aligned} B_{N_{\mathrm{t}}}^{\mathrm{d}}(g_k^{\mathrm{th,d}}) &\triangleq \int_0^{g_k^{\mathrm{th,d}}} \left(1 - \frac{g}{g_k^{\mathrm{th,d}}}\right) f_{N_{\mathrm{t}}}(g)\,\mathrm{d}g \\ &= \left(1 - \frac{N_{\mathrm{t}}}{g_k^{\mathrm{th,d}}}\right) e^{-g_k^{\mathrm{th,d}}} \sum_{n=0}^{N_{\mathrm{t}}-1} \frac{\left(g_k^{\mathrm{th,d}}\right)^n}{n!} + \\ &\quad e^{-g_k^{\mathrm{th,d}}} \frac{\left(g_k^{\mathrm{th,d}}\right)^{N_{\mathrm{t}}-1}}{(N_{\mathrm{t}}-1)!}. \end{aligned} \quad (18)$$

where $f_{N_{\mathrm{t}}}(\cdot)$ is the distribution of small-scale channel gain given the number of antennas at the BS, and $g_k^{\mathrm{th,d}}$ is defined as

$$g_k^{\mathrm{th,d}} \triangleq \frac{\phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} P_k^{\mathrm{th,d}}}. \quad (19)$$

For Rayleigh fading, $f_{N_{\mathrm{t}}}(g) = \frac{g^{N_{\mathrm{t}}-1}}{(N_{\mathrm{t}}-1)!} e^{-g}$ [42]. To guarantee the DL proactive packet dropping probability, the following constraint should be satisfied,

$$B_{N_{\mathrm{t}}}^{\mathrm{d}}\left(\frac{\phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} P_k^{\mathrm{th,d}}}\right) = \varepsilon^{\mathrm{p,d}}. \quad (20)$$

*4) Ensuring UL Decoding Error and Proactive Packet Dropping Probabilities:* Let $P_{\max}^{\mathrm{u}}$ and $\gamma_i^{\mathrm{u}}$ be the maximal transmit power of each sensor and the SNR threshold required to ensure UL decoding error probability $\varepsilon^{\mathrm{c,u}}$ of the $i$th sensor, respectively. By substituting (8) into $r_i^{\mathrm{u}} = 1$ (i.e., one packet is uploaded in one frame), we can derive the SNR threshold as follows,

$$\gamma_i^{\mathrm{u}} = \exp\left(\frac{u \ln 2}{\tau W_i^{\mathrm{u}}} + \frac{Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,u}}\right)}{\sqrt{\tau W_i^{\mathrm{u}}}}\right) - 1. \quad (21)$$

To save energy, we introduce a threshold of UL transmit power, i.e., $P_i^{\mathrm{th,u}} \leq P_{\max}^{\mathrm{u}}$, and ensure that the reliability can be satisfied with $P_i^{\mathrm{th,u}}$. Then, we can reduce the average UL transmit power by optimizing the value of $P_i^{\mathrm{th,u}}$.

If the received SNR with $P_i^{\mathrm{th,u}}$ is less than the required SNR, the packet is dropped proactively, and no power is transmitted from the sensor. Otherwise, channel inverse is applied to

[4] Another way to reduce packet loss probability is retransmission. However, whether retransmission outperforms proactive packet dropping under the same E2E delay and overall reliability requirements is unclear and deserves further investigation.

achieve the target SNR in (21). Such a power control policy is given by

$$P_i^{\mathrm{u}} = \begin{cases} 0, & \text{if } g_i^{\mathrm{u}} < g_i^{\mathrm{th,u}}, \\ \frac{\phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} g_i^{\mathrm{u}}}, & \text{if } g_i^{\mathrm{u}} \geq g_i^{\mathrm{th,u}}, \end{cases} \tag{22}$$

where $g_i^{\mathrm{th,u}}$ is a threshold of channel gain for the $i$th sensor defined as follows,

$$g_i^{\mathrm{th,u}} \triangleq \frac{\phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} P_i^{\mathrm{th,u}}}. \tag{23}$$

The packet will be discarded when $g_i^{\mathrm{u}} < g_i^{\mathrm{th,u}}$, and the UL proactive packet dropping probability of the $i$th sensor is given by

$$B_{N_{\mathrm{t}}}^{\mathrm{u}}(g_i^{\mathrm{th,u}}) \triangleq \int_0^{g_i^{\mathrm{th,u}}} f_{N_{\mathrm{t}}}(g)\,\mathrm{d}g = 1 - e^{-g_i^{\mathrm{th,u}}} \sum_{n=0}^{N_{\mathrm{t}}-1} \frac{\left(g_i^{\mathrm{th,u}}\right)^n}{n!}. \tag{24}$$

To guarantee the UL proactive packet dropping probability and UL the decoding error, the following constraint should be satisfied,

$$B_{N_{\mathrm{t}}}^{\mathrm{u}}\left(\frac{\phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} P_i^{\mathrm{th,u}}}\right) = \varepsilon^{\mathrm{p,u}}. \tag{25}$$

### *B. Objective Function*

In what follows, we derive the expression of $P_{\mathrm{tot}}$ in (12) for the considered policies to ensure the QoS. Given the power control policies in (17) and (22), we can obtain $\mathbb{E}\{P_k^{\mathrm{d}}\}$ and $\mathbb{E}\{P_i^{\mathrm{u}}\}$, which however are too complicated to obtain graceful results in resource optimization. To simplify the expressions, we consider two tight upper bounds of $\mathbb{E}\{P_k^{\mathrm{d}}\}$ and $\mathbb{E}\{P_i^{\mathrm{u}}\}$, respectively.

From the second terms in (17) and (22), we can see that the power control policies are the same as the channel inverse. When the channel is in deep fading, the transmit power with policies (17) and (22) are less than that with channel inverse due to the maximal transmit power constraints. Thus, we have $P_k^{\mathrm{d}} \leq \frac{\phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} g_k^{\mathrm{d}}}$ and $P_i^{\mathrm{u}} \leq \frac{\phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} g_i^{\mathrm{u}}}$.

From $P_k^{\mathrm{d}} \leq \frac{\phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} g_k^{\mathrm{d}}}$, the upper bound of the average DL transmit power of the $k$th user can be obtained as follows,

$$\mathbb{E}\{P_k^{\mathrm{d}}\} = \xi_k \int_0^{\infty} P_k^{\mathrm{d}} f_{N_{\mathrm{t}}}(g)\,\mathrm{d}g \leq \frac{\xi_k \phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}}(N_{\mathrm{t}}-1)}, \tag{26}$$

where $\xi_k = \lambda_k / E_k^{\mathrm{B}}$ is the probability that the queue at the BS for the $k$th user is nonempty.

From $P_i^{\mathrm{u}} \leq \frac{\phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} g_i^{\mathrm{u}}}$, the upper bound of the average UL transmit power of the $i$th sensor can be obtained as,

$$\mathbb{E}\{P_i^{\mathrm{u}}\} = \kappa \int_0^{\infty} P_i^{\mathrm{u}} f_{N_{\mathrm{t}}}(g)\,\mathrm{d}g \leq \frac{\kappa \phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}}(N_{\mathrm{t}}-1)}. \tag{27}$$

The upper bounds in (26) and (27) are very tight since the probabilities that $g_k^{\mathrm{d}} < g_k^{\mathrm{th,d}}$ and $g_i^{\mathrm{u}} < g_i^{\mathrm{th,u}}$ are extremely small in order to meet the ultra-high reliability requirements in (20) and (24). The expressions of $\mathbb{E}\{P_k^{\mathrm{d}}\}$ and $\mathbb{E}\{P_i^{\mathrm{u}}\}$ and their gaps with the corresponding upper bounds in (26) and (27) can be found in Appendix D.

Finally, an upper bound of $P_{\mathrm{tot}}$ can be obtained by substituting (26) and (27) into (12), i.e.,

$$\begin{aligned} P_{\mathrm{tot}} \leq & \theta \max_{i \in \mathcal{S}} \left( \frac{\kappa \phi N_0 W_i^{\mathrm{u}} \gamma_i^{\mathrm{u}}}{\rho^{\mathrm{u}} \alpha_i^{\mathrm{u}}(N_{\mathrm{t}}-1)} + P_i^{\mathrm{c}} \right) + \\ & (1-\theta) \left( \sum_{k \in \mathcal{U}} \frac{\xi_k \phi N_0 W_k^{\mathrm{d}} \gamma_k^{\mathrm{d}}}{\rho^{\mathrm{d}} \alpha_k^{\mathrm{d}}(N_{\mathrm{t}}-1)} + P^{\mathrm{nt}} N_{\mathrm{t}} + P_0^{\mathrm{c}} \right) \triangleq P_{\mathrm{tot}}^{\mathrm{UB}}. \end{aligned} \tag{28}$$

### *C. Problem Formulation*

As shown in (28), $P_{\mathrm{tot}}^{\mathrm{UB}}$ depends on the bandwidth allocated to each sensor or user in UL or DL, as well as the number of active antennas at the BS. Although the transmit power thresholds in UL and DL do not affect $P_{\mathrm{tot}}^{\mathrm{UB}}$, they control the feasible region of the bandwidth and the number of antennas. The optimization problem, which minimizes the upper bound of the average total power consumption under the QoS constraint, can be formulated as follows,

$$\min_{\substack{P_i^{\mathrm{th,u}}, P_k^{\mathrm{th,d}}, \\ W_i^{\mathrm{u}}, W_k^{\mathrm{d}}, N_{\mathrm{t}}}} \quad P_{\mathrm{tot}}^{\mathrm{UB}} \tag{29}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{S}} W_i^{\mathrm{u}} + \sum_{k \in \mathcal{U}} W_k^{\mathrm{d}} \leq W_{\max}, \tag{29a}$$

$$\sum_{k \in \mathcal{U}} P_k^{\mathrm{th,d}} \leq P_{\max}^{\mathrm{d}}, \tag{29b}$$

$$P_i^{\mathrm{th,u}} \leq P_{\max}^{\mathrm{u}}, \tag{29c}$$

$$\varepsilon^{\mathrm{c,u}} = \varepsilon^{\mathrm{p,u}} = \varepsilon^{\mathrm{c,d}} = \varepsilon^{\mathrm{q,d}} = \varepsilon^{\mathrm{p,d}} = \varepsilon_{\max}/5, \tag{29d}$$

$$(16), (20), (21), (25), P_i^{\mathrm{th,u}}, P_k^{\mathrm{th,d}}, W_i^{\mathrm{u}}, W_k^{\mathrm{d}} > 0,$$

where (29a) is the maximal bandwidth constraint, (29b) and (29c) are the maximal DL and UL transmit power constraints, respectively, (29d) is a near optimal combination of packet loss/error probabilities that ensures the overall packet loss probability in (14),[5] and (20) and (25) guarantee the DL and UL proactive packet dropping probabilities, respectively.

**Remark 1.** Similar to (3), (16) and (21) are derived from the achievable rate in (1) (more exactly, the achievable packet service rate in (7) and (8)), respectively. In the next section, we will show how to apply the two properties provided in section II.B in solving problem (29).

## V. Energy Efficient Resource Allocation Optimization

Problem (29) is a non-convex mixed-integer optimization problem, where the number of active antennas $N_{\mathrm{t}}$ is an integer. We find the global optimal solution of this problem with a three-step method. First, we optimize the power control policy with given bandwidth and the number active of antennas. With the optimal power control policy, problem (29) degenerates into a new problem, which however is still non-convex. Then,

[5]Finding the optimal combination of these five probabilities is intractable. As discussed in [18], for DL transmission, compared with the optimal combination, setting these components equal only causes minor performance loss in transmit power. It is not hard to see that if any of these components goes to zero, then the required UL or DL transmit power goes to infinite. Setting them equal is a reasonable way to simplify the resource allocation optimization.

by applying Properties 1 and 2, we find the global optimal bandwidth allocation for a given number of active antennas. Finally, with the optimal power control policy and bandwidth allocation, we show how to obtain the optimal number of active antennas.

### A. Optimal Power Control Policy

In this subsection, we consider the case where $N_{\mathrm{t}}$ is large enough such that problem (29) is feasible, and find the optimal solution of $P_i^{\mathrm{th,u}}$ and $P_k^{\mathrm{th,d}}$ for given $W_i^{\mathrm{u}}, W_k^{\mathrm{d}}$ and $N_{\mathrm{t}}$.

Let $g^{\mathrm{th,d}*}$ be the solution of $B_{N_{\mathrm{t}}}^{\mathrm{d}}\left(g_k^{\mathrm{th,d}}\right)=\varepsilon_{\max}/5$. Substituting $g_k^{\mathrm{th,d}}=g^{\mathrm{th,d}*}$ into (19), we can obtain the expression of the DL transmit power threshold,

$$P_k^{\mathrm{th,d}}=\frac{\phi N_0 W_k^{\mathrm{d}}\gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} g^{\mathrm{th,d}*}}, \tag{30}$$

which depends on the DL bandwidth allocation. Then, constraint (29b) can be expressed as

$$\sum_{k\in\mathcal{U}}\frac{\phi N_0 W_k^{\mathrm{d}}\gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}} g^{\mathrm{th,d}*}}\le P_{\max}^{\mathrm{d}}. \tag{31}$$

Similarly, let $g^{\mathrm{th,u}*}$ be the solution of $B_{N_{\mathrm{t}}}^{\mathrm{u}}\left(g_i^{\mathrm{th,u}}\right)=\varepsilon_{\max}/5$. Substituting $g_i^{\mathrm{th,u}}=g^{\mathrm{th,u}*}$ into (23), we can obtain the expression of the UL transmit power threshold,

$$P_i^{\mathrm{th,u}}=\frac{\phi N_0 W_i^{\mathrm{u}}\gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} g^{\mathrm{th,u}*}}, \tag{32}$$

which depends on the UL bandwidth allocation. Then, constraint (29c) is equivalent to

$$\frac{\phi N_0 W_i^{\mathrm{u}}\gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}} g^{\mathrm{th,u}*}}\le P_{\max}^{\mathrm{u}}. \tag{33}$$

For notational simplicity, we denote

$$y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)\triangleq\frac{\phi N_0 W_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}}}\left\{\exp\left[\frac{E_k^{\mathrm{B}}u\ln 2}{\tau W_k^{\mathrm{d}}}+\frac{Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)}{\sqrt{\tau W_k^{\mathrm{d}}}}\right]-1\right\}, \tag{34}$$

$$y_i^{\mathrm{u}}(W_i^{\mathrm{u}})\triangleq\frac{\phi N_0 W_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}}}\left\{\exp\left[\frac{u\ln 2}{\tau W_k^{\mathrm{u}}}+\frac{Q_{\mathrm{G}}^{-1}(\varepsilon^{\mathrm{c,u}})}{\sqrt{\tau W_i^{\mathrm{u}}}}\right]-1\right\}, \tag{35}$$

which are obtained by substituting (16) and (21) into $\frac{\phi N_0 W_k^{\mathrm{d}}\gamma_k^{\mathrm{d}}}{\alpha_k^{\mathrm{d}}}$ and $\frac{\phi N_0 W_i^{\mathrm{u}}\gamma_i^{\mathrm{u}}}{\alpha_i^{\mathrm{u}}}$, respectively. Then, the optimal bandwidth allocation can be found from the following problem,

$$\min_{W_i^{\mathrm{u}},W_k^{\mathrm{d}}}\quad\frac{1}{N_{\mathrm{t}}-1}\left[\frac{\theta}{\rho^{\mathrm{u}}}\max_{i\in\mathcal{S}}\{\kappa y_i^{\mathrm{u}}(W_i^{\mathrm{u}})\}+\frac{1-\theta}{\rho^{\mathrm{d}}}\sum_{k\in\mathcal{U}}\xi_k y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)\right] \tag{36}$$

$$\text{s.t.}\quad\sum_{k\in\mathcal{U}}\frac{y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)}{g^{\mathrm{th,d}*}}\le P_{\max}^{\mathrm{d}}, \tag{36a}$$

$$\frac{y_i^{\mathrm{u}}(W_i^{\mathrm{u}})}{g^{\mathrm{th,u}*}}\le P_{\max}^{\mathrm{u}}, \tag{36b}$$

$$(29a),\ W_i^{\mathrm{u}}>0,\ W_k^{\mathrm{d}}>0,$$

where constraints (36a) and (36b) are obtained by substituting (34) and (35) into (31) and (33), respectively.

It is worthy to note that some constraints in problem (29) are not in problem (36). Specifically, constraints (16) and (21) in problem (29) are used to obtain $y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)$ and $y_i^{\mathrm{u}}\left(W_i^{\mathrm{u}}\right)$. Constraints (20) and (25) are used to obtain channel gain thresholds $g^{\mathrm{th,d}*}$ and $g^{\mathrm{th,u}*}$, respectively. Therefore, the optimal solution of problem (36) satisfies constraints (16), (21), (20) and (25), and hence these constraints are not included in problem (36). Since $P_k^{\mathrm{th,d}}$ and $P_i^{\mathrm{th,u}}$ are determined by $W_k^{\mathrm{d}}$ and $W_i^{\mathrm{u}}$ according to (30) and (32), respectively, they are not included in problem (36).

### B. Bandwidth Allocation Optimization

In this subsection, we still consider $N_{\mathrm{t}}$ as a sufficient large constant and find the global optimal bandwidth allocation of problem (36).

Noticing that $y_i^{\mathrm{u}}\left(W_i^{\mathrm{u}}\right)$ and $y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)$ have the same form as $y\left(W\right)$ defined in (3), Property 1 is also applicable to $y_i^{\mathrm{u}}\left(W_i^{\mathrm{u}}\right)$ and $y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)$. Although problem (36) is slightly different from problem (4) in structure, Property 2 is still applicable. Let $W_i^{\mathrm{th,u}}$ and $W_k^{\mathrm{th,d}}$ denote the bandwidth that minimizes $y_i^{\mathrm{u}}\left(W_i^{\mathrm{u}}\right)$ and $y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)$, respectively. According to Property 1 and Property 2, an optimal solution of problem (36) can be obtained by solving the following convex problem,[6]

$$\min_{W_i^{\mathrm{u}},W_k^{\mathrm{d}}}\ \frac{1}{N_{\mathrm{t}}-1}\left[\frac{\theta}{\rho^{\mathrm{u}}}\max_{i\in\mathcal{S}}\{\kappa y_i^{\mathrm{u}}(W_i^{\mathrm{u}})\}+\frac{1-\theta}{\rho^{\mathrm{d}}}\sum_{k\in\mathcal{U}}\xi_k y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)\right] \tag{37}$$

$$\text{s.t. (29a), (36b), (36a)}, 0<W_i^{\mathrm{u}}\le W_i^{\mathrm{th,u}}, 0<W_k^{\mathrm{d}}\le W_k^{\mathrm{th,d}}.$$

Let $W_i^{\mathrm{u}*}$ and $W_k^{\mathrm{d}*}$ denote the optimal bandwidth allocation.

### C. Optimal Number of Active Antennas

In this subsection we study how to obtain the minimal value of $N_{\mathrm{t}}$ that makes problem (37) feasible. Then, we show how to find the optimal solution for $N_{\mathrm{t}}$.

Previous optimal bandwidth allocation and power control policy are obtained when $N_{\mathrm{t}}$ is sufficiently large such that optimization problem (37) (equivalently, problem (36)) is feasible. To find out whether or not the problem is feasible, we study whether or not the required transmit powers in UL and DL exceed maximal transmit powers by solving the following convex optimization problem,

$$\min_{W_i^{\mathrm{u}},W_k^{\mathrm{d}}}\ z=\max\left\{\sum_{k\in\mathcal{U}}\frac{y_k^{\mathrm{d}}\left(W_k^{\mathrm{d}}\right)}{g^{\mathrm{th,d}*}}-P_{\max}^{\mathrm{d}},\max_{i\in\mathcal{S}}\left\{\frac{y_i^{\mathrm{u}}\left(W_i^{\mathrm{u}}\right)}{g^{\mathrm{th,u}*}}\right\}-P_{\max}^{\mathrm{u}}\right\} \tag{38}$$

$$\text{s.t. (29a)},\ 0<W_k^{\mathrm{d}}\le W_k^{\mathrm{th,d}},\ 0<W_i^{\mathrm{u}}\le W_i^{\mathrm{th,u}}.$$

Let $z^*$ denote the minimum value of $z$ in problem (38). Then, problem (37) is feasible if and only if $z^*\le 0$.

Let $N_{\mathrm{t}}^{\min}$ denote the minimum value of $N_{\mathrm{t}}$ that makes the optimization problem feasible. When $N_{\mathrm{t}}\ge N_{\mathrm{t}}^{\min}$, $P_{\mathrm{tot}}^{\mathrm{UB}}$ in (28)

[6]Pointwise maximum preserves the convexity of functions [43], and thus $\max_{i\in\mathcal{S}}\left\{\kappa y_i^{\mathrm{u}}\left(W_i^{\mathrm{u}}\right)\right\}$ is convex.

can be expressed as follows,

$$P_{\text{tot}}^{\text{UB}} = \frac{C_1}{N_{\text{t}} - 1} + C_2 N_{\text{t}} + \theta P^{\text{c,u}} + (1-\theta)\, P^{\text{c,d}}, \tag{39}$$

where $C_1 = \frac{\theta}{\rho^{\text{u}}} \max_{i \in \mathcal{S}} \left\{ \kappa y_i^{\text{u}} \left( W_i^{\text{u}*} \right) \right\} + \frac{1-\theta}{\rho^{\text{d}}} \sum_{k \in \mathcal{U}} \xi_k y_k^{\text{d}} \left( W_k^{\text{d}*} \right)$ and $C_2 = (1-\theta)\, P^{\text{nt}}$.

$C_2$ is a parameter that does not change with $N_{\text{t}}$. The value of $C_1$ is proportional to the objective function of problem (37). If the optimal bandwidth allocation depends on $N_{\text{t}}$, then $C_1$ changes with $N_{\text{t}}$. Otherwise, $C_1$ is a constant. Whether $W_i^{\text{u}*}$ and $W_k^{\text{d}*}$ changing with $N_{\text{t}}$ depends on whether constraints (36a) and (36b) are active (i.e., the equality holds [43]), because the other constraints of problem (37) do not depend on $N_{\text{t}}$. If $N_{\text{t}}$ is large enough, such that constraints (36a) and (36b) are inactive for the optimal solution, then $C_1$ does not change with $N_{\text{t}}$.

Let $\widetilde{N_{\text{t}}}$ denote the minimal value of $N_{\text{t}}$ that makes constraints (36a) and (36b) inactive, which can also be obtained via binary search. For the case where the optimal number of antennas $N_{\text{t}}^*$ that minimizes $P_{\text{tot}}^{\text{UB}}$ is larger than $\widetilde{N_{\text{t}}}$, constraints (36a) and (36b) are inactive and the optimal bandwidth allocation does not change with $N_{\text{t}}$. Since $C_1$ is a constant, $N_{\text{t}}^*$ can be obtained as,

$$N_{\text{t}}^* = \left\lceil \frac{1}{2} \left( 1 + \sqrt{1 + \frac{4C_1}{C_2}} \right) \right\rceil, \tag{40}$$

where $\lceil x \rceil$ is the minimal integer not less than $x$. For the case where $N_{\text{t}}^* \leq \widetilde{N_{\text{t}}}$, $N_{\text{t}}^*$ can be found via exhaustive search in $[N_{\text{t}}^{\min}, \widetilde{N_{\text{t}}}]$.

## VI. Simulation and Numerical Results

In this section, we first illustrate Property 1 with numerical results. Then, the packet loss probability achieved by the proactive packet dropping mechanism and a reactive policy are compared. After that, we show the impact of medium and high SNR approximation ($V \approx 1$) on the average total power consumption. Finally, the power consumption and EE achieved by the proposed policy are shown with numerical results.

TABLE I
Parameters [44]

| | |
|---|---|
| $\varepsilon^{\text{c,u}} = \varepsilon^{\text{p,u}} = \varepsilon^{\text{c,d}} = \varepsilon^{\text{q,d}} = \varepsilon^{\text{p,d}} = \varepsilon_{\max}/5$ | $2 \times 10^{-8}$ |
| E2E delay requirement $D_{\max}$ | 1 ms |
| Duration of each frame $T_{\text{f}}$ | 0.1 ms |
| Duration of transmission $\tau$ | 0.05 ms |
| SNR loss coefficient $\phi$ | 1.5 (around 2 dB [45]) |
| DL Queueing delay requirement $D_{\max}^{q}$ | 0.7 ms |
| Single-sided noise spectral density $N_0$ | $-173$ dBm/Hz |
| Available bandwidth $W_{\max}$ | 100 MHz [2] |
| Packet size $u$ | 20 bytes (160 bits) [2] |
| Path loss model $10 \lg(\alpha_k)$ | $35.3 + 37.6 \lg(d_k)$ |
| Maximum transmit power of sensor $P_{\max}^{\text{u}}$ | 23 dBm (200 mW) |
| Maximum transmit power of BS $P_{\max}^{\text{d}}$ | 40 dBm (10000 mW) |

We consider a single-cell scenario with one BS, which serves $50 \sim 300$ sensors and $10 \sim 100$ users. The distances from the BS to the accessed users and sensors are uniformly distributed from 50 m to 250 m. Each user needs the packets from its nearby sensors with sensor-user distance less than 50 m. Other parameters are listed in Table I, unless otherwise specified.

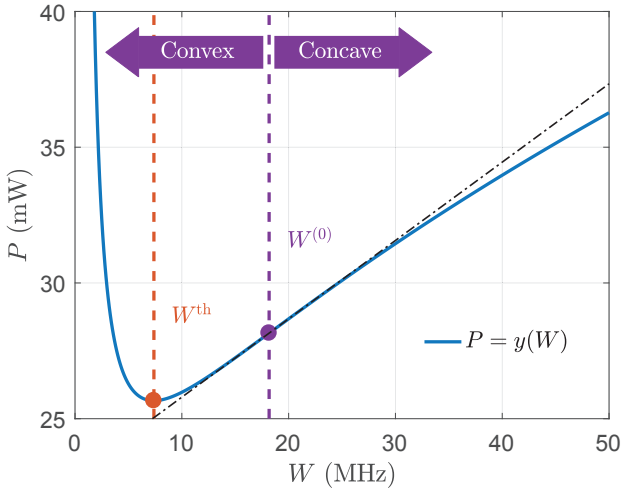



Fig. 4. Demonstration of $y(W)$, where one packet is transmitted to a user locates at cell edge.

The values of $y(W)$ in (3) are illustrated in Fig. 4, where $\alpha$ is set as the path loss of a user at the edge of the cell, $g$ is set as the average small-scale channel gain when $N_{\text{t}} = 4$. The results show that $y(W)$ first decreases and then increases with $W$, and is minimized at $W^{\text{th}}$. We also see that $y(W)$ is concave when $W_k > W_k^{(0)}$, and convex when $W \leq W^{(0)}$. In the region $W \leq W^{\text{th}}$, $y(W)$ is convex in $W$.

Numerical results in Table II show the impact of approximation $V \approx 1$ on the SNR and corresponding average transmit power that are required to ensure the QoS. The results are obtained in a single user scenario, where the distance between the transmitter and receiver is 250 m. The required packet service rate is set to be 1 packet/frame. By changing the bandwidth allocated to the transmitter, the required SNR increases from 5 dB to 20 dB. By using the approximation $V \approx 1$, we can obtain an upper bound of the average transmit power. The results show that the upper bound is very tight when the required SNR is equal to or higher than 5 dB. Even for the scenario that the required SNR is 5 dB, the gap between the upper bound and the accurate average transmit power is around 1%.

Simulation results in Table III validate that the required overall packet loss/error probability can be achieved by the optimal solution. The simulation is carried out in the scenario with 20 users and 200 sensors. Resources are allocated according to the optimal solution of problem (37). The requirement on UL and DL decoding error probabilities are set as $\varepsilon_{\max}/5$. The UL and DL proactive packet dropping probabilities and queueing delay violation probability of sensors and users are obtained through $10^{11}$ Monte Carlo trails. In each trail, Raleigh fading channel gains are randomly generated, packets randomly arrive at the sensors with the average packet arrival rate 100 packets/s, and each user desires the packets from the sensors within the range of 50 m. The highest value of each packet loss component

TABLE II
IMPACT OF THE APPROXIMATION $V \approx 1$

| | | | | |
|---|---|---|---|---|
| Required SNR with approximated $V \approx 1$ (dB) | 5 | 10 | 15 | 20 |
| Required SNR with accurate $V$ in (2) (dB) | 4.913 | 9.986 | 14.998 | 20.000 |
| Average transmit power with approximated $V \approx 1$ (dBm) | 8.110 | 10.427 | 13.557 | 17.179 |
| Average transmit power with accurate $V$ in (2) (dBm) | 8.023 | 10.413 | 13.555 | 17.179 |

TABLE III
EVALUATING THE RELIABILITY WITH THE OPTIMAL SOLUTION

| | | | | |
|---|---|---|---|---|
| Required overall packet loss/error probability $\varepsilon_{\max}$ | $10^{-8}$ | $10^{-7}$ | $10^{-6}$ | $10^{-5}$ |
| Achieved overall packet loss/error probability $\varepsilon'_{\max}$ | $4.42\times10^{-9}$ | $4.43\times10^{-8}$ | $4.46\times10^{-7}$ | $4.45\times10^{-6}$ |

among sensors or users is selected and summed up to obtain the achieved overall packet loss/error probability $\varepsilon'_{\max}$. We can see that with the optimal resource allocation, the overall packet loss probability satisfies the reliability requirement.

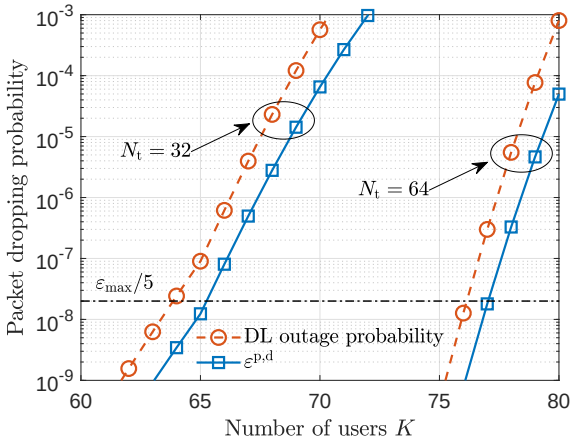



Fig. 5. Proactive packet dropping probabilities v.s. number of users.

To show the relation between the proactive packet dropping probability and outage probability, we compare the service policy proposed in [18] with the policy that discards all packets when the received SNR is lower than $\gamma_k^{\mathrm{d}}$. As mentioned ahead, the proactive packet dropping probability of the later policy equals to the outage probability. In DL transmission, the outage probability can be computed as $\Pr\{g_k^{\mathrm{d}} < g^{\mathrm{th,d}*}\}$. The results in Fig. 5 show that the DL proactive packet dropping probability of the service policy in [18] is much lower than the DL outage probability. For any number of users, the gap is around one order of magnitude. This means that with the service policy in [18], the reliability can be improved significantly. In UL transmission, since there is only one packet to be transmitted in each frame, the packet is dropped when the received SNR is lower than $\gamma_i^{\mathrm{u}}$. Thus, the UL proactive packet dropping probability is the same with the UL outage probability.

The ratio of the circuit power at the BS to the DL transmit power is shown in Fig. 6 (the results in UL transmission are similar and hence are not shown). Due to the development of hardware technologies, the circuit power of each antenna will decrease in the future. To obtain useful insights that do not change over time, we change the value of $P^{\mathrm{nt}}$. The numerical results show that the circuit power is dominated in a wide range of $P^{\mathrm{nt}}$ when the traffic load is not high (say $10 \sim 100$ packets/s). When $P^{\mathrm{nt}} = 2000$ mW, $P^{\mathrm{c,d}} = 120$ mW, $P^{\mathrm{c,u}} = 57$ mW and $\rho^{\mathrm{u}} = \rho^{\mathrm{d}} = 0.5$ (i.e., the values of parameters in 2020 as

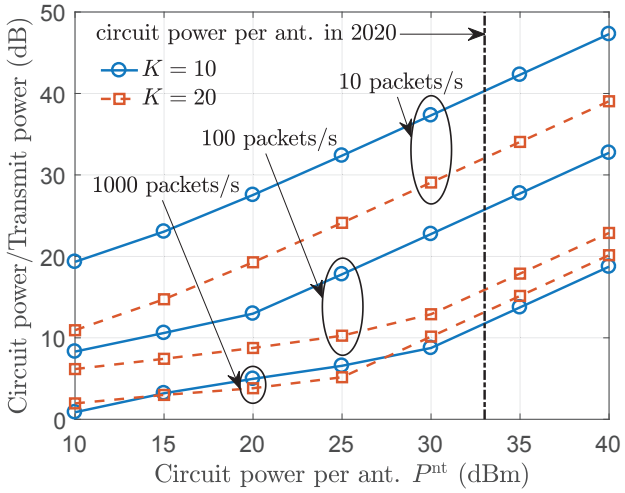
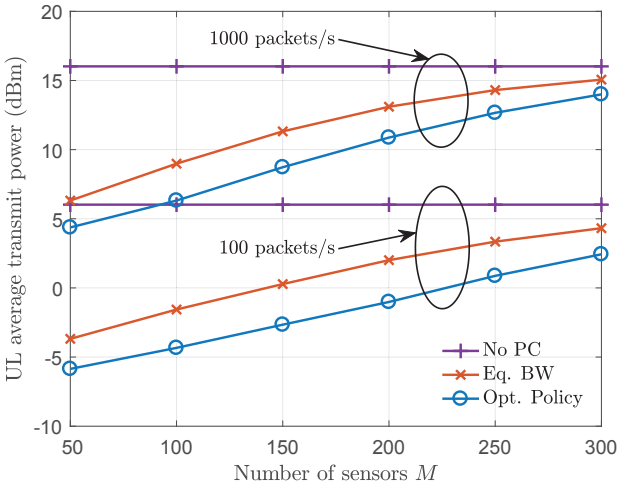


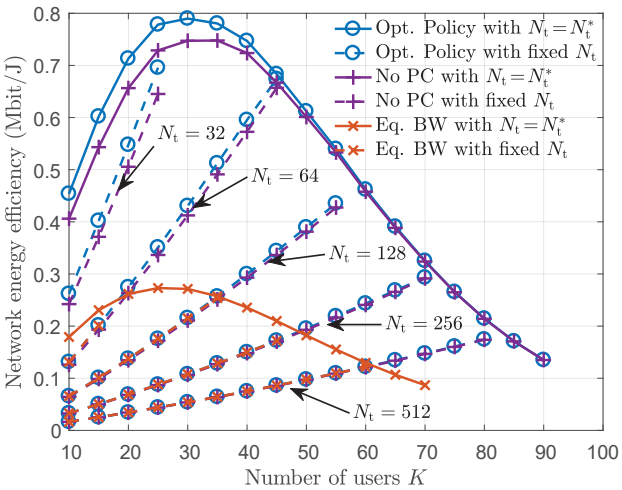

Fig. 6. The ratio of the circuit power to the transmit power in DL when $\theta = 0.99$.

predicted in [37]), the circuit power dominates in all considered traffic loads. Such results seem to be counterintuitive since the required transmit power is very high when the channel is in deep fading. Nonetheless, there are two factors leading to this result. First, to satisfy the stringent requirements with limited transmit power, a large number of antennas is required, which increases with traffic load. Second, the packets arrive randomly, such that the packet arrival time is unpredictable. In some frames with no transmission request, the transmit power is zero, but the BS and sensors cannot be turned into idle mode in order to satisfy the low latency requirements of randomly arrived packets. Thus, circuit powers at the BS and sensors are not zero even there is no packet to transmit.

To show the performance gain with bandwidth allocation and power control/allocation, in Fig. 7 we compare the minimal UL average transmit power and the maximal EE achieved by the optimal resource allocation policy (with legend "Opt. Policy") with two baselines. The first one equally allocates the UL and DL bandwidth among different sensors and users, respectively, while the total UL and DL bandwidth are still optimized (with legend "Eq. BW"). The second one optimizes bandwidth allocation without power control/allocation (with legend "No PC"). In other words, BS always allocates $P_k^{\mathrm{th,d}}$ to the $k$th user if the buffer of the $k$th user is not empty, and a sensor always uses maximal transmit power $P_{\max}^{\mathrm{u}}$ if it has packets to upload. Components of circuit powers are set as the predicted values in 2020. To reflect the impact of traffic load on the performance, different numbers of sensors and users are considered. We only




(a) UL average transmit power vs. number of sensors, where the number of users is 10 and $P^{\mathrm{c,u}} = 57$ mW (17.6 dBm).




(b) Energy efficiency vs. number of users, where the packet arrival rate of each user is 100 packets/s.

Fig. 7. UL average transmit power and network EE, where $\theta = 0.99$.

provide the results when the required number of antennas does not exceed 1024.

The gap between "Opt. Policy" and "Eq. BW" policy in Fig. 7(a) is not large. This means that bandwidth allocation does not have much impact on UL average transmit power. By contrast, by comparing "Opt. Policy" and "No PC" policy, we can see that power control can save a large portion of average transmit power. However, the circuit power is much higher than the average transmit power when the packet arrival rate at each sensor is around 100 packets/s. As a result, in a low traffic load scenario, power control in UL transmission contributes little to reducing the power consumption at each device.

The results in Fig. 7(b) show that when the number of antennas is fixed (i.e., not optimized), much more users can be supported with the optimized bandwidth allocation policy. As a result, optimizing bandwidth allocation is helpful for reducing the required minimum number of antennas to guarantee the QoS. In the case where the optimal number of antennas equals to the minimum number of required antennas, the EE achieved by the optimal bandwidth allocation policy is much higher than the baseline that allocates bandwidth equally. Furthermore, by comparing the solid curves with the dash curves, we can see that adjusting the number of active antennas according to the number of users can increase EE significantly.

## VII. Conclusion

In this paper, we aimed to find the global optimal resource allocation for URLLC, where the achievable rate in the short blocklength regime should be employed, which is non-convex in the radio resources. By analyzing the properties of the required transmit power for ensuring the decoding error probability and the transmission delay of URLLC, we proved that a global optimal solution lies in a convex-subset of the original feasible region. By narrowing down the feasible region, the optimization problem becomes a convex optimization problem. Then, we illustrated how to apply these properties by considering an example problem, where the antenna configuration, the bandwidth allocation, and the power control were optimized to minimize a weighted sum of UL and DL average power consumption under the constraints on ultra-low E2E delay and ultra-low packet loss probability. Simulation and numerical results validated our analysis. When the average inter-arrival time between packets at each device is much larger than the required delay bound, the total power consumption is dominated by circuit power in both UL and DL transmissions. As a consequence, optimizing antenna configuration and bandwidth allocation can improve the EE of the system and the battery life of devices significantly, and adjusting transmit power according to channel fading is not necessary.

## Appendix A
## Proof of Property 1

*Proof:* $y(W)$ can be simplified as $y(W) = aW\left(e^{\frac{b}{W}+\frac{c}{\sqrt{W}}}-1\right)$, where $a$, $b$ and $c$ are positive constants. The first order and second order derivatives of $y(W)$ can be derived as follows,

$$y'(W) = a\left[\left(1-\frac{b}{W}-\frac{c}{2\sqrt{W}}\right)e^{\frac{b}{W}+\frac{c}{\sqrt{W}}}-1\right], \tag{A.1}$$

$$y''(W) = a\left(-cW^{3/2}+c^2W+4bc\sqrt{W}+4b^2\right)\frac{e^{\frac{b}{W}+\frac{c}{\sqrt{W}}}}{4W^3}. \tag{A.2}$$

Let us denote $x(\sqrt{W}) = -cW^{3/2}+c^2W+4bc\sqrt{W}+4b^2$. Then, $y''(W) = 0$ if and only if $x(\sqrt{W}) = 0$. To check when $x(\sqrt{W}) = 0$, we replace $\sqrt{W}$ with $z > 0$. The derivative of $x(z)$ is

$$x'(z) = -3cz^2 + 2c^2z + 4bc. \tag{A.3}$$

It is not hard to verify that there exists a unique $z_0 > 0$ such that $x'(z_0) = 0$, and $x'(z_0) > 0$ in $[0, z_0)$, $x'(z_0) < 0$ in $(z_0, +\infty)$. Considering that $x(0) = 4b^2 > 0$, and $x'(z)$ is non-negative in $[0, z_0]$, we can easily check that $x(z)$ is non-decreasing in $[0, z_0]$, and hence $x(z)$ is positive in $[0, z_0]$. On the other hand, when $z \to +\infty$, $x(z) \to -\infty$. Since $x'(z)$ is negative in $(z_0, +\infty)$, $x(z)$ strictly decreases with $z$. As a result, there is a unique solution in $(z_0, +\infty)$ that satisfies $x(z) = 0$. Therefore,

a unique value of $W^{(0)}$ can be found such that

$$y''(W)\begin{cases}>0 & \text{if } 0<W<W^{(0)},\\ =0 & \text{if } W=W^{(0)},\\ <0 & \text{if } W>W^{(0)}.\end{cases} \tag{A.4}$$

Thus, $y(W)$ is non-convex when $W\in(0,\infty)$.

Now we analyze the properties of $y'(W)$. It can be found that $y'(W)\to 0$ when $W\to+\infty$. Since $y''(W)$ is negative in $(W^{(0)},\infty)$, $y'(W)$ strictly decreases with $W$. Therefore, $y'(W)>0,\forall\, W\in[W^{(0)},\infty)$. Moreover, $y''(W)$ is positive in $(0,W^{(0)})$, which means $y'(W)$ strictly increases with $W$. since $y'(W)\to-\infty$ when $W\to 0^+$, and $y'(W^{(0)})>0$, there is a unique solution in $(0,W^{(0)})$ that satisfies $y'(W)=0$. Denote the solution of $y'(W)=0$ as $W^{\rm th}$. Then, we have,

$$y'(W)\begin{cases}<0 & \text{if } 0<W<W^{\rm th},\\ =0 & \text{if } W=W^{\rm th},\\ >0 & \text{if } W^{\rm th}<W.\end{cases} \tag{A.5}$$

According to (A.5), $y(W)$ decreases with $W$ when $0<W\leq W^{\rm th}$, and increases with $W$ when $W\geq W^{\rm th}$. Therefore, $y(W)$ is minimized at $W^{\rm th}$. Since $W^{\rm th}\in(0,W^{(0)})$, from (A.4) we know that $y(W)$ is strictly convex in $W$ when $0<W\leq W^{\rm th}$. This completes the proof.

□

## Appendix B
## Proof of Property 2

*Proof:* In this appendix, we prove the property in a frequency division multiple access(say OFDMA) system, where the number of users is $K$. If all feasible solutions of problem (4) satisfy condition (5), then Property 2 is obviously true. Otherwise, we denote the feasible solutions that do not satisfy condition (5) as $\widetilde{\boldsymbol{x}}\triangleq(\widetilde{W}_1,\cdots,\widetilde{W}_K,\widetilde{P}_1,\cdots,\widetilde{P}_K)$, where $\widetilde{W}_j>W_j^{\rm th}$ for $j\in\mathcal{J}$.

To prove the property, for any $\widetilde{\boldsymbol{x}}$, we construct another feasible solution of problem (4) $\widetilde{\boldsymbol{x}}^a\triangleq(\widetilde{W}_1^a,\cdots,\widetilde{W}_K^a,\widetilde{P}_1^a,\cdots,\widetilde{P}_K^a)$ that satisfies condition (5) and $f(\widetilde{\boldsymbol{x}}^a)\leq f(\widetilde{\boldsymbol{x}})$.

In the following we show that $\widetilde{\boldsymbol{x}}^a$ can be obtained by replacing $\widetilde{W}_j, j\in\mathcal{J}$ in $\widetilde{\boldsymbol{x}}$ with $W_j^{\rm th}$.

Since $\widetilde{\boldsymbol{x}}$ is a feasible solution, we have $\widetilde{P}_j\geq y_j(\widetilde{W}_j)$. According to Property 1, $y_j(W_j)$ is minimized when $W_j=W_j^{\rm th}$. As a result, $\widetilde{P}_j\geq y_j(\widetilde{W}_j)\geq y_j(W_j^{\rm th})$, and the constrains in (3) are satisfied.

The above analysis indicates that $\widetilde{\boldsymbol{x}}^a$ is also a feasible solution of problem (4). Since the cost function $f(W_1,\cdots,W_K,P_1,\cdots,P_K)$ is non-decreasing, we have $f(\widetilde{\boldsymbol{x}}^a)\leq f(\widetilde{\boldsymbol{x}})$. Hence, we find another feasible solution $\widetilde{\boldsymbol{x}}^a$ that satisfies condition (5) and $f(\widetilde{\boldsymbol{x}}^a)\leq f(\widetilde{\boldsymbol{x}})$.

Therefore, we can always find an optimal solution that satisfies (5). This completes the proof. □

## Appendix C
## The upper bound of DL proactive packet dropping probability in (18)

*Proof:* From the derivation in [18], the DL packet dropping probability of the $k$th user can be approximated by

$$F_{N_{\rm t}}^{\rm d}(g_k^{\rm th,d})\triangleq\int_0^{g_i^{\rm th,d}}\left[1-C(g)\right]f_{N_{\rm t}}(g)\,{\rm d}g, \tag{C.1}$$

where $C(g)=\frac{\ln\left(1+\frac{g\gamma_k^{\rm d}}{g_k^{\rm th,d}}\right)}{\ln\left(1+\gamma_k^{\rm d}\right)}$. Since $C(g)$ is strictly convex with respect to $g$ in $(0,g_k^{\rm th,d})$, we have $C\left((1-\theta)0+\theta g_k^{\rm th,d}\right)<(1-\theta)C(0)+\theta C(g_k^{\rm th,d})$. Thus, $C\left(\theta g_k^{\rm th,d}\right)<1-\theta$, $(0<\theta<1)$. By selecting $\theta=\frac{g}{g_k^{\rm th,d}}$, we obtain $C(g)<1-\frac{g}{g_k^{\rm th,d}}$ in $(0,g_k^{\rm th,d})$. Then, by comparing (C.1) with (18), we have $F_{N_{\rm t}}^{\rm d}\left(g_k^{\rm th,d}\right)<B_{N_{\rm t}}^{\rm d}\left(g_k^{\rm th,d}\right)$. □

## Appendix D
## The tightness of the upper bounds in (26) and (27)

*Proof:* The expressions of $\mathbb{E}\{P_k^{\rm d}\}$ and $\mathbb{E}\{P_i^{\rm u}\}$ can be derived as

$$\mathbb{E}\{P_k^{\rm d}\}=\left[1-B_{N_{\rm t}-1}^{\rm d}\left(g_k^{\rm th,d}\right)\right]\frac{\xi_k\phi N_0W_k^{\rm d}\gamma_k^{\rm d}}{\alpha_k^{\rm d}(N_{\rm t}-1)}, \tag{D.1}$$

$$\mathbb{E}\{P_i^{\rm u}\}=\left[1-B_{N_{\rm t}-1}^{\rm u}\left(g_k^{\rm th,d}\right)\right]\frac{\kappa\phi N_0W_i^{\rm u}\gamma_i^{\rm u}}{\alpha_i^{\rm u}(N_{\rm t}-1)}. \tag{D.2}$$

We can find that the differences between (D.1) and (D.2) and the upper bounds in (26) and (27) lie in the terms $\left[1-B_{N_{\rm t}-1}^{\rm d}\left(g_k^{\rm th,d}\right)\right]$ and $\left[1-B_{N_{\rm t}-1}^{\rm u}\left(g_k^{\rm th,d}\right)\right]$. To guarantee the reliability requirement, $B_{N_{\rm t}-1}^{\rm d}\left(g_k^{\rm th,d}\right)$, $B_{N_{\rm t}-1}^{\rm u}\left(g_k^{\rm th,d}\right)$ and $\varepsilon_{\max}$ are in the same order of magnitude. Since $\varepsilon_{\max}$ is extremely small in URLLC, the upper bounds in (26) and (27) are very tight. □

## References


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

[2] 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.

[3] O. N. C. Yilmaz, Y.-P. E. Wang, N. A. Johansson, *et al.*, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *IEEE ICC Workshops*, 2015.

[4] A. Aijaz, M. Dohler, A. H. Aghvami, *et al.*, "Realizing the tactile internet: Haptic communications over next generation 5G cellular networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[5] P. Schulz, M. Matthe, H. Klessig, *et al.*, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[6] F. Capozzi, G. Piro, L.A. Grieco, *et al.*, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2013.

[7] 3GPP, *Study on New Radio (NR) Access Technologies*. Technical Specification Group Radio Access Network, Technical Report 38.802, Release 14, Mar. 2017.

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

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

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

[11] D. Öhmann, A. Awada, I. Viering *et al.*, "SINR model with best server association for high availability studies of wireless networks," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 60–63, Feb. 2016.

[12] J. Jia, Y. Deng, J. Chen, A.-H. Aghvami, and A. Nallanathan, "Availability analysis and optimization in CoMP and CA-enabled hetnets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2438–2450, Jun. 2017.

[13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[14] 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.

[15] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2823–2835, Aug. 2015.

[16] Y. Hu, A. Schmeink, and J. Gross, "Blocklength-limited performance of relaying under quasi-static Rayleigh channels," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4548–4558, Jul. 2016.

[17] S. Xu and T.-H. Chang and S.-C. Lin, *et al.*, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[18] 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.

[19] ——, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[20] S. Zhang, Q. Wu, S. Xu, *et al.*, "Fundamental green tradeoffs: Progresses, challenges, and impacts on 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 33–56, First Quarter 2017.

[21] R. A. Berry, "Optimal power-delay tradeoffs in fading channels—small-delay asymptotics," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3939–3952, Jun. 2013.

[22] Y. Xu, C. Shen, T.-H. Chang, *et al.*, "Energy-efficient non-orthogonal transmission under reliability and finite blocklength constraints," in *IEEE Globecom Workshops*, 2017.

[23] O. L. A. López, H. Alves, R. D. Souza, and E. M. G. Fernndez, "Ultrareliable short-packet communications with wireless energy transfer," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 387–391, April 2017.

[24] T. A. Khan, R. W. Heath, and P. Popovski, "Wirelessly powered communication networks with short packets," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5529–5543, Dec 2017.

[25] A. Aijaz, "Towards 5G-enabled tactile internet: Radio resource allocation for haptic communications," in *Proc. IEEE WCNC*, 2016.

[26] Z. Xu, C. Yang, G. Y. Li, *et al.*, "Energy-efficient configuration of spatial and frequency resources in MIMO-OFDMA systems," *IEEE Trans. Commun.*, vol. 28, no. 2, pp. 564–575, Feb. 2013.

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

[28] X. Liu, S. Han, and C. Yang, "Energy-efficient training-assisted transmission strategies for closed-loop MISO systems," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 2846–2860, Jul. 2015.

[29] C. Li, S. Yan, and N. Yang, "On channel reciprocity to activate uplink channel training for downlink wireless transmission in tactile internet applications," in *IEEE ICC Workshops*, 2018. [Online]. Available: https://arxiv.org/pdf/1711.05912v2.pdf

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

[31] G. Wu, C. Yang, S. Li, *et al.*, "Recent advance in energy-efficient networks and its application in 5G systems," *IEEE Wireless Commun. Mag.*, vol. 22, no. 2, pp. 145–151, Apr. 2015.

[32] G. Zhang, T. Q. S. Quek, M. Kountouris, *et al.*, "Fundamentals of heterogeneous backhaul design—analysis and optimization," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 876–889, Feb. 2016.

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

[34] G. R1-120056, "Analysis on traffic model and characteristics for MTC and text proposal." Technical Report, TSG-RAN Meeting WG1#68, Dresden, Germany, 2012.

[35] P. Popovski, *et al.*, "Deliverable d6.3 intermediate system evaluation results." ICT-317669-METIS/D6.3, 2014. [Online]. Available: https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D6.3_v1.pdf

[36] 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.

[37] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *Proc. IEEE VTC Spring*, 2015.

[38] C. She and C. Yang, "Energy efficiency and delay in wireless systems: Is their relation always a tradeoff?" *IEEE Trans. on Wireless Commun.*, vol. 15, no. 11, pp. 7215–7228, Nov. 2016.

[39] G. Zhang, T. Q. S. Quek, A. Huang, *et al.*, "Delay modeling for heterogeneous backhaul technologies," in *Proc. IEEE VTC Fall*, 2015.

[40] 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.

[41] Y. Hu, M. Ozmen, M. C. Gursoy, and A. Schmeink, "Optimal power allocation for QoS-constrained downlink multi-user networks in the finite blocklength regime," *IEEE Trans. Wireless Commun., early access*, 2018.

[42] I. E. Telatar, *Capacity of multi-antenna Gaussian channels*, 1995.

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

[44] A. Osseiran, F. Boccardi and V. Braun, *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag*, vol. 52, no. 5, pp. 26–35, May 2014.

[45] M. Shirvanimoghaddam, M. Sadegh Mohamadi, R. Abbas, *et al.*, "Short block-length codes for ultra-reliable low-latency communications," *IEEE Commun. Mag., submitted*. [Online]. Available: https://arxiv.org/abs/1802.09166