Radio Resource Management for Ultra-reliable and Low-latency Communications

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

Tony Q. S. Quek
Information Systems Technology and Design Pillar,
Singapore University of Technology and Design, Singapore
Email: tonyquek@sutd.edu.sg

***Abstract*—Supporting ultra-reliable and low-latency communications (URLLC) is one of the major goals in fifth generation (5G) communication systems. Previous studies focus on ensuring end-to-end delay requirement by reducing transmission delay and coding delay, and only consider reliability in data transmission procedure. However, the reliability reflected by overall packet loss also includes other components such as queueing delay violation. Moreover, which tools are appropriate to design radio resource allocation under constraints on delay, reliability and availability is not well-understood. As a result, how to optimize resource allocation for URLLC is still unclear. In this article, we first discuss the delay and packet loss components in URLLC and the network availability on supporting the quality-of-service of users. Then, we present tools for resource optimization in URLLC. Lastly, we summarize the major challenges related to resource management for URLLC, and perform a case study.**

I. INTRODUCTION

Ultra-high reliability (say 10^{-7} packet loss probability) and ultra-low latency (say 1 ms end-to-end (E2E) delay) are required by a variety of applications such as autonomous vehicles, factory automation, virtual and augmented reality, remote control and healthcare [1, 2]. As summarized in [3], ultra-reliable and low-latency communications (URLLC) lies in the overlapped area of internet-of-things and tactile internet, which is one of major research directions for the fifth generation (5G) cellular networks [4].

Some technical issues of the network architecture, wireless access and resource allocation for tactile internet have been discussed in [2, 3], where E2E delay consists of transmission delay, coding delay, computing delay and propagation delay, and reliability is captured by transmission error. These studies focus on global communications, where the communication distance ranges from hundreds to thousands kilometers, and the propagation delay dominates the E2E delay.

It is worth noting that guarantee the stringent quality of service (QoS) in terms of both latency (defined as E2E delay) and reliability (defined as overall packet loss probability) for URLLC is not easy even in the local communications scenario, where the users are associated with a few adjacent base stations (BS) and the communication distance is less than a few kilometers. In [2], resource allocation with mixed tactile internet and regular traffic was discussed. However, resource management for URLLC in radio access network is challenging even if the system only supports one class of traffic. When designing radio resource allocation for traditional human-to-human (H2H) communications, the blocklength of channel codes is sufficiently large such that Shannon's capacity is an accurate approximation of the error-free achievable rate. However, this is not true for URLLC, where small packets are transmitted. Since only a small amount of bits is transmitted in one coding block and the transmission delay should be very low, the transmission is not error-free with finite blocklength channel codes. Therefore, when designing resource allocation for URLLC to control the packet loss caused by transmission error, Shannon formula can no longer be applied, which cannot characterize the maximal achievable rate with given error probability [5].

Moreover, packet loss may result from factors other than transmission error, such as queueing delay violation. Since some event-driven packets generated by different mobile users (MU) arrive at a BS randomly and the inter-arrival time between packets may be shorter than the transmission duration of each packet, there is a need to consider queueing delay [6]. As a result, the overall packet loss not only comes from uplink (UL) and downlink (DL) transmission errors, but also from queueing delay violation. Because E2E delay and overall reliability are respectively composed of multiple components, the queueing delay should be characterized by a delay bound and a delay bound violation probability for URLLC. Then, tools for analyzing average queueing delay cannot be used. There are two kinds of tools that have been applied in analyzing queueing delay of URLLC in existing literatures. One is network calculus [7], and the other is effective bandwidth and effective capacity [8]. Yet when these tools are applicable (and even whether or not they can be applied) on imposing the constraint on queueing delay for URLLC are not well-understood.

Different from latency and reliability that are the QoS required by each MU, availability is from network perspective, and is another key performance metric for URLLC. Availability is defined as the probability that the network can support a MU with a target QoS requirement on latency and reliability [1]. For the applications such as factory automation and autonomous vehicle, extremely high network availability should be guaranteed. For instance, if the required network availability is 99.999%, then the QoS of one MU in a hundred thousand MUs cannot be satisfied. For another instance, when an autonomous vehicle is moving, in around 10^{-5} fraction of overall service duration, the QoS requirement of the MU cannot be satisfied.

Since the study for URLLC is still in early stage, this article aims to elaborate the design aspects and open problems in radio resource management to achieve the unique performance of URLLC. Because resource management in radio access network cannot deal with propagation delay, we focus on local communication scenarios. The contributions are as follows:

- We elaborate various components of the E2E delay, overall packet loss probability and network availability for URLLC. Because only with appropriate tools, the requirements on latency, reliability and availability can be formulated as constraints on resource optimization, we summarize the state-of-the-arts of analytical tools to characterize the delay and packet loss components for URLLC, and address challenges in resource allocation.
- We discuss design aspects and identify open problems related to radio resource management for URLLC, such as control overhead, network availability guaranteeing, and resource usage efficiency.
- With a case study, we illustrate how to design resource management for UL transmission in URLLC. The results show that retransmission is helpful for reducing required transmit power when the number of antennas is small, otherwise transmitting a packet once but with longer duration is a better solution.

II. REQUIREMENTS OF URLLC IN LOCAL COMMUNICATION SCENARIOS

A typical local communication scenario for URLLC is illustrated in Fig. 1(a), where each MU is served by one of adjacent BSs, which are linked with a single-hop backhaul. When packets are generated at a MU, it first uploads the packets to its own BS. The BS then forwards these packets to the other BSs, where the target MUs are associated with. Finally, the BSs send packets from their buffers to the target MUs.

A. E2E delay

As illustrated in Fig. 1(b), the E2E delay includes UL transmission delay, backhaul delay, queueing delay, and DL transmission delay. With fiber backhaul, backhaul delay is much shorter than 1 ms, and hence will not be discussed in the following. For the case that the MUs are associated to a single BS, the transmission process is simpler and without backhaul delay. The transmission delay could be the durations of multiple frames, depending on the transmission policy, e.g., whether retransmission is allowed or not among subsequent frames. As shown in Fig. 1(c), the control signaling also occupies some time/frequency resources, and leads to extra delay.

In Long Term Evolution systems, the transmission time interval (TTI) is 1 ms, and a frame consists of 10 TTIs. As a result, the transmission delay far exceeds the required E2E delay for URLLC. To reduce transmission delay, a short frame structure was proposed in [9], whose duration equals to one TTI, and each frame includes a phase for control signaling except the phases for UL and DL data transmission. The relationship among the E2E delay, frame duration and blocklength is illustrated in Fig. 1(c). With short blocklength channel codes, coding delay does not exceed transmission duration, and hence is no need to consider.

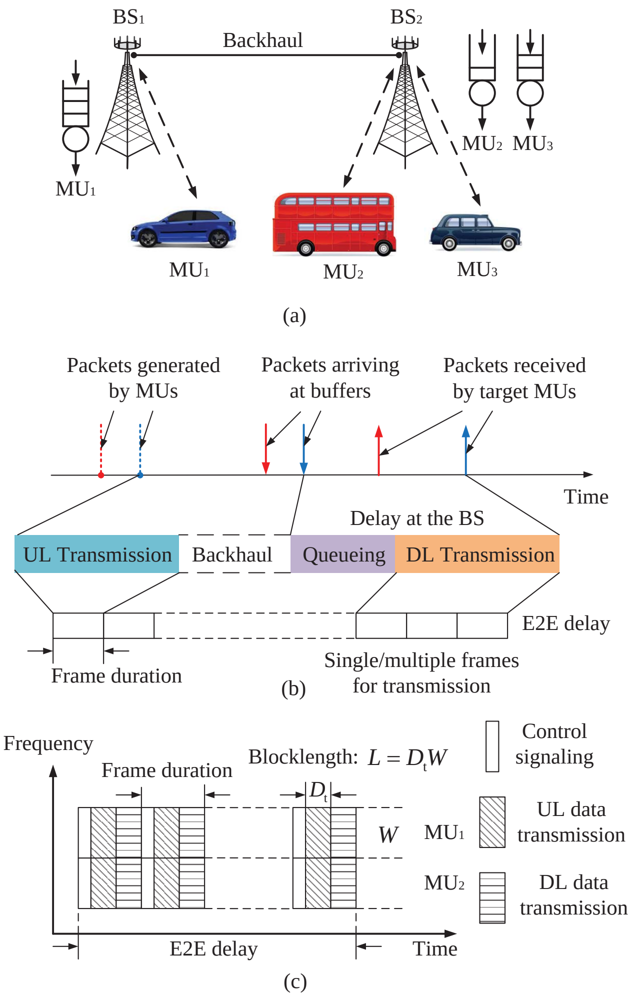


Fig. 1. Example system of URLLC: (a) Local communication scenario. (b) Components of E2E delay. (c) Short frame structure.

B. Overall Packet Loss

The packet loss components are closely related to the delay components in Fig. 1(b).

One of the components is transmission error, which highly depends on the resource allocation. In order to optimize resource allocation with transmission error constraints, we need a mathematical tool to characterize the relationship among achievable rate, transmission error probability, and resources.

Another component comes from queueing delay violation. When serving randomly arrived packets with a wireless link, it is difficult to guarantee a delay bound with probability one if not impossible. To optimize resource allocation with the constraint on queueing delay bound and queueing delay violation probability, we need an analytical tool to translate the queueing delay requirement into the constraint on resource optimization.

In typical application scenarios of URLLC, the required delay is shorter than the channel coherence time. To ensure queueing delay requirement (and also transmission error probabilities), the transmit power may become unbounded in fading channels. To satisfy the queueing delay requirement with finite transmit power, when channel is in deep fading, some packets that cannot be transmitted even with the maximal transmit power can be discarded proactively [10]. Hence, the third component is from the proactive packet dropping.

C. Network Availability

Network availability is closely related to coverage or outage probability [1]. In [11, 12], signal-to-interference-and-noise (SINR) is used to characterize availability, where multi-connectivity is exploited to reduce the probability that the SINR is lower than a required threshold (i.e., outage probability). However, network availability is not always equivalent to outage probability. This is because whether a packet can be successfully transmitted with short delay and high reliability depends not only on SINR but also on the allocated time/frequency resources. For example, given SINR, by increasing the transmission duration of a packet, transmission error probability can be reduced. According to its definition, network availability can be divided into UL availability and DL availability.

III. RESOURCE OPTIMIZATION FOR URLLC

To ensure the QoS and support the availability for URLLC, the total amount of resources (e.g., number of antennas and maximal transmit power at the BS, and system bandwidth) need to be optimized. To use the resources efficiently when guaranteeing a target QoS, the resource allocation (e.g., power, bandwidth, and transmission duration allocation) among UL and DL data transmission and control signalling and further among multiple MUs needs optimization. To formulate the optimization problems with tractable solutions, the QoS constraints should be obtained in closed-form, which is hard as detailed in what follows.

A. Ensuring Transmission Error Requirement

To guarantee the reliability for short packet transmission with the stringent delay, we need the relationship between the achievable rate in finite blocklength regime and transmission error probability $\varepsilon_{\rm t}$, which cannot be characterized by the Shannon's capacity. Unfortunately, the maximal achievable rate with finite blocklength channel codes cannot be obtained in closed-form, as shown in [5] and the references therein. Since it is a building block for deriving the constraints for optimization, appropriate approximations are necessary.

1) Normal Approximation: For single-input-single-output (SISO), single-input-multiple-output (SIMO) and multiple-input-single-output (MISO) systems, the achievable rate with finite blocklength can be accurately approximated as [5]

$$R(\varepsilon_{\rm t}) \approx \frac{W}{\ln 2}\left[\ln\left(1+\frac{\alpha P_{\rm t} g}{N_0 W}\right) - \sqrt{\frac{V}{D_{\rm t} W}} f_{\rm Q}^{-1}\left(\varepsilon_{\rm t}\right)\right] \quad \text{(bits/s)}, \quad (1)$$

where W is the bandwidth, $P_{\rm t}$ is the transmit power, α is the average channel gain that captures path loss and shadowing, g is the normalized instantaneous channel gain, N_0 is the single-side noise spectral density, $D_{\rm t}$ is the transmission duration, $f_{\rm Q}^{-1}(x)$ is the inverse of Gaussian-Q function, and $V = 1 - \frac{1}{\left(1+\frac{\alpha P_{\rm t} g}{N_0 W}\right)^2}$. The differences among SISO, SIMO, and MISO systems lie in the distribution of channel gain g. For multiple-input-multiple-output (MIMO) systems, the achievable rate is similar to that in (1) with the only difference in the instantaneous channel gain, which is replaced by $\mathbf{H}\mathbf{H}^{\dagger}$, where $\mathbf{H}$ is the channel matrix and $(.)^{\dagger}$ is complex conjugate transpose [5].

The first term in (1) is the Shannon's capacity. As shown in Fig. 1(b), if UL transmission of a packet is finished in one frame, the number of symbols for transmitting one packet is $L = D_{\rm t} W$, which is also referred to as blocklength of channel codes in [5]. When the blocklength is large, the achievable rate in (1) approaches the Shannon's capacity.

From (1), the constraint on transmission error probability $\varepsilon_{\rm t}$ for transmitting packets of a coding block can be obtained.

While interference is one of the key factors affecting reliability, existing studies on achievable rate with finite blocklength have not taken interference into consideration. However, even for the scenarios without interference, the constraint on $\varepsilon_{\rm t}$ is neither convex nor concave in $P_{\rm t}$, W and $D_{\rm t}$. As a result, the global optimal power, bandwidth and time allocation with such a constraint is hard to obtain.

2) Simplified Approximations of Achievable Rate: As validated in [7], when the signal-to-noise ratio (SNR) is higher than 10 dB, $V \approx 1$. By introducing such approximation into (1), $R(\varepsilon_{\rm t})$ becomes strictly concave in $P_{\rm t}$, which can be used for deriving the constraint on $\varepsilon_{\rm t}$ and yields optimal power control policy. Since the required SNR should be high to ensure the strict QoS requirement, the high SNR approximation is usually accurate. However, even with the simplified approximation, $R(\varepsilon_{\rm t})$ is still not jointly concave in $P_{\rm t}$ and W. Therefore, global optimal solution is still not easy to derive if we jointly optimize transmit power and bandwidth.

B. Ensuring Queueing Delay Requirement

In this subsection, we address the state-of-the-arts of existing tools that can analyze queueing delay bound $D_{\rm q}$ and its violation probability $\varepsilon_{\rm q}$ for URLLC.

1) Network Calculus: One way to analyze the delay bound and delay violation probability is network calculus [7]. The basic idea of network calculus is converting the accumulatively transmitted data and arrived data from bit domain to SNR domain. In the SNR domain, an upper bound of delay bound violation probability can be obtained [7].

One problem with network calculus is that a data rate requirement is equivalent to a SNR requirement only when the bandwidth of the system is given. If one needs to design both bandwidth and transmit power allocation, a requirement in SNR domain cannot reflect the requirement in bit domain. Moreover, even for power allocation, it is hard to obtain closed-form relation between transmit power and delay bound violation probability for unbounded arrival processes such as Poisson process. As a result, it will be difficult to apply this tool to derive queueing delay constraints for resource allocation optimization.

2) Effective Bandwidth: Different from network calculus, effective bandwidth can be used to design resource allocation in bit domain. Effective bandwidth is the minimal constant service rate that is needed to serve a random arrivals under queueing delay requirement, which is a function of D_{q} and ε_{q} [13]. For URLLC, the delay bound for each packet is usually less than 1 ms, which is shorter than the channel coherence time in typical scenarios. As such, the channel is constant within the delay bound, and the service rate is constant given a resource allocation policy. Therefore, the queueing delay requirement can be satisfied when the constant service rate equals to effective bandwidth. When only one packet is transmitted within a coding block, a constraint on D_{q} and ε_{q} for resource optimization can be imposed by setting the service rate required to transmit a packet equal to (1), as detailed in [10]. When the coherence time is shorter than the delay bound, effective capacity, a dual concept of effective bandwidth, can be used together with effective bandwidth as in [8].

Since effective bandwidth is derived based on the large deviation principle [13], it is widely believed that it can only be used in the scenarios when the delay bound is large. Otherwise, the approximation on the queueing delay violation probability derived from the effective bandwidth is inaccurate. However, simulation results in [14] show that for Poisson process and for arrival processes that are more bursty than Poisson process (e.g., Interrupted Poisson Process), the approximated probability is an exact upper bound of the queueing delay violation probability. Numerical and simulation results in [15] show that when the delay bound is longer than five TTIs, the upper bound is tight. This implies that effective bandwidth can be applied in resource allocation for URLLC with bursty arrival process. Lastly, for Poisson process or Interrupted Poisson Process, the effective bandwidth is with closed-form expression. For these arrival processes, the queueing delay requirement can be represented as a closed-form constraint [10].

C. *Ensuring Network Availability*

To ensure network availability, we need to ensure both UL and DL availability. The transmission error probability in (1) depends on average channel gain α, which is determined by shadowing and the distance from transmitter to receiver, and instantaneous channel gain g. Therefore, both the transmission error probabilities in UL and DL are also random variables. A proper and rigorous framework to impose availability as a constraint on resource allocation is still missing in existing literatures, even not with closed-form expression.

IV. OPEN PROBLEMS IN RESOURCE MANAGEMENT FOR URLLC

In this section, we discuss the major issues and open problems in resource optimization for URLLC.

A. *Overhead for Control Signaling*

As illustrated in Fig. 1(c), control signaling occupies resources in each frame. For H2H communications, overhead for control signaling is trivial compared with data transmission. However, for URLLC this is no longer true due to the small packet size (e.g., 20 bits) [4]. Thus, the resources allocated to control signaling should be designed carefully [7], which ought to be jointly optimized with short packet transmission.

Before the transmission of each event-driven packet that are randomly generated by each MU, a MU first sends a scheduling request to a BS. After receiving the request, a scheduling grant is sent to the MU. Finally, data packet is transmitted to the BS. If any of these three transmissions fails, then the packet is lost. Simulation in [6] shows that the reliability of control signaling can be improved by proper selection of time/frequency resources, transmit power, number of antennas, modulation schemes, and channel codes. Yet how to optimize resource allocation for UL/DL control signaling and data transmission has not been considered in existing literatures.

Another part of control overhead comes from channel estimation in a closed loop transmission strategy. If more training resources are used to estimate channel state information (CSI), transmission errors can be reduced with more accurate CSI, but less resources are remained for data transmission. Similar to H2H communications, the issue of allocating resource for training and data transmission needs to be investigated, but Shannon formula should be replaced by the achievable rate in (1). Alternatively, we can consider an open loop DL transmission strategy, e.g., the BS simply broadcasts the received packets to all MUs without the need to estimate CSI. However, to ensure the QoS of every MU, more DL resources may be required. It is unclear which one requires less resources: open loop or closed loop strategy?

B. *Network Availability Guaranteeing*

For URLLC, the requirement on network availability is much higher than traditional services (around 95 % [1]). Then, shadowing becomes a bottleneck for achieving the extremely high availability. One possible way to deal with shadowing is to exploit macro-diversity [11]. However, the shadowing of closely located links is highly correlated. The outage probability is hard to derive because it depends on the joint probability distribution of the shadowing and fast fading of each link. On the other hand, simulation results in [11] show that the correlation of shadowing has large impact on the outage probability. For example, without correlation the outage probability equals to 10^{-5} with three links, but if the correlation coefficients (defined as the covariance of two shadowing normalized by their standard deviations) exceed 0.3, the probability will be around 10^{-4}. Recall that network availability is not always equal to outage probability, it is unknown whether availability can be ensured with macro-diversity.

Furthermore, multi-user and inter-cell interference leads to low SINR and further deteriorates the network availability. Some possible solutions have been mentioned in [12], such as transmitting the same information from nearby BSs synchronously and using frequency reuse with $1/3$ reuse factor. However, these schemes lead to higher control overhead or lower spectrum efficiency.

C. *Resource Usage Efficiency*

While ensuring stringent QoS requirement with extremely high availability is not an easy task, the spectrum efficiency and energy efficiency should not be compromised in URLLC.

As discussed earlier, the E2E delay of URLLC is shorter than the channel coherence time in typical scenarios. To ensure queueing delay requirement, the transmit power may become unbounded due to deep fading [10], which leads to very low energy efficiency. This is largely overlooked in existing studies on URLLC. Some 5G radio technologies can help alleviate this problem incurred by channel fading. For example, in massive MIMO, millimeter wave and visible light communication systems, the probability that channels are in deep fading is low. Yet using these technologies for URLLC raises new problems. For example, when using millimeter wave and visible light communications, the coverage area of each cell is small. This will lead to frequent handover and will need coordination among BSs. Besides, for massive MIMO and millimeter wave communications, whether or not themselves are energy efficient is problematic.

To ensure the E2E delay and overall packet loss probability with a given amount of resources for URLLC, there is a tradeoff between UL and DL resource allocation, which should be jointly allocated. For example, given the E2E delay requirement, if more time is used for UL transmission, the remaining time for queueing delay and DL transmission decreases. A joint UL and DL resource allocation has been studied for bi-directional haptic communications [8], where queueing delay and queueing delay violation is guaranteed by using effective bandwidth and effective capacity, but transmission errors were not considered. Even by using Shannon's Capacity to formulate the optimization problem, it is still intractable due to the joint resource allocation [8].

With the short frame structure, resource management becomes more flexible. For example, to ensure the reliability with the delay requirement, the transmitter can transmit a packet without retransmission either with longer duration or with larger bandwidth, or retransmit the packet in subsequent frames. How to exploit such flexibility to minimize the required resources to ensure the QoS and availability deserves further investigation.

D. *Other Issues*

Different kinds of services such as enhanced Mobile Broadband and URLLC will co-exist in future mobile networks [4]. How to design resource management for URLLC when co-existing with other services has been addressed in [2], which is challenging.

Device-to-device (D2D) transmission (say vehicle-to-vehicle transmission in vehicular networks) is an option mode to reduce transmission delay [1]. However, the disadvantage of such mode is that the communication distance is limited. Besides, how to control the interference among different links to guarantee the availability in this mode is unclear. One possible way of extending the service distance with ensured availability is to use D2D links together with cellular links for each packet transmission, but how to manage the resources of two types of links remains unknown.

V. UPLINK RESOURCE MANAGEMENT: A CASE STUDY

In this section, we take UL transmission as an example to show how to guarantee the transmission delay, transmission error probability and network availability of URLLC with efficient resource management.

To ensure network availability, we introduce the following approach to formulate network availability as a constraint on resource allocation. With given transmit duration $D_{\rm t}$ and bandwidth W, we can obtain the average error probabilities of transmitting a packet with b bits, $\bar{\varepsilon}_{\rm t}$, first from $D_{\rm t}R(\varepsilon_{\rm t}) = b$ and (1) and then by taking the expectation over small-scale channel gain g conditioned on the average channel gain α. Since α is a random variable depending on the locations of MUs and propagation environment, $\bar{\varepsilon}_{\rm t}$ is also random. We find a threshold $\alpha_{\rm th}$ such that $\Pr\{\alpha < \alpha_{\rm th}\} = 1 - \eta$, where η is the required network availability includes UL and DL availabilities. Then, the availability constraint on UL transmission can be formulated as $\Pr\{\bar{\varepsilon}_{\rm t} > \varepsilon_{\rm req}\} \leq (1-\eta)/2$, where $\varepsilon_{\rm req}$ is the required transmission error probability for UL.

A. *System Setup*

Packet size is set to be 20 bytes [4]. Consider the short frame structure, where the duration for UL transmission in each frame (i.e., TTI) is 0.1 ms. Each MU has one antenna and the BS has $N_{\rm t}$ antennas. The maximal transmit power of each MU is 23 dBm. The average channel gain α is determined by path loss and shadowing according to $10\lg\alpha = -35.3 - 37.6\lg(d) + \text{Shadowing}$, where d is the MU-BS distance. The small-scale channel fading is Rayleigh fading. The required network availability is $\eta = 99.999\%$, and the required transmission error probability for UL is $\varepsilon_{\rm req} = 10^{-7}$. All MUs require URLLC. Frequency division multiple access is adopted to avoid interference among MUs. We fix the bandwidth for each MU, hence the number of users does not affect the following results.

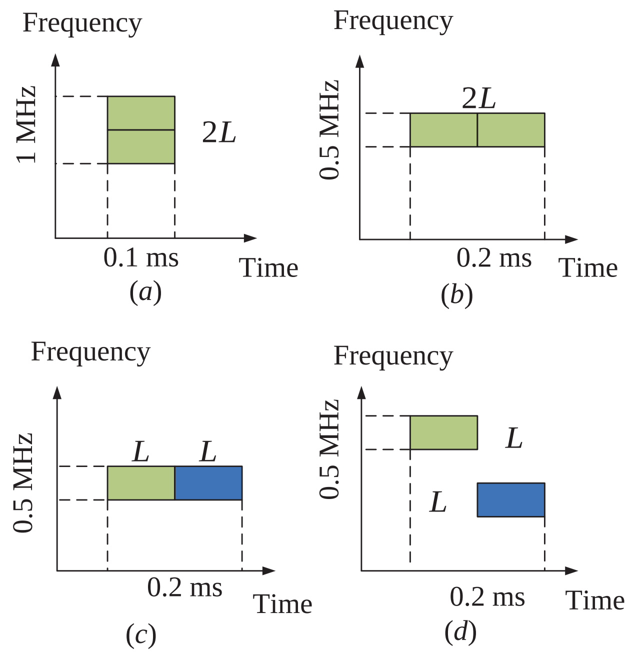


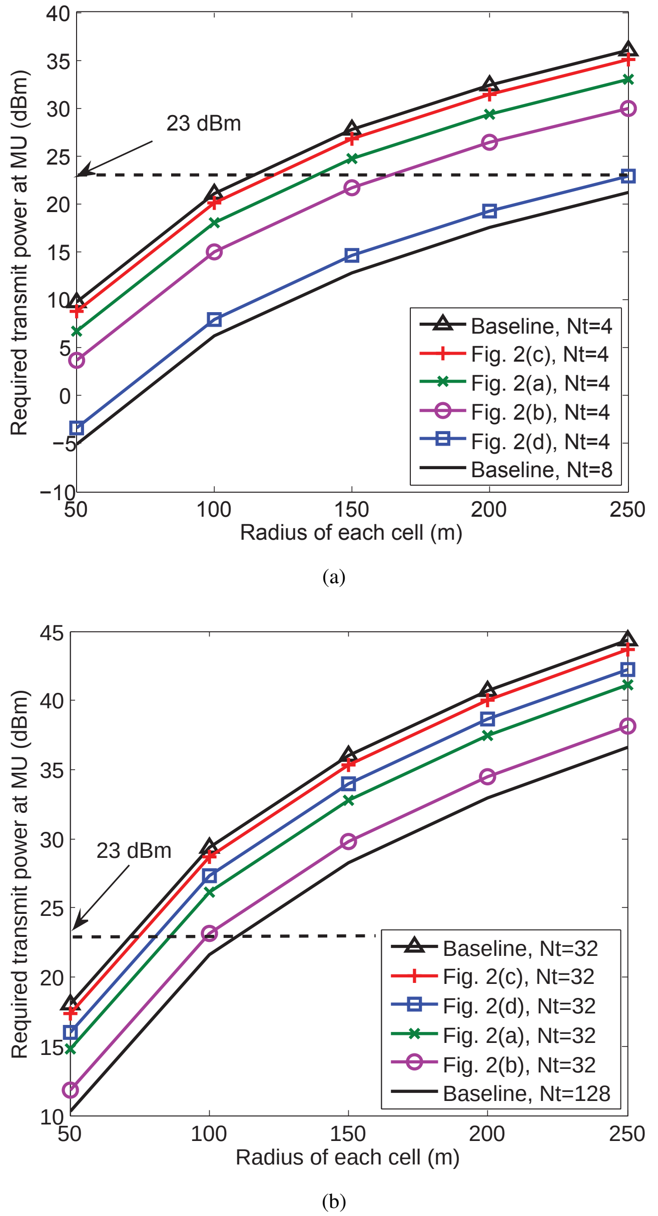
Fig. 2. Four resource management policies. (a) Increasing bandwidth. (b) Increasing transmission duration. (c) Simple retransmission. (d) Retransmission with frequency hopping.

We compare a baseline policy, with which a packet from each MU is transmitted with 0.5 MHz bandwidth within one frame, with four policies shown in Fig. 2. To improve reliability, one simple way is to increase bandwidth (e.g., 1 MHz in Fig. 2(a)). Another way is transmitting a packet with longer duration (e.g., two frames in Fig. 2(b)). Alternatively, we can resort retransmission, where a packet can be retransmitted in subsequent frames as illustrated in Fig. 2(c), or retransmitted with frequency hopping over separated subchannels with different channel gains as shown in Fig. 2(d).

B. Results and Discussion

The required transmit power to ensure the QoS with the target availability is shown in Fig. 3. To ensure the availability, we consider the worst case where all MUs are located at the cell-edge. The number of antennas in the two sub-figures are chosen to be different values such that the required transmit powers are in the same order of the maximal transmit power of a MU.

The results in Fig. 3(a) show that retransmission with frequency hopping is the best policy (this is still true when shadowing is considered, which is not shown for conciseness). Considering that the channel coherence time is longer than 0.2 ms in most cases, there is no diversity gain with the simple retransmission policy. However, the results in Fig. 3(b) show that increasing transmission duration is the best policy to achieve extremely high availability. The reason is that when N_{t} is large the probability that the channel is in deep fading becomes small, and hence the frequency diversity gain is marginal. This suggests that we only need to optimize time/frequency resources allocation for the policies without retransmission if N_{t} is large, say greater than 32. On the other hand, we can see that the required UL transmit power with longer transmission duration (i.e., the policy in Fig. 2(b)) is less than that with larger bandwidth (i.e., the policy in Fig. 2(a)) given the same blocklength. Since with longer UL transmission duration, the queueing delay and DL transmission delay decreases, the required DL transmit power may increase. This implies that the UL and DL transmission resources should be jointly optimized for the policies in Figs. 2(a) and (b).



(a)

(b)

Fig. 3. Required transmit power v.s. radius of each cell, where the MUs are in cell-edge. (a) Without shadowing. (b) Shadowing is lognormal distributed with zero mean and 8 dB standard deviation.

Furthermore, the results in Fig. 3 indicate that by increasing the density of BSs (i.e., reducing the radius of cells) and the number of antennas, the target availability can be supported.

If the cell radius is 100 m, then the network availability requirement of 99.999% can be satisfied with $N_{\mathrm{t}} = 128$ and $P_t = 23$ dBm.

VI. CONCLUSION

In this article, we addressed major technical issues on how to ensure the E2E delay and overall packet loss with high availability by radio resource management for URLLC. We first elaborated the delay and packet loss components in local communications, and the networks availability in supporting the QoS in terms of latency and reliability required by every user. Then, mathematical tools for optimizing resource allocation under constraints on transmission error probability and queueing delay violation probability were presented, their application scenarios were discussed. Next, we identified relevant open problems including reducing signaling overhead, ensuring network availability, and improving resource usage efficiency. Finally, we performed a case study for resource management of URLLC.

REFERENCES

[1] P. Popovski, *et al.*, *Deliverable D6.3 Intermediate system evaluation results*, 2014, https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D6.3_v1.pdf. Accessed on Feb. 2017.

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

[3] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, “The tactile internet: Vision, recent progress, and open challenges,” *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 138–145, May 2016.

[4] 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.

[5] 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.

[6] S. A. Ashraf, F. Lindqvist, R. Baldemair, and B. Lindoff, “Control channel design trade-offs for ultra-reliable and low-latency communication system,” in *IEEE Globecom Workshop*, 2015.

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

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

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

[10] C. She, C. Yang, and T. Q. S. Quek, “Cross-layer transmission design for tactile internet,” in *Proc. IEEE Globecom*, 2016.

[11] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, “Achieving high availability in wireless networks by inter-frequency multi-connectivity,” in *Proc. IEEE ICC*, 2016.

[12] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen, and P. Mogensen, “Signal quality outage analysis for ultra-reliable communications in cellular networks,” in *IEEE Globecom Workshops*, 2015.

[13] 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.

[14] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, “Squeezing the most out of ATM,” *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 203–217, Feb. 1996.

[15] C. She and C. Yang, “Energy efficient design for tactile internet,” in *Proc. IEEE/CIC ICCC*, 2016.