Burstiness Aware Bandwidth Reservation for Ultra-reliable and Low-latency Communications (URLLC) in Tactile Internet

Zhanwei Hou, *Student Member, IEEE*, Changyang She, *Member, IEEE*, Yonghui Li, *Senior Member, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, and Branka Vucetic, *Fellow, IEEE*

***Abstract*—The Tactile Internet that will enable humans to remotely control objects in real time by tactile sense has recently drawn significant attention from both academic and industrial communities. Ensuring ultra-reliable and low-latency communications (URLLC) with limited bandwidth is crucial for Tactile Internet. Recent studies found that the packet arrival processes in Tactile Internet are very bursty. This observation enables us to design spectrally efficient resource management protocol to meet the stringent delay and reliability requirements while minimizing the bandwidth usage. In this paper, both model-based and data-driven methods are applied in classifying the packet arrival process of each user into high or low traffic states, so that we can design efficient bandwidth reservation schemes accordingly. However, when the traffic state classification is inaccurate, it is very challenging to satisfy the ultra-high reliability requirement. To tackle this problem, we formulate an optimization problem to minimize the reserved bandwidth subject to the delay and reliability requirements by taking into account the classification errors. Simulation results show that the proposed methods can save 40% to 70% bandwidth compared with the conventional method that is not aware of burstiness, while guaranteeing the delay and reliability requirements. Our results are further validated by the practical packet arrival processes acquired from experiments using a real tactile hardware device.**

***Index Terms*—Tactile Internet, ultra-reliable and low-latency communications, bandwidth reservation**

I. Introduction

A. Backgrounds and Motivations

With ultra-reliable and low-latency connectivity, Tactile Internet enables delivering real-time control and tactile feedback in both human-to-human and human-to-machine communications [1]. Typical applications include human-machine collaborations in industry automation, tele-surgery in healthcare, virtual reality for remote education or in gaming industry [2]. To provide satisfactory user experience, the latency of tactile feedback should be around 1 ms [3]. For some remote control applications in Tactile Internet, achieving ultra-high reliability is crucial [1]. As a result, one of the key challenges in Tactile Internet is how to achieve ultra-reliable and low-latency communications (URLLC), e.g., 1 ms end-to-end (E2E) delay and 10^{-5} packet loss probability [4].

Z. Hou, C. She, Y. Li, and B. Vucetic are with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia (email: {zhanwei.hou, changyang.she, yonghui.li, branka.vucetic}@sydney.edu.au).

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).

The traditional uplink (UL) multiple access protocols are based on request-and-grant mechanism. The users first need to transmit a scheduling request and then wait for the scheduling grant from the base station. This process introduces significant delay and cannot meet the stringent latency requirement of Tactile Internet. An effective solution to reduce latency is to pre-reserve a certain amount of bandwidth for delay-sensitive users, so that they can transmit their data immediately on the bandwidth reserved for them whenever they have the data to send without going through the time consuming scheduling request and grant process [5]. The bandwidth reservation was originally designed for VoIP services, where the transmission rate is generally fixed and known in advance. However, the packet arrival process of Tactile Internet is completely different and very bursty, i.e., the high traffic state with the peak packet arrival rate occurs aperiodically [6]. This uncertain traffic arrival rate makes the bandwidth reservation very challenging. This is because reserving dedicated bandwidth for all users regardless of their real packet arrival rate will result in low resource usage efficiency.

Motivated by the above issues, we classify a packet arrival process of Tactile Internet into high and low traffic states, and investigate their statistic properties, based on which we then design the corresponding bandwidth reservation and transmission. By sharing the reserved bandwidth among users in low traffic state, instead of reserving dedicate bandwidth for all users regardless of their traffic states, we can achieve efficient bandwidth utilization. The major challenge is that we cannot sacrifice latency and reliability to achieve high bandwidth utilization. Specifically, we will answer the following questions in this paper: 1) Is it possible to ensure ultra-high reliability (i.e., 10^{-5} packet loss probability) when the traffic state classification is inaccurate (e.g., 10^{-2} classification error probability)? 2) If yes, how to design transmission schemes and optimize bandwidth reservations when taking into account the burstiness properties of traffic and the classification errors? 3) How much bandwidth can be saved by exploiting burstiness characteristics of traffic while guaranteeing ultra-low latency and ultra-high reliability required?

It is challenging to answer the above questions, because a variety of delay components and packet loss factors should be considered, resulting from queueing [7], short packet transmissions [8], and random access [9]. Specifically, the delay components might include queueing delay, transmission delay, and access delay. The packet loss factors might include

delay violations in queueing, decoding errors, collisions in random access, and classification errors. As such, we need a cross-layer design framework that takes different delay components and packet loss factors into consideration, especially the classification errors and their impacts on the packet loss probability. However, such a framework is still missing in existing literatures.

B. Our Solutions and Contributions

To address the above challenge, we will design and optimize bandwidth reservation for users in high and low traffic states subject to the delay and reliability constraints of all users, by taking into consideration the traffic state classification errors. To the best knowledge of the authors, this is the first work for exploiting burstiness to optimize bandwidth reservation for URLLC in Tactile Internet. The main contributions of this paper are summarized as follows.

- We establish a cross-layer framework for designing burstiness-aware bandwidth reservation for UL transmission of URLLC in Tactile Internet, where multiple delay components and packet loss factors resulting from queueing, short packet transmissions, and random access, are taken into account. The bandwidth reservations are optimized for users classified into high or low traffic state subject to the UL delay and reliability requirements.
- Two types of classification methods are applied in traffic state classification: a model-based Neyman-Pearson (N-P) method [10] and a data-driven unsupervised learning method, i.e., k-means method [11]. Experiment results show that if the traffic model is inaccurate, the data-driven method outperforms the model-based method.
- The traffic state classification errors are taken into account when designing transmission schemes and optimizing bandwidth reservations. Simulation results show that with around 10^{-2} classification error probability, our method can achieve 10^{-5} packet loss probability. Meanwhile, 40% to 70% bandwidth can be saved compared with the baseline method.

It should be noted that part of the work was presented in a conference version [12]. In the journal version, we further apply a data-driven method in traffic state classification, study impacts of the classification errors on the packet loss probability, and provide experiments with a real tactile device to validate our theoretical work.

II. Related Work

URLLC is considered as one of the new application scenarios in 5G [13]. Some possible solutions for ensuring the quality-of-service (QoS) of URLLC have been proposed in existing literatures [7], [9], [14]–[17]. The studies in [14] optimized packet scheduling scheme to maximize energy efficiency under the delay constraint, where the decoding error in short blocklength regime is considered. The throughput of a decode-and-forward relay system with limited blocklength was studied in [7], where effective capacity was applied in characterizing queueing delay. A framework for cross-layer optimization in URLLC was proposed in [15], where the required maximal transmit power to ensure the latency and reliability requirements is minimized. To improve reliability, multi-path diversity was applied in [16]. The authors of [17] proposed a tactile data quantization scheme, and discussed how to reduce quantization errors. A grant-free uplink transmission scheme was proposed in [9] to avoid scheduling delay.

Recent studies in [6] indicated that the traffic of Tactile Internet is very bursty. The methods in [7], [9], [14], [16], [17] did not take burstiness into account in their resource allocation design. Although the effective bandwidth of some bursty arrival processes was used in resource allocation [15], the reserved resources were not adjusted according to traffic states. In a most relevant reference [6], the authors designed radio access protocol in control-plane that was aware of burstiness. How to classify traffic states in computing systems and cognitive radio networks were studied in [18] and [19], respectively. Since [18] and [19] did not consider URLLC, the impacts of classification errors on the reliability were not analyzed. Therefore, it is still not clear how to leverage burstiness in the data-plane for more effective URLLC design of tactile communications.

III. System Model and QoS Requirements

A Tactile Internet system is composed of three parts: the master device, the network domain and the slave device [1]–[3]. The master device is a tactile device, which converts human inputs to tactile inputs via various coding schemes and sends them to the network domain. The network domain couples the human to the remote environment and a local BS is considered in this paper as the network domain. The slave device is a teleoperator, which interacts with various objects in the remote environment. The master device sends control commands (e.g., 3D locations and velocity of the joints) to the slave device and gets feedback (e.g., force or torques) from it.

In this paper, we focus on the UL transmission in the air interface between N single-antenna tactile users and a BS with N_{r} antennas.[1] Our goal is to minimize the required bandwidth of URLLC by designing transmission schemes and optimizing bandwidth reservation.

A. Burstiness of Traffic

As observed in [6] and the references therein, burstiness is one of the major features of the packet arrival processes in Tactile Internet. Intuitively, the burstiness of Tactile Internet can be explained by the behavior of tactile sense. In most applications of Tactile Internet, the intense force feedback only happens when the slave device touches the surface of an object. Otherwise, the slave device just needs to send few short packets to inform the BS that it is associated with the BS. For the control commands from the master device, it only needs to update its realtime locations and velocity to the BS when it is moving. When it is static, the packet rate is very low. As a result, the packet arrival rate of a moving master device is much higher than a static one. Considering that the burstiness of the packet arrival processes of some applications

[1]How to jointly design UL and downlink (DL) delay components have been studied in [20] and will not be discussed in this paper.

in Tactile Internet may not be very high, we will illustrate the required bandwidth of the proposed methods with various burstiness levels in our simulations.

In the considered system, time is discretized into slots. The duration of each slot is denoted as T_{s}. The burstiness of the packet arrival process is characterized by the squared coefficient of variation (SCV), which is defined as $I_{\mathrm{s}} = \sigma^2/\nu$, where σ^2 and ν are the variance and the mean of the number of packets arriving within a slot, respectively [21]. The burstiness of an arrival process increases with SCV. As such, we will use burstiness and SCV interchangeably hereafter.

To capture the burstiness of the arrival process in Tactile Internet, we use the switched Poisson process (SPP), including a high traffic state and a low traffic state, in our model-based method [22]. In each state of SPP, the packet arrival process is a Poisson process with an average packet arrival rate λ_{H} (i.e., high traffic state) or λ_{L} (i.e., low traffic state). The durations of high and low traffic states follow exponential distributions with the average durations μ_{H} and μ_{L}, respectively.

It is worth noting that for Tactile Internet, there is no well-established traffic model. SPP is a widely adopted model that can characterize the burstiness and the auto-correlation of a packet arrival process. When the model of the packet arrival process is unavailable, a data-driven method will be applied, and the burstiness can be estimated by calculating I_{s} from the historical packet arrival processes.

B. QoS requirements

We use a delay bound and a packet loss probability to characterize the UL delay and reliability requirements, i.e. $(D^{\mathrm{u}}, \epsilon^{\mathrm{u}})$.

In the high traffic state, the packet arrival rate of each user can be higher than 1000 packets/s [2], so that the inter-arrival time between packets may be less than the delay bound requirement, i.e., 1 ms. As a result, there is a queue in the buffer of each user. Let D^{q} and D^{t} be the queueing and transmission delays, respectively. Then, the delay components should satisfy

$$D^{\mathrm{q}} + D^{\mathrm{t}} \leq D^{\mathrm{u}}. \tag{1}$$

For users in the low traffic state, the inter-arrival time between packets is much longer than the delay bound requirement. As a result, there is no queue at the user's buffer, and a user may stay silent for a relatively long time between the transmissions of two packets. To improve bandwidth usage efficiency, a resource pool is shared by all the users in the low traffic state. When a user has a packet to transmit, it randomly selects some time/frequency resources in the resource pool. To ensure the ultra-high reliability and avoid feedback overhead, M-Repetition scheme in [23] is applied, with which a packet is transmitted multiple times repetitively without acknowledgement from the BS.[2] To exploit the frequency-diversity gain, and to avoid continuous transmission collisions, frequency-hopping is adopted. With frequency-hopping, different repetitions of one packet select different frequency resources randomly. Let D^{r} be the duration of each repetition, and M be the number of repetitions. Then, the delay requirement can be satisfied with the following constraint,

$$MD^{\mathrm{r}} \leq D^{\mathrm{u}}. \tag{2}$$

The factors that lead to packet loss depend on traffic states and bandwidth reservation schemes, which will be discussed in Section V.

IV. Traffic State Classification

Different users are assigned with orthogonal pilots for user detection and channel estimation at the BS [24]. When there are transmission collisions, the BS knows which users were sending packets according to their orthogonal pilots.[3] Besides, from the number of pilots, the BS knows how many packets were transmitted from users.

The BS classifies the packet arrival process of each user based on the number of packets arrived in the past sample window. As illustrated in Fig. 1, a sample window is defined as the past T_{w} slots. The number of packets arrived in the past sample window is referred to as a sample and is denoted as k. The BS classifies a user in the high or low traffic states from k. A prediction window follows the sample window. According to [6] and our experiment in section VII-B, the durations of high and low traffic states are in the order of hundreds of milliseconds. So we assume that the total duration of the sample window and the prediction window is much smaller than the average duration that a user stays in each traffic state. As such, the traffic state does not change within the adjacent sample and prediction windows.

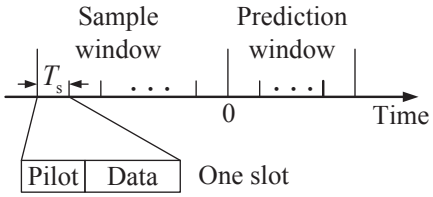

Fig. 1: Illustration of sample and prediction windows.

A. Model-based Method

When the parameters of SPP are available, model-based methods are applicable. It should be noted that there are two types of errors from traffic state classification: classifying a user with high traffic into the low traffic state, and classifying a user with low traffic into the high traffic state. In URLLC, the first kind of error leads to conservative bandwidth reservation. While the second kind of error leads to packet loss or long delay, i.e., the QoS requirements cannot be satisfied due to this kind of error.

To address this issue, we utilize the Neyman-Pearson (N-P) method to classify the traffic state, where the probability of classifying a high traffic state as a low traffic state is bounded by a threshold [10]. Following the N-P method, we denote H_0

[2]In [23], the repetition scheme is referred to as K-Repetition. However, we will use the k-means method in our work. To avoid abuse of notations, the repetition scheme is referred to as M-Repetition.

[3]We assume that some time-frequency resources are reserved for pilots. How to optimize pilot overhead has been studied in [25], and will not be discussed in this work.

and H_1 as the hypotheses that the sample is taken from the high and low traffic state, respectively. Denote D_0 and D_1 are the binary decisions that a user is classified into high and low traffic states, respectively. Then, the probability of false alarm is defined as $P_{\rm f} = P(D_1|H_0)$, i.e., a user with high traffic is classified into low traffic state. Similarly, the probabilities of missing detection and successful detection are defined as $P_{\rm s} = P(D_0|H_1)$ and $P_{\rm d} = P(D_1|H_1)$, respectively.

The idea is to maximize the successful detection rate subject to the constraint on the false alarm probability. To guarantee the reliability requirement of URLLC, $P_{\rm f}$ should be less than a certain requirement $\epsilon^{\rm f}$. In this way, the bandwidth for successfully detected users in low traffic state can be saved and the reliability requirement can be satisfied. The likelihood ratio is defined as $\Gamma(k) = \frac{P(k|H_1)}{P(k|H_0)}$. The relationship between the likelihood ratio and a threshold γ is used to classify traffic states. When $\Gamma(k) \geq \gamma$, the current arrival process is classified into low traffic state. Otherwise, it is classified into high traffic state. The threshold γ can be obtained from the following optimization problem [10],

$$\begin{aligned} (\mathbf{P1}): \quad & \max_{\gamma} P_{\rm d} \\ \text{s.t.} \quad & P_{\rm f} \leq \epsilon^{\rm f}. \end{aligned} \tag{3}$$

According to [10], the threshold of likelihood ratio can be derived as $\gamma = \frac{P(k|H_1)}{P(k|H_0)} = e^{(\lambda_H - \lambda_L)} \left(\frac{\lambda_L}{\lambda_H}\right)^k$, which is equivalent to a threshold in terms of k, i.e., $\hat{\gamma} \triangleq \frac{\ln(\gamma) - (\lambda_H - \lambda_L)}{\ln(\lambda_L) - \ln(\lambda_H)}$. When $k \leq \hat{\gamma}$, it is classified into low traffic. Otherwise, it is classified into high traffic state. As such, we can rewrite the probability of false alarm as

$$P_{\rm f} = P(D_1|H_0) = P\left(k \leq \hat{\gamma}|H_0\right). \tag{4}$$

When H_0 is true, k follows Poisson distribution with mean $T_{\rm w}\lambda_{\rm H}$. From the cumulative distribution function (CDF) of Poisson distribution, it is not hard to find the largest $\hat{\gamma}$ that satisfies $P_{\rm f} \leq \epsilon^{\rm f}$. When $\hat{\gamma}$ is given, the missing detection rate can be calculated from $P_{\rm s} = P(D_0|H_1) = P\left(k > \hat{\gamma}|H_1\right)$.

B. Data-driven Method

When the traffic model is unavailable, data-driven machine learning methods can be used [11]. Considering that the labels of training data are not available at the BS, supervised learning methods cannot be applied. In contrast, the k-means method is an effective unsupervised learning method that can be applied without the labels of training data. In this part, we will show how to apply the k-means method in traffic state classification.

First, the k-means method is trained by historical arrival processes, i.e., $\mathcal{K} = \{k_1, k_2, \cdots, K_{\rm s}\}$, which denotes the number of packets arrived in each of $K_{\rm s}$ sample windows. The target of the k-means method is to partition the input data into a certain number of clusters. According to the real packet arrival processes in [6], we assume that there are two clusters, which means there are two traffic states.

With k-means clustering method, we need to find a centroid of each cluster. We denote the centroids as $\nu_j, j = \{1, 2\}$, which are randomly selected initially. Then, the algorithm proceeds by iteratively running the following two steps until each of the centroid is convergent.

Classifying step: In this step, we classify training data according to their distances to each of the centroid. Let $c_i \in \{1, 2\}$ be the cluster index of k_i. If the c_ith centroid is the nearest one to k_i, then k_i is classified in the c_ith cluster, i.e.,

$$c_i = \arg\min_j ||k_i - \nu_j||^2. \tag{5}$$

Updating centroid step: In this step, we update the location of each cluster centroid according to the following expression,

$$\nu_j = \frac{\sum_{i=1}^{K_{\rm T}} \mathbf{1}\{c_i = j\} k_i}{\sum_{i=1}^{K_{\rm T}} \mathbf{1}\{c_i = j\}}, \tag{6}$$

where $\mathbf{1}\{c_i = j\}$ is an indicator function. If k_i was classified in cluster j, then $\mathbf{1}\{c_i = j\} = 1$. Otherwise, $\mathbf{1}\{c_i = j\} = 0$. So (6) is the mean of the data classified into the jth cluster.

After the training process, we obtain two centroids ν_1^* and ν_2^* that will be used to classify traffic state according to (5).

It is possible to further reduce classification error probability by applying more advanced machine learning algorithms, which deserves further study [11]. However, the performance limit of the classification accuracy is determined by the data. If the data in the two clusters are partly overlapped, classification errors are unavoidable.

V. Transmission Schemes for Different Traffic States

After traffic state classification, we will show how to design transmission schemes for users in high and low traffic states.

A. High Traffic State

To ensure QoS requirements, the dedicated bandwidth is reserved for each user in high traffic state.

1) Queueing Analysis: We use effective bandwidth[4] to characterize the queueing delay requirement. Effective bandwidth is defined as the minimal constant service rate (i.e., packets/slot) that is required to meet the requirements of maximum queueing delay $D_m^{\rm q}$ and maximum queueing delay violation probability $\epsilon_m^{\rm q}$ for the mth user. It is widely believed that effective bandwidth is applicable when the delay bound is large. However, the studies in [15] validated that it is applicable in short delay regime when the arrival process is more bursty than a Poisson process, i.e., the scenario we considered in this paper. According to [15], the effective bandwidth of Poisson process with average packet rate $\lambda_{\rm H}$ is given by

$$E_{\rm B}^m = \frac{T_{\rm s} \ln\left(1/\epsilon_m^{\rm q}\right)}{D_m^{\rm q} \ln\left[\frac{T_{\rm s}\ln\left(1/\epsilon_m^{\rm q}\right)}{\lambda_{\rm H} D_m^{\rm q}} + 1\right]} \text{ packets/slot.} \tag{7}$$

We can see from (7) that larger average traffic arrival rate and more strict delay requirements (i.e., smaller $D_m^{\rm q}$ or $\epsilon_m^{\rm q}$) lead to larger effective bandwidth.

[4]It should be noted that the concept of effective bandwidth is widely adopted in the community of computer networking [26], which is different from the concept of bandwidth of physical spectrum used in the community of telecommunications.

2) Transmission Mode and Bandwidth Reservation: Denote the number of users that are classified in high traffic state as N_{H}. Orthogonal subchannels are assigned to the N_{H} users. Let B_m^{H} be the bandwidth reserved for the mth user. It is assumed that B_m^{H} is less than the coherent bandwidth W_{c}, so that each subchannel is frequency-flat fading.

In Tactile Internet, the packet size is much smaller than that of traditional video services. To transmit short packets with low latency, the blocklength of channel codes is short. As a result, the Shannon capacity is no longer applicable [27]. From [8], an accurate approximation of the achievable data rate of the mth user can be expressed as follows,

$$R_m^{\mathrm{H}} \approx \frac{\rho_{\mathrm{d}} T_{\mathrm{s}} B_m^{\mathrm{H}}}{\ln 2}\left[\ln\left(1+\frac{\alpha_m g_m P_{\mathrm{t}}}{\phi N_0 B_m^{\mathrm{H}}}\right)-\sqrt{\frac{V_m}{D_m^{\mathrm{t}}\rho_{\mathrm{d}} B_m^{\mathrm{H}}}} f_{\mathrm{Q}}^{-1}(\epsilon_m^{\mathrm{t}})\right] \text{ bits/slot}, \quad (8)$$

where ρ_{d} denotes the fraction of time that is used for data transmission in each slot, α_m denotes the large-scale channel gain, $g_m = \boldsymbol{h}_m^{\mathrm{H}}\boldsymbol{h}_m$ is the instantaneous small-scale channel gain, $[\cdot]^H$ is the conjugate transpose and $\boldsymbol{h}_m \in \mathcal{C}^{N_{\mathrm{r}}\times 1}$ is the channel vector whose elements are independent and identically complex Gaussian distributed with zero mean and unit variance. To avoid feedback overhead, the channel state information (CSI) is not available at the users. P_{t} denotes the transmit power of each user, $\phi > 1$ presents the signal-to-noise ratio loss due to inaccurate channel estimation at the BS, N_0 is the noise power spectral density, $V_m^{\mathrm{H}} = 1-[1+(\alpha_m g_m P_{\mathrm{t}})/(\phi N_0 B_m^{\mathrm{H}})]^{-2}$ [8], $f_{\mathrm{Q}}^{-1}(\cdot)$ is the inverse function of the Q-function, and ϵ_m^{t} is the decoding error probability (i.e., block error probability) of the mth user.

Let b be the number of bits in each packet. To guarantee the delay bound and delay bound violation probability, at least E_{B}^m packets should be transmitted within one slot. So the wireless link should transmit bE_{B}^m bits within one slot, i.e., bE_{B}^m bits/slot. By substituting $R_m^{\mathrm{H}} = bE_{\mathrm{B}}^m$ into (8), the conditional decoding error probability when the instantaneous channel gain is g_m can be expressed as follows,

$$\epsilon_m^{\mathrm{t}} = f_{\mathrm{Q}}\left\{\sqrt{\frac{D_m^{\mathrm{t}}\rho_{\mathrm{d}} B_m^{\mathrm{H}}}{V_m^{\mathrm{H}}}}\left[\ln\left(1+\frac{\alpha_m g_m P_{\mathrm{t}}}{\phi N_0 B_m^{\mathrm{H}}}\right)-\frac{bE_{\mathrm{B}}^m \ln 2}{\rho_{\mathrm{d}} T_{\mathrm{s}} B_m^{\mathrm{H}}}\right]\right\}. \quad (9)$$

From (9), the packet loss probability due to decoding error over Rayleigh fading channel can be expressed as [8]

$$\epsilon_m^{\mathrm{c}} = \int_0^{\infty} \epsilon_m^{\mathrm{t}} f_g(x) dx, \quad (10)$$

where $f_g(x) = \frac{1}{(N_{\mathrm{r}}-1)!}x^{N_{\mathrm{r}}-1}e^{-x}$ is the distribution of the instantaneous channel gain [8].

3) Packet Loss Probability: The uplink packet loss comes from the queueing violation ϵ_m^{q} and the decoding error ϵ_m^{c}. To meet the reliability requirement, we have

$$1-(1-\epsilon_m^{\mathrm{q}})(1-\epsilon_m^{\mathrm{c}}) \le \epsilon^{\mathrm{u}}. \quad (11)$$

Since ϵ_m^{q} and ϵ_m^{c} are small (i.e., in the order of $10^{-5}\sim 10^{-8}$), (11) can be accurately approximated by

$$\epsilon_m^{\mathrm{q}} + \epsilon_m^{\mathrm{c}} \le \epsilon^{\mathrm{u}}. \quad (12)$$

B. Low Traffic State

Reserving dedicated bandwidth for users in low traffic state results in low resource usage efficiency. To this end, we reserve a resource pool with K_p orthogonal subchannels that are shared by N_{L} users in the low traffic state. A synchronized multi-channel slotted ALOHA protocol is considered. Without scheduling, each packet randomly selects a subchannel to transmit its message. Transmission collisions happen if two or more packets try to access the same subchannel in the same slot. When the collisions happen, packets cannot be decoded successfully.[5] To reduce the collision probability and improve the reliability, the repetition scheme is adopted [23]. To exploit frequency diversity, frequency-hopping is applied, and different repetitions of each packet occupy different subchannels.

In our analysis, the worst-case of large-scale channel gain is considered to ensure the QoS requirements, i.e., the users are located at the edge of the cell. We will find the minimal K_p and the bandwidth of each subchannel that are required to guarantee the QoS requirements of cell-edge users.

1) Collision Probability: In this paper, a multi-channel slotted ALOHA protocol is considered, where the resource pool is divided into multiple subchannels [9]. During each time slot, K_p subchannels in the resource pool are shared by N_{L} users in low traffic state that select subchannels randomly. Since M copies of each packet are transmitted, the equivalent packet arrival rate of each user is $M\lambda_{\mathrm{L}}$. In the multi-channel slotted ALOHA protocol, the probability that all the M repetitions of a packet collide with the other packets is [9]

$$P_{\mathrm{c}}^0 = 1-\left(\frac{e^{-M\lambda_{\mathrm{L}}}+K_p^M-1}{K_p^M}\right)^{N_{\mathrm{L}}-1}. \quad (13)$$

We can see from (13) that P_{c}^0 increases with λ_{L} and N_{L}, and decreases with K_p.

Due to classification errors, some of the N_{L} users classified into low traffic state are high traffic users. Let n be the number of misclassified high traffic users. For each high traffic user, it may need to transmit multiple packets over multiple subchannels within the same slot. For example, when there are two packets generated by the user in one slot, it randomly selects two subchannels for the packets as if there were two users in low traffic state sending packets in the same slots.

With the above random access mechanism, there is no queueing delay, and each misclassified high traffic user can be approximated by $\lceil \lambda_{\mathrm{H}}/\lambda_{\mathrm{L}} \rceil$ low traffic users, where $\lceil x \rceil$ denotes the smallest integer that is larger than or equal to x. As a result, the equivalent number of low traffic users that share the K_p subchannels is given by $\hat{N}_{\mathrm{L}}(n) = n\lceil \lambda_{\mathrm{H}}/\lambda_{\mathrm{L}} \rceil + N_{\mathrm{L}} - n$. Similar to (13), the conditional collision probability with n misclassified high traffic users is given by

$$P_{\mathrm{c}|n} = 1-\left(\frac{e^{-M\lambda_{\mathrm{L}}}+K_p^M-1}{K_p^M}\right)^{\hat{N}_{\mathrm{L}}(n)-1}. \quad (14)$$

Then, the probability that all the M repetitions of a packet collide with other packets can be obtained from the law of

[5] It is possible to decode the conflicted packets with successive interference cancellation (SIC), which is not considered in this paper.

total probability, i.e.,

$$P_{\rm c} = \sum_{n=0}^{N_{\rm L}} P(n|N_L) P_{{\rm c}|n}, \tag{15}$$

where $P(n|N_{\rm L})$ is the probability that n of the $N_{\rm L}$ users classified into low traffic state are misclassified. $P(n|N_{\rm L})$ is given by

$$P(n|N_{\rm L}) = \binom{N_{\rm L}}{n} P_{\rm h}^n (1-P_{\rm h})^{N_{\rm L}-n}, \tag{16}$$

where $\binom{N_{\rm L}}{n}$ is Binomial coefficient and $P_{\rm h}$ is the probability that a user classified into low traffic state is actually a high traffic user. It is not hard to derive that

$$P_{\rm h} = \frac{\mu_{\rm H} P_{\rm f}}{\mu_{\rm H} P_{\rm f} + \mu_{\rm L}(1-P_{\rm s})}. \tag{17}$$

2) Decoding Error and Packet Loss Probabilities: Now, we analyze the decoding error probabilities of a packet that is transmitted without collision. For each of the M repetitions, the decoding error probability is denoted as $\epsilon_m^{\rm r}$. Similar to (10), $\epsilon_m^{\rm r}$ is given by

$$\epsilon_m^{\rm r} = \int_0^\infty \epsilon_m^{\rm l} f_g(x) dx, \tag{18}$$

where $\epsilon_m^{\rm l}$ is the conditional decoding error probability when the instantaneous channel gain is g_m [8]. Similar to (9), it can be expressed as follows,

$$\epsilon_m^{\rm l} = f_{\rm Q}\left\{ \sqrt{\frac{D_m^{\rm r}\rho_{\rm d} B_{\rm L}}{V_m^{\rm p}}} \left[\ln\left(1 + \frac{\alpha_m^{\rm p} g_m P_{\rm t}}{\phi N_0 B_{\rm L}}\right) - \frac{b \ln 2}{\rho_{\rm d} T_{\rm s} B_{\rm L}} \right] \right\}, \tag{19}$$

where $B_{\rm L}$ is the bandwidth of each subchannel, and $D_m^{\rm r}\rho_{\rm d}$ is the data transmission duration of each repetition.

In the low traffic state, a packet will be successfully received if at least one of the repetitions is transmitted without collision and it is successfully decoded. Since $P_{\rm c}$ is the probability that all the M repetitions of a packet collide with other packets, the overall packet loss probability can be satisfied with the following constraint, given by

$$P_{\rm c} + (1-P_{\rm c})\epsilon_m^{\rm r} \le P_{\rm c} + \epsilon_m^{\rm r} \le \epsilon^{\rm u}, \tag{20}$$

where the upper bound is very tight since $P_{\rm c}$ is extremely small (i.e., ranges from 10^{-5} to 10^{-8}).

C. Error Correction Mechanism for False Alarm Errors

Classifying high traffic users into low traffic state leads to high collision probability. In the cases that the false alarm probability is much higher than the reliability requirement $\epsilon^{\rm u}$, it is very difficult to guarantee reliability. To handle this problem, except for taking into account the classification errors in our transmission scheme design, we propose an error correction mechanism.

As mentioned in Section II, the key difference between high and low traffic states is that only the high traffic users may need to transmit multiple packets in one slot. If the BS received multiple packets from one user in one slot, then the user is misclassified and can be detected by the BS. Then, the BS reserves dedicated bandwidth to the user, and adjusts the number of subchannels in the shared resource pool.

Next, we analyze the collision probability with the proposed error correction mechanism. Before the detection of a misclassified high traffic user, it either stays silent or transmit one packet in a slot. This is because if more than one packet of a user is transmitted in one slot, the user can be detected by the BS. Then, the average packet arrival rate of a high traffic user before being detected can be obtained from the following conditional expectation,

$$\begin{aligned}\lambda_{\rm H}^e &= \Pr\{k_f = 1 | k_f \le 1\} \cdot 1 + \Pr\{k_f = 0 | k_f \le 1\} \cdot 0 \\ &= \frac{\Pr\{k_f \le 1\} - \Pr\{k_f \le 0\}}{\Pr\{k_f \le 1\}},\end{aligned} \tag{21}$$

where k_f can be approximated by a Poisson distribution with the parameter $T_{\rm w}\lambda_{\rm H}$, and thus $\Pr\{k_f \le 1\}$, $\Pr\{k_f \le 0\}$ and $\Pr\{k_f \le 1\}$ can be calculated by the corresponding CDF of the Poisson distribution.

Similar to (15) and (20), we can derive the collision probability of all the M repetitions and the packet loss probability. The decoding error probability without collision is the same as that with no correction scheme. It is not hard to see that $\lambda_{\rm H}^e < \lambda_{\rm H}$, and thus the overall packet loss probability with error correction mechanism is smaller.

VI. Bandwidth Reservation Optimizations

Based on the transmission schemes in the previous section, we formulate an optimization problem to optimize bandwidth reservation that minimizes the required bandwidth with the constraints on delay and reliability.

A. Problem Formulation

The minimal required bandwidth to ensure the QoS requirements of URLLC can be obtained by solving the following optimization problem:

$$\begin{aligned}
\min_{\substack{K_p, B_{\rm L}, B_{\rm H}, \\ M, D_m^{\rm r}, D_m^{\rm q}, D_m^{\rm t}, \\ \epsilon_m^{\rm q}, \epsilon_m^{\rm c}, P_{\rm c}, \epsilon_m^{\rm r}}} \quad & K_p B_{\rm L} + N_{\rm H} B_{\rm H}, && (22)\\
\text{s.t.} \quad & 0 \le B_{\rm L}, B_{\rm H} \le W_{\rm c}, && (22a)\\
& K_p \ge 1, K_p \in \mathbb{Z}, && (22b)\\
& M D_m^{\rm r} \le D^{\rm u}, M \in \mathbb{Z} && (22c)\\
& D_m^{\rm q} + D_m^{\rm t} \le D^{\rm u}, && (22d)\\
& \epsilon_m^{\rm q} + \epsilon_m^{\rm c} \le \epsilon^{\rm u}, && (22e)\\
& P_{\rm c} + \epsilon_m^{\rm r} \le \epsilon^{\rm u}, && (22f)\\
& (7), (9), (10), (13), (18) \text{ and } (19),
\end{aligned}$$

where $N_{\rm H}$ and $N_{\rm L}$ are the numbers of users classified in high and low traffic states, respectively. In the objective function (22), the first term is the bandwidth of the resource pool shared by low traffic users for random access. The second term is the bandwidth reserved for high traffic users.

Since (22) is a mixed-integer optimization problem, it is non-convex and hard to solve directly. To optimize bandwidth reservation, we set some optimization variables as constant system parameters. For the low traffic state, the duration of each repetition is set to be one slot and the number of

repetitions M is fixed. To find the optimal value of M, we can search M in the range of $0 < M < D^{\mathrm{u}}/T_{\mathrm{f}}, M \in \mathbb{Z}$. Therefore, constraint (22c) is simplified as follows,

$$D_m^{\mathrm{r}} = T_{\mathrm{s}}. \tag{23}$$

For the high traffic state, to serve a queueing system that requires a fixed packet rate E_{B}^m, the optimal configuration of transmission delay is one slot [20]. According to this result, the constraint in (22d) can be replaced by

$$D_m^{\mathrm{t}} = T_{\mathrm{s}}, D_m^{\mathrm{q}} = D^{\mathrm{u}} - T_{\mathrm{s}}. \tag{24}$$

Optimization results in [15] indicate that optimizing the value of packet loss probabilities (i.e., ϵ_m^{q}, ϵ_m^{c}, ϵ_m^{r} and P_{c} in problem (22)) is not necessary. Setting $\epsilon_m^{\mathrm{q}} = \epsilon_m^{\mathrm{c}} = \epsilon^{\mathrm{u}}/2$ and $P_{\mathrm{c}} = \epsilon_m^{\mathrm{r}} = \epsilon^{\mathrm{u}}/2$ leads to minor performance loss.

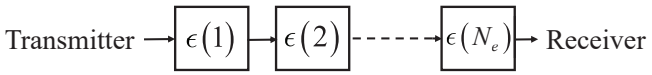

Fig. 2: A general model on packet loss factors.

Now we provide an intuitive explanation for setting $\epsilon_m^{\mathrm{q}} = \epsilon_m^{\mathrm{c}} = \epsilon^{\mathrm{u}}/2$ and $P_{\mathrm{c}} = \epsilon_m^{\mathrm{r}} = \epsilon^{\mathrm{u}}/2$. A general model of packet loss factors is illustrated in Fig. 2. When a packet is sent from a transmitter to a receiver, there are N_e factors leading to packet loss. The packet loss probabilities are denoted as $\epsilon(1), \epsilon(2), ..., \epsilon(N_e)$, respectively. Then, the overall packet loss probability requirement can be satisfied under the constraint

$$1 - \prod_{n_e=1}^{N_e} \left[1 - \epsilon(n_e)\right] \leq \epsilon^{\mathrm{u}}.$$

Similar to (11) and (12), the above constraint can be accurately approximated by

$$\sum_{n_e=1}^{N_e} \epsilon(n_e) \leq \epsilon^{\mathrm{u}}, \tag{25}$$

when $\epsilon(1), \epsilon(2), ..., \epsilon(N_e)$ are small (i.e., in the order of $10^{-5} \sim 10^{-8}$). On the one hand, if any $\epsilon(n_e) \to 0$, $n_e = 1, 2, ..., N_e$, then the required resources approach infinite. For example, if ϵ_m^{q} in (7) approaches to zero, then effective bandwidth (i.e., the required data rate) approaches to infinity. So we cannot set any $\epsilon(n_e)$ to be zero. On the other hand, according to (25), if one of the packet loss probabilities is much higher than the other packet loss probabilities, then the overall packet loss probability is dominated by this maximal packet loss probability, and can be larger than the required overall packet loss probability. To avoid low resource usage efficiency and large overall packet loss probability, the optimal values of $\epsilon(n_e)$, $n_e = 1, 2, ..., N_e$, should be in the same order of magnitude [15]. Setting all the packet loss probabilities in (25) as equal is a good approximation of this constraint.

Therefore, the constraints in (22e) and (22f) are replaced by the following two constraints, given by

$$\epsilon_m^{\mathrm{q}} = \epsilon_m^{\mathrm{c}} = \epsilon^{\mathrm{u}}/2, \tag{26}$$

$$P_c = \epsilon_m^{\mathrm{r}} = \epsilon^{\mathrm{u}}/2, \tag{27}$$

respectively.

With the above settings, the optimization problem (22) can be simplified as follows,

$$\min_{K_p, B_{\mathrm{L}}, B_{\mathrm{H}}} K_p B_{\mathrm{L}} + N_{\mathrm{H}} B_{\mathrm{H}}, \tag{28}$$

$$\text{s.t.} \quad (22a), (22b), (7), (9), (10), (13), (18), (19), (23), (24), (26) \text{ and } (27).$$

B. Optimizations for Bandwidth Reservation

To solve optimization problem (28), a three-step method is proposed.

First, the minimal bandwidth B_{H}^* for a high traffic user can be obtained from (7), (9), and (10). As we have discussed before, to ensure the delay and reliability requirements, we should set $D_m^{\mathrm{t}} = T_{\mathrm{s}}$ and $D_m^{\mathrm{q}} = D^{\mathrm{u}} - T_{\mathrm{s}}$ and $\epsilon_m^{\mathrm{q}} = \epsilon_m^{\mathrm{c}} = \epsilon^{\mathrm{u}}/2$. By substituting the above values into (7), (9), and (10), B_{H}^* is the solution of the following equation:

$$\Phi(B_{\mathrm{H}}) = (\epsilon^{\mathrm{u}} - \epsilon^{\mathrm{f}})/2 - \int_0^{\infty} \epsilon_m^{\mathrm{t}} f_g(x) dx = 0, \tag{29}$$

where

$$\epsilon_m^{\mathrm{t}} = f_{\mathrm{Q}} \left\{ \sqrt{\frac{T_{\mathrm{s}} \rho_{\mathrm{d}} B_{\mathrm{H}}}{V_m^{\mathrm{H}}}} \left[\ln\left(1 + \frac{\alpha_m g_m P_{\max}}{\phi N_0 B_{\mathrm{H}}}\right) - \frac{b E_{\mathrm{B}}^m \ln 2}{\rho_{\mathrm{d}} T_{\mathrm{s}} B_{\mathrm{H}}} \right] \right\}, \tag{30}$$

and E_B^m can be re-expressed as

$$E_{\mathrm{B}}^m = \frac{T_{\mathrm{s}} \ln\left(3/\epsilon^{\mathrm{u}}\right)}{(D^{\mathrm{u}} - T_{\mathrm{s}}) \ln\left[\frac{T_{\mathrm{s}} \ln(3/\epsilon^{\mathrm{u}})}{\lambda_{\mathrm{H}}(D^{\mathrm{u}} - T_{\mathrm{s}})} + 1\right]}. \tag{31}$$

It should be noted that $\Phi(B_{\mathrm{H}})$ is an increasing function of B_{H}. This is because ϵ_m^{t} decreases as the bandwidth B_{H} increases [14]. Therefore, we can use binary search to find B_{H}^*.

Second, similar to searching B_{H}^* in the first step, by substituting $D_m^{\mathrm{r}} = T_{\mathrm{s}}$ and $\epsilon_m^{\mathrm{r}} = \epsilon^{\mathrm{u}}/2$ into (18) and (19), we can find B_{L}^* using binary search. The details are omitted for conciseness.

Third, K_p is an integer number, we can exhaustively search $K_p \geq 1$ to get the minimal K_p^* that satisfies $P_{\mathrm{c}} \leq \epsilon^{\mathrm{u}}/2$.

In summary, to ensure the latency and the reliability requirements during the queueing (i.e., D_m^{q} and ϵ_m^{q}) and the transmission processes (i.e., D_m^{t} and ϵ_m^{c}), the bandwidth reserved for a user in the high traffic state is B_{H}^*. For the users in the low traffic state, to ensure the latency and the reliability requirements during the random access and the transmission processes (i.e., MD_m^{p}, P_{c} and ϵ_m^{r}), the bandwidth required for each subchannel is B_L^* and the number of subchannels in the resource pool is K_p^*. As such, the required minimal bandwidth is $B_{\min}^* = K_p^* B_{\mathrm{L}}^* + N_{\mathrm{H}} B_{\mathrm{H}}^*$. The optimal solution can be obtained by the three-step method because the optimal B_{L}^*, B_{H}^* and K_p^* can be obtained from three decoupled problems.

C. Implementation and Complexity Discussions

From the orthogonal pilots of different users, the BS records the number of packets arrived in the past sample window, i.e., k. When the N-P method is applied, the BS only needs to compare k with the threshold $\hat{\gamma}$, which was derived from SPP. For the k-means method, the BS only needs to compare the

TABLE I: Parameters [13]

Parameters	Values
Number of users N	20
Maximal transmit power of a user P_{t}	23 dBm
Single-sided noise spectral density N_0	-174 dBm/Hz
Coherence bandwidth W_{c}	0.5 MHz
Number of received antenna N_{r} at BS	16
Packet size b	32 bytes
Slot duration T_{s}	0.1 ms
Overall uplink packet loss probability ϵ^{u}	1×10^{-5}
Overall uplink latency D_{u}	0.5 ms
Sample window size N_{w}	100 T_{s}
Prediction window size N_{p}	100 T_{s}
Average packet rate generated by each user	1000 packets/s [2]

distance from k to the two centroids ν_1^* and ν_2^*, which were obtained from the historical training data.

After classifications, the BS calculates the optimal B_{L}^*, B_{H}^* and K_p^* by solving (28). B_{L}^* and B_{H}^* are obtained via binary search for the user with the worst large-scale channel gain, and hence the searching complexity is low and does not increase with the number of users. K_p^* are obtained via exhaustive searching, and the searching complexity is linear with N. Therefore, the complexity for solving problem (28) is $O(N)$.

At the beginning of each prediction window, the BS assigns bandwidth to each user that is classified into the high traffic state, and reserves bandwidth in the resource pool that is shared by the users classified into the low traffic state. During the prediction window, grant-free transmission is performed. It should be noted that the resource reservation is not updated in each slot. Instead, it is updated at the beginning of each prediction window, so that the signaling overheads will not be too large.

VII. Performance Evaluation

In this section, simulations and experiments are conducted to verify the effectiveness of the proposed methods.

A. Simulations

In simulations, we first show how to choose the upper bound of the false alarm probability with the N-P method. Then, we demonstrate the required bandwidth for different methods. The proposed model-based and data-driven methods are compared with another two methods. The baseline method is reserving dedicated bandwidth according to the high packet arrival rate for all the users, which can be considered as a direct extension of existing method in [15]. The other method knows perfectly the traffic states of all the users. Such an idealistic method can obtain a lower bound of the required bandwidth. Finally, we illustrate the delay and reliability achieved by the proposed methods.

Simulation settings are listed in Table I. To ensure QoS requirements for all users, we consider the worst case of large-scale fading, i.e., the users are at the edge of the cell. The path loss model is $10\log_{10}(\alpha_m) = 35.3 + 37.6\log_{10}(d_m)$, where $d_m = 200$ m is the radius of the cell. For the k-means method,

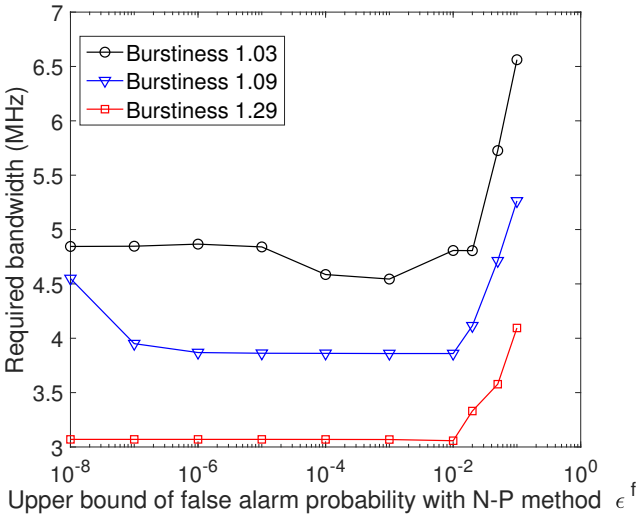

Fig. 3: Required bandwidth v.s. ϵ^{f}.

we can obtain the empirical false alarm and missing detection probabilities from historical data by comparing if the classified traffic states within the sample window and prediction window are the same.

The packet arrival processes are generated according to the SPP model in simulations. Recall that the burstiness is defined as $I_{\mathrm{s}} = \sigma^2/\nu$. To investigate the impacts of burstiness on the required bandwidth, the average packet rate of each user and the sum of average durations of the high traffic state and the low traffic state are set to be constant, i.e., $\bar{\lambda} = (\lambda_{\mathrm{H}}\mu_{\mathrm{H}} + \lambda_{\mathrm{L}}\mu_{\mathrm{L}})/(\mu_{\mathrm{H}} + \mu_{\mathrm{L}}) = 1000$ packets/s, and $\mu_{\mathrm{L}} + \mu_{\mathrm{H}} = 1$ s. The packet arrival rate in low traffic state is set to be a constant, i.e., $\lambda_{\mathrm{L}} = 10$ packets/s. We change λ_{H} from 1000 packets/s to 4000 packets/s. To keep the average packet arrival rate constant, μ_{H} decreases with λ_{H}.

To show how to choose the value of ϵ^{f} in the N-P method, we illustrate the minimal bandwidth that is required to ensure the reliability and the latency requirements with different ϵ^{f}. The results in Fig. 3 indicate that when the burstiness is given, e.g., 1.09, the minimal bandwidth first decreases and then increases with ϵ^{f}. The optimal ϵ^{f} ranges from 10^{-3} to 10^{-2}, and is much larger than the reliability requirement ϵ^{u}. Such an observation is counterintuitive. Since classification errors are considered in bandwidth reservation, the reliability can be satisfied in the case $\epsilon^{\mathrm{f}} > \epsilon^{\mathrm{u}}$. On the other hand, a small ϵ^{f} leads to a high missing detection probability. As a result, a lot of bandwidth are reserved for the low traffic users that are classified into high traffic state, and the bandwidth usage efficiency is low. In this section, we set $\epsilon^{\mathrm{f}} = 10^{-2}$.

The relation between the required bandwidth and burstiness is shown in Fig. 4. The reason why the curves in Fig. 4 are not smooth is because K_p is integer, and is not due to insufficient number of Monte Carlo simulations. The results show that for the baseline method, the required bandwidth increases with the burstiness rapidly. This is because to guarantee the QoS in different traffic states, the BS needs to reserve bandwidth by assuming that all the users are in the high traffic state. While the proposed method without prediction errors

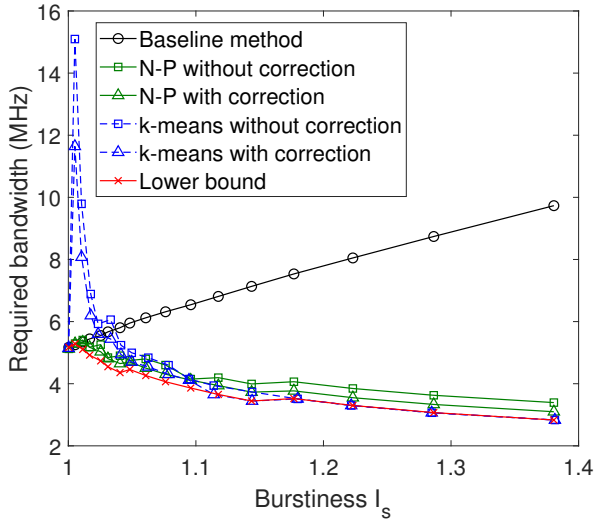

Fig. 4: Minimal bandwidth required v.s. burstiness.

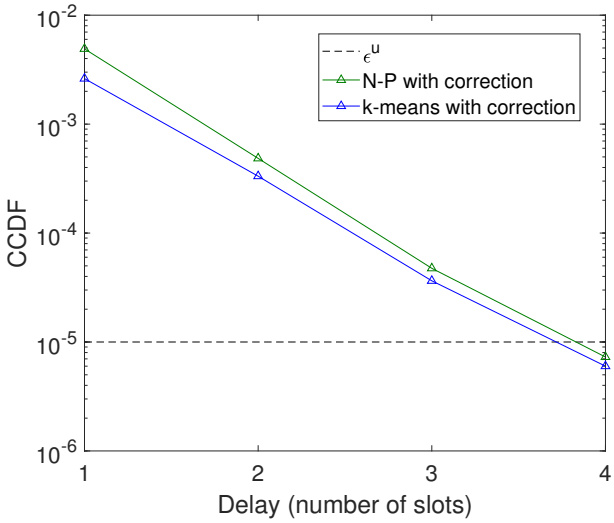

Fig. 5: CCDF of the delay.

provides the best performance and the required bandwidth decreases with the burstiness. This is because the average durations of low traffic state increase with burstiness, and bandwidth is saved by sharing the resource pool among the users in the low traffic state. With the N-P method or the k-means method, the required bandwidth also decreases with the burstiness and is close to the lower bound when the burstiness is large. This demonstrates the effectiveness of the proposed methods. Moreover, the proposed methods with error correction mechanism require less bandwidth compared with the methods without error correction. Finally, the results also indicate that when the burstiness is less than 1.1, the k-means method is worse than the N-P method; when the burstiness is larger than 1.1, the k-means method outperforms the N-P method and is very close to the lower bound. This is because the gap between the average packet arrival rates in low and high traffic states increase with burstiness, and hence the classification error probability of k-means method decreases with burstiness rapidly.

To investigate the delay and reliability achieved by the proposed methods, Fig. 5 shows the complementary cumulative distribution function (CCDF) of delay experienced by the users that are classified into the low traffic state. For the users that are classified into the high traffic state, the delay and reliability have been validated in [15], and hence is omitted in this work. In this simulation, the delay of a packet equals to the interval between the slot when the packet is generated by a user and the first slot when the packet is decoded at the BS. For example, if the first two repetitions of a packet fail, and the third repetition is successfully decoded, then the delay equals to 3 slots, where the duration of each repetition is one slot. The reliability requirement (i.e., $\epsilon^{\mathrm{u}} = 10^{-5}$) is also provided as a reference. The results in Fig. 5 show that the probability that the delay of a packet is longer than 4 slots is less than the reliability requirement. In other words, with probability $1-\epsilon^{\mathrm{u}}$ a packet can be successfully received at the BS after 4 repetitions. Therefore, the UL delay and reliability requirements in Tactile Internet can be satisfied with the proposed methods.

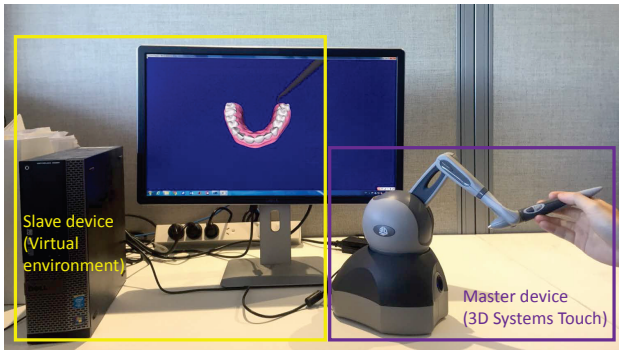

Fig. 6: Experiment to gain real packet arrival process of Tactile Internet.

B. Experiment: Bandwidth Reservation for Real Packet Arrival Processes

Since the packet arrival processes in simulation part are generated according to the SPP model, it is not clear whether our methods are still helpful when the real packet arrival processes do not match SPP perfectly. To further demonstrate the effectiveness of the proposed methods, we illustrate the required bandwidth with packet arrival processes that are generated from a real tactile hardware in Fig. 6. In this experiment, a typical application of Tactile Internet is considered (i.e., training dentists), where the dental students can use a probe to feel the hardness of the teeth. As shown in Fig. 6, a real 3D System Touch (also named Phantom Omni, Geomagic) tactile device is used as a master device. It is connected to a slave device in a virtual environment via a wire link. The slave device in the virtual environment is a probe that sends force feedback to the master device.

In the experiment, ten participants are invited to control the slave device in the virtual environment. The packet arrival process of force feedback from each participant is recorded in ten minutes. It can be observed that the traffic is very bursty, which is similar to the observations from [6]. The recorded ten packet arrival processes are used in the optimization of bandwidth reservation for ten users working simultaneously.

Next, we apply the N-P method and the k-means method in traffic state classifications. For the N-P method, we assume that the arrival process is SPP. Then, the bandwidth is reserved for the users according to their traffic states. The minimal bandwidth that is required to ensure the delay and reliability requirements is illustrated in Table II. The results in Table II indicate that the N-P method and the k-means method can save 33.9% and 43.2% of the bandwidth that is needed with the baseline method, respectively. Since the real packet arrival processes do not match SPP perfectly, the required bandwidth of the N-P method is higher than the k-means method, but is still much lower than the baseline method. The results indicate that the proposed methods outperform the baseline method with real packet arrival processes.

TABLE II: Bandwidth Reservations

Applied Methods	Total Bandwidth (MHz)	Bandwidth Saving
Baseline	1.95	N/A
N-P	1.29	33.9%
k-means	1.11	43.2%
Lower bound	0.96	50.5%

VIII. Conclusions

In this paper, we established a framework for designing burstiness aware bandwidth reservation for uplink transmissions in Tactile Internet. We first classified packet arrival processes of Tactile Internet into high and low traffic states with the model-based method and the data-driven method. Then, we designed bandwidth reservation schemes according to the traffic states of users. Furthermore, the total bandwidth is minimized by optimizing bandwidth reservation under the delay and reliability requirements. Simulation results validate the effectiveness of the proposed methods and around 40% to 70% bandwidth can be saved compared with a baseline method. Besides, when the classification error probability is around 10^{-2}, our method can achieve 10^{-5} packet loss probability. Experiment results show that if the traffic model is inaccurate, the data-driven method outperforms the model-based method.

References

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

[2] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing the Tactile Internet: haptic communications over next generation 5G cellular networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 82–89, Apr. 2017.

[3] G. P. Fettweis, "The Tactile Internet: applications & challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[4] P. Popovski, *et al.*, *Deliverable D6.3 Intermediate system evaluation results*, Aug. 2014.

[5] 3GPP TR 36.881 V0.5.0, "Evolved universal terrestrial radio access(E-UTRA)," Nov. 2015.

[6] M. Condoluci, T. Mahmoodi, E. Steinbach, and M. Dohler, "Soft resource reservation for low-delayed teleoperation over mobile networks," *IEEE Access*, vol. 5, pp. 10 445–10 455, May 2017.

[7] 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.

[8] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[9] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-based access for ultra-reliable low latency uplink transmissions," *IEEE Wireless Commun. Lett.*, Oct. 2017.

[10] S. M. Kay, *Fundamentals of statistical signal processing, Volume II: Detection theory*. Prentice Hall, 1998.

[11] M. B. Christopher, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2016.

[12] Z. Hou, C. She, Y. Li, T. Q. S. Quek, and B. Vuectic, "Burstiness aware bandwidth reservation for uplink transmission in tactile internet," in *IEEE ICC workshops*, May 2018.

[13] 3GPP TSG RAN TR38.913 R14, "Study on scenarios and requirements for next generation access technologies," Jun. 2017.

[14] S. Xu, T.-H. Chang, S.-C. Lin, C. Shen, and G. Zhu, "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.

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

[16] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication (URLLC) using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2017.

[17] M. Takeya, Y. Kawamura, and S. Katsura, "Data reduction design based on delta-sigma modulator in quantized scaling-bilateral control for realizing of haptic broadcasting," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1962–1971, Mar. 2016.

[18] S. Zhang, Z. Qian, Z. Luo, J. Wu, and S. Lu, "Burstiness-aware resource reservation for server consolidation in computing clouds," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 4, pp. 964–977, Apr. 2016.

[19] M. E. Ahmed, J. B. Song, Z. Han, and D. Y. Suh, "Sensing-transmission edifice using Bayesian nonparametric traffic clustering in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 27, no. 4, pp. 964–977, Apr. 2016.

[20] C. She, C. Yang, and T. Q. S. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Trans. Commun., early access*, 2018.

[21] 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.

[22] J. Wu, Y. Bao, G. Miao, S. Zhou, and Z. Niu, "Base-station sleeping control and power matching for energy–delay tradeoffs with bursty traffic," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3657–3675, May 2016.

[23] 3GPP TR 38.802 V2.0.0, "Study on new radio (NR) access technology; physical layer aspects (release 14)," 2017.

[24] 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 Workshops*, Dec. 2015.

[25] M. Mousaei and B. Smida, "Optimizing pilot overhead for ultra-reliable short-packet transmission," in *Proc. IEEE ICC*, May 2017.

[26] C. S. 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.

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