Improved Grant-Free Access for URLLC via Multi-Tier-Driven Computing: Network-Load Learning, Prediction, and Resource Allocation

Zixiao Zhao, *Graduate Student Member, IEEE*, Qinghe Du, *Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*— Grant-Free (GF) access has been recognized as a promising candidate for Ultra-Reliable and Low-Latency Communications (URLLC). However, even with GF access, URLLC still may not effectively gain high reliability and millimeter-level latency, simultaneously. This is because the network load is typically time-varying and not known to the base station (BS), and thus, the resource allocated for GF access cannot well adapt to variations of the network load, resulting in low resource utilization efficiency under light network load and leading to severe collisions under heavy network load. To tackle this problem, we propose a multi-tier-driven computing framework and the associated algorithms for URLLC to support users with different QoS requirements. Especially, we concentrate on K- repetition GF access in light of its simplicity and well-balanced performance for practical systems. In particular, our framework consists of three tiers of computation, namely *network-load learning*, *network-load prediction*, and *adaptive resource allocation*. In the first tier, the BS can learn the network-load information from the states (success, collision, and idle) of random-access resources in terms of resource blocks (RB) and time slots. In the second tier, the network-load variation is effectively predicted based on estimation results from the first tier. Finally, in the third tier, by deriving and weighing the failure probabilities of different groups of users, their QoS requirements, and the predicted network loads, the BS is able to dynamically allocate sufficient resources accommodating the varying network loads. Simulation results show that our proposed approach can estimate the network load more accurately compared with the baseline schemes. Moreover, with the assistance of network-load prediction, our adaptive resource allocation offers an effective way to enhance the QoS for different URLLC services, simultaneously.**

***Index Terms*— URLLC, grant-free access, multi-tier-driven computing, network-load estimation, adaptive resource allocation.**

Manuscript received 15 May 2022; revised 6 September 2022; accepted 25 October 2022. Date of publication 5 January 2023; date of current version 16 February 2023. This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1807700. An earlier version of this paper was presented in part at the IEEE PIMRC 2021 [DOI: 10.1109/PIMRC50174.2021.9569425]. *(Corresponding author: Qinghe Du.)*

Zixiao Zhao and Qinghe Du are with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Shaanxi Smart Networks and Ubiquitous Access Research Center, Xi'an 710049, China (e-mail: zzx67120787@stu.xjtu.edu.cn; duqinghe@mail.xjtu.edu.cn).

George K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54 124 Thessaloniki, Greece (e-mail: geokarag@auth.gr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2023.3234696.

Digital Object Identifier 10.1109/JSAC.2023.3234696

I. Introduction

ULTRA-RELIABLE Low-Latency Communications (URLLC), together with enhanced Mobile Broadband (eMBB) and massive Machine-Type Communications (mMTC) are the three main application scenarios of the fifth generation (5G) of mobile communications networks. As the 5G had inspired widespread research efforts, some researchers begin to envisage the next steps in wireless networking [1], [2], [3]. In beyond 5G (B5G) and 6th generation (6G), URLLC will keep evolving to its advanced version, while still encountering many challenges. The typical quality-of-service (QoS) of URLLC services require the user-plane latency between the base station (e.g., evolved Node B (eNB) in 4G and the next generation Node B (gNB) in 5G) and the user to be confined within 1 ms. In the meantime, for transmission of short packets, the reliability needs to be guaranteed with a probability equal to 99.999% [4].

A. State-of-the-Art

The overall latency mainly comes from several factors, consisting of handshake procedures in random access, scheduling latency introduced by the base station (BS), retransmission in case of collision, transmission delay, hardware processing delay at the receiver, etc. Aside from transmission delay and hardware processing delay, which can already be confined within 0.5 through 1 ms currently, even delay caused by the standard handshake procedures for random access [5] will inevitably exceed 1 ms. In order to shorten the overall latency, Cheng et al. [6] proposed an adaptive block-length transmission framework considering the tradeoff between the queuing delay and the transmission delay. In [7], Qiao et al. derived the maximum throughput that can be supported under statistical queuing delay constraints. In [8], Gu et al. analyzed the effective capacity for machine-type communications with statistical delay constraints. The above research efforts mainly concentrated on the queuing delay. However, the most challenging concerned in URLLC lies in the delay introduced during the random access phase rather than the transmission phase. Moreover, URLLC typically serves the short-packet yet sparse transmissions for each user, the queuing delay of each user's transmission does not play an essential role in contributing to the overall latency.

To assure millisecond-level latency, URLLC typically employs grant-free (GF) access mode [9], [10], where the delay can be significantly shortened by avoiding too many handshake procedures. Typical GF access approaches include the Reactive scheme [11], [12], the K-repetition scheme [12], [13], [14], and the Proactive scheme [12], [14], which use redundant transmissions (retransmission and/or repetition transmission) to combat collisions and improve the reliability. In the reactive scheme, retransmission begins when a negative feedback from the BS is received by the user. In the K-repetition scheme, each packet is directly and repeatedly transmitted K times within a subframe over K different resource units (e.g., K resource blocks), which can well tradeoff between the reliability and retransmission delay. In the proactive scheme, the BS immediately notifies the user at once upon the successful packet arrival, such that the repetition transmission can stop as early as possible and the occupied random-access resources can be released.

Based on the existing research and results, it is evident that GF access integrates random access and data transmission together and thus avoiding time-consuming handshakes and shortening waiting delay for BS's scheduling information. Yet GF access still faces the essential issues in random access, i.e., collisions across users, especially given the fact that the network load of URLLC is time varying. As the network load is typically not known to the system, the resource allocated for random access of URLLC is fixed. If too many resources are reserved, the utilization efficiency would be very poor; on the contrary, if random-access resources are insufficient, frequent collisions under bursty requests will harm the reliability and prolong the delay by retransmission. Consequently, the key to effectively assure reliability and lower delay is to allocate sufficient resources to URLLC well matching the network load, i.e., the number of active users. Therefore, it is highly desirable to develop network-load estimation techniques and then enable adaptive resource allocation for URLLC, benefiting both latency and reliability quality-of-service (QoS).

There have been some research focusing on network-load estimation in random access, but mainly for machine-to-machine (M2M) communications or mMTC [15], [16], [17]. Particularly, reference [15] proposed the traffic-load estimation framework and scheme based on the Markov Chain model. Reference [16] derived the joint PDF of the number of successful and collided access attempts. Reference [17] proposed the estimation approach through minimizing the Euclidian distance between the theoretical means and the observed preamble access states. However, it is worth noting that in M2M or mMTC networks, random backoff strategies are applied as response to collisions. Consequently, the correlation between access status across adjacent time slots is weakened. In contrast, for URLLC, repetition and immediate retransmission in fact introduce much stronger correlation across adjacent time slots, making estimation approaches for mMTC less inaccurate for URLLC.

Multi-tier computing has been regarded as an open topic and powerful tool to attain excellent network performance. Multi-tier computing concept can be either fit for task optimization (e.g., scheduling, caching, power allocation, and/or offloading) [18], [19] or used for network topology partition [20]. Joint optimization for energy, cost, computing, storage resources, etc., in communications system has been widely discussed with multi-tier approaches. Reference [21] considered task offloading and green energy scheduling simultaneously in multi-tier edge-computing systems. The task offloading needs to search suitable edge servers owning sufficient computing capacity, while the green energy scheduling aims to minimize the system cost by configuring computing resources effectively across different tiers. It provides a competent joint optimization algorithm in the multi-tier architecture, especially for the scenarios of mobile edge computing. Reference [22] investigated user scheduling with the target of cost-minimization in multi-tier fog-computing networks. The cost studied in this paper mainly includes downloading delay from service nodes to end users and payments charged by service nodes. Reference [23] proposed edge caching methods based on the numbers of times to request and cache, respectively. Simulation results verify that these caching schemes can well reduce the energy consumption and improve backhaul rates, and thus will assist to carry more task loads among fog nodes.

Moreover, due to the fact that joint optimization typically requires to be executed timely, some researchers paid more attention to additional latency introduced by cross-layer computing. Reference [24] discussed allocation of both the resource and tasks in the multi-tier fog-cloud ecosystems under hard constrains of computing resource and battery consumption. The ecosystem mentioned in [24] consists of mobile devices, some proximate fog nodes, and a remote cloud node. The experiment results indicates the proposed algorithm can achieve better energy and latency performances in the ecosystem, compared with some state-of-the-art benchmark solutions. Reference [25] proposed a SDN-based multi-tier architecture for healthcare applications, in which both edge computing and cloud computing were incorporated and the software defined networking (SDN) was employed for controlling across multiple function layers. The authors evaluated the architecture via demonstration based on real data collected from patients, conducting a real machine learning-based fall risk assessment application. The results show that by offloading computing and storing tasks to different function tiers, this architecture can achieve high accuracy and low latency specifically targeted for healthcare applications. Reference [26] proposed an adaptive decision-making framework for federated learning (FL) from the global perspective, which was proved to be capable of decreasing the average delay in multi-tier computing.

The powerful capability of multi-tier computing in network optimization comes from the idea of decoupling the serving function into sub-functions over. This idea as well as the multi-tier computing architecture can not only efficiently serve cloud and edge computing architectures, but also fit diverse networking optimization scenarios. We in this paper show that multi-tier computing can be designed to effectively combat the aforementioned existing problems in random access of URLLC. In the meantime, the decoupled multi-tier functions can be equipped at different levels of network nodes, and

therefore in nature our design well aligns with the structure of multi-tier computing architecture of cloud and edge networks.

B. Contribution

Motivated by previous discussion, we propose a multi-tier-driven computing framework and the associated algorithms for GF random access in URLLC. The multi-tier-driven computing framework consists of three tiers, namely, *network-load learning*, *network-load prediction*, and *adaptive resource allocation*. In this paper, we concentrate on K-repetition GF access in light of its simplicity and well-balanced reliability and delay performances, which are highly desirable to practical systems. The idea behind the three-tier computing framework comes from the following principles. The random access resources (e.g., resource blocks) will be randomly selected by users. Consequently, the access states (success, collision, or idle) of resource blocks somehow will carry the information of network load (the active number of access users) in an implicit and hidden manner, which, however, is often neglected. Following this thought, the first-tier computing is designed to learn network-load information from the access states of resource blocks. Then, with the extracted network-load information and the recorded history data, the network load in the coming time slot can be forecast via the second-tier computing. The predicated results will then be injected into the third-tier computing to yield the amount of resources required to accommodate the coming network load, such that QoS assurance for URLLC can be well fulfilled. Also conducted is a set of simulation results to verify the superiority of our proposal compared with the existing baseline approaches.

The main contributions of this paper are summarized as follows:

- We propose the three-tier computing framework for URLLC adopting K-repetition access, serving for network-load learning, prediction, and resource allocation, respectively, towards effectively accommodating the varying access load and fulfilling all users' access with stringent yet differentiated QoS requirements. This framework divides the challenging task, i.e., solving for the resource amount needed to support users' QoS under the unknown network load, into three tractable tiers
- We propose a spectrum of estimation schemes for network-load learning over URLLC based on the access states (success, collision, or idle) of resources blocks, which are suitable for two variant modes of K-repetition GF access, termed adjacent-occupation K-repetition access and arbitrary-occupation K-repetition access. Simulation results demonstrated the superiority of our proposed approach compared with the existing baseline schemes.
- We design an adaptive resource allocation scheme driven by differentiated QoS requirements. In particular, we successfully derive the analytical expressions of access -failure probability within 1 ms for the two K-repetition access schemes as a function of network load. Then, the system can precisely allocate resources to different types of services according to their QoS requirements and predicated traffic loads. We validate our proposed algorithms via abundant simulations under various network conditions.

Part of our research work reported in this paper was presented in [27], and the differences compared with our preliminary research in [27] are summarized as follows: 1) Based on the preliminary research design in [27], this paper develops a multi-tier computing framework, which consists of three function layers, namely, traffic-load learning, traffic-load prediction, and resource allocation. 2) A more flexible yet more efficient K-repetition GF access mode, namely, arbitrary-occupation K-repetition mode, is studied in this paper, especially on its load estimation scheme, while in [27] we merely discussed a simple mode called adjacent-occupation K-repetition. 3) The load-prediction function is enhanced with better accuracy by statistical model considering the history data and random error, while the work in [27] taking the current value as the predicted result. 4) The analytical expressions of failure access probability are newly derived, which can help to quickly calculate the required amount of resources. 5). The allocation strategy considering stringent and differentiated QoS requirements of events is newly discussed in this paper. The previous work in [27] concentrated on services with only single type of QoS requirement.

C. Structure

The rest of the paper is organized as follows. Section II describes the system model and proposes the three-tier computing framework for URLLC. Section III proposes a spectrum of network-load estimation schemes in the first-tier computing for K-repetition access based URLLC. Section IV concentrates on the second-tier computing and presents the network-load prediction schemes. Section V derives the third-tier computing to develop the adaptive resource allocation scheme. Section VI presents the simulation results. The paper concludes with Section VII.

II. System Model

A. System Description

We consider a multi-user network for URLLC which is composed of N_{al} users and one base station (BS). These users can be in only two states, i.e., *active* and *inactive* states. The number of active users is represented as N_{tr}. The users access and transmit in a synchronized but grant-free (GF) manner coordinated by the BS. It is well-known that URLLC requires high successful access probability and low latency, which significantly relies on whether there are sufficient resources allocated for GF access compared to N_{tr}. However, N_{tr} is typically unknown to the BS, thus causing the major hurdle to release the potentials of GF access. In this paper, we concentrate on estimating N_{tr}, then enabling adaptive resource allocation for better supporting URLLC with assured QoS.

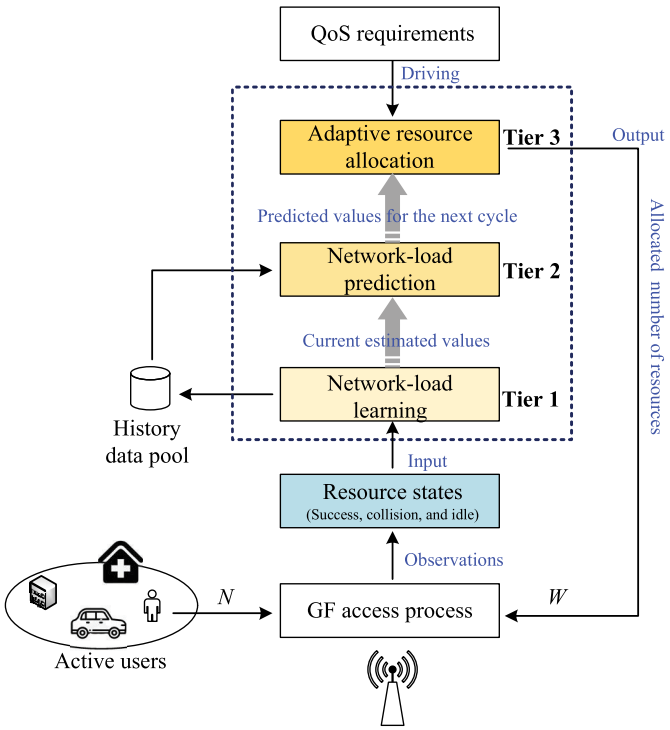


Fig. 1. Multi-tier-driven computing framework, which consists of three function layers, namely, traffic-load learning, traffic-load prediction, and resource allocation.

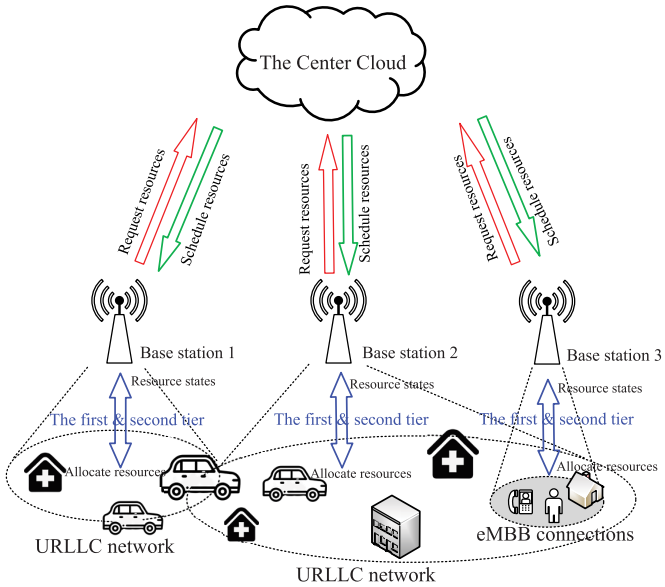


Fig. 2. Center cloud will proactively schedule for different base stations sharing the overall resources. If a BS has sufficient autonomy in resource allocation, our framework can be implemented locally. Otherwise, the BS needs to forward and wait for the center cloud's instructions.

B. The Multi-Tier-Driven Computing Framework

The access states (success, collision, or idle) of resource blocks can include hidden information of network-loads, which, however, are typically neglected. In this paper, we propose a multi-tier-driven computing framework which consists of three tiers, i.e., network-load learning, network-load prediction, and adaptive resource allocation, in order to assure the QoS requirements of URLLC applications. The network-load learning is implemented via a Markov-chain model based estimation technique, and thus in this paper we use the terminologies *network-load learning* and *network-load estimation* interchangeably. As shown in Fig. 1, firstly the BS can get the knowledge of resource states in every slot (success, collision, and idle). Based on these observations, we can get the estimated number of current active users via the first tier. The estimated values will be recorded in the history data pool and also utilized as the input together with a selected series of history data to the second tier. Based on the prediction values, the third tier will adaptively allocate resources driven by different QoS requirements. In particular, we propose several network-load estimation schemes for adjacent-occupation K-repetition and arbitrary-occupation K-repetition, respectively. We also employ the auto-regressive integrated moving average (ARIMA) model to achieve accurate and timely predictions. With the assistance of analytical formulations towards access failure probability, we design a resource negotiation algorithm driven by QoS requirements.

Typically, the entire network needs to dynamically allocate resources in order to better support the varying needs of diverse connections, e.g., wireless connections for self-driving, factory automation, and access to multimedia services. It is composed of the center cloud, a BS which performs as a edge server, and end users. Though our proposed multi-tier computing framework can provide suggestions to the allocated amount of resource for users located within each single BS's coverage, the realistic allocation policies may still be controlled by the center cloud globally. In particular, while the first and second tier of our framework can be implemented in each single BS of the network, the functions of third tier can be equipped in different entities, including the following two cases. If a BS has sufficient autonomy in resource allocation, then the computing function at the third tier of our framework can be implemented locally in this BS via dynamic slicing functions. Otherwise, the BS needs to forward all related information to the cloud center, and wait for cloud center's instructions on resource allocation policies, which jointly consider requests and burdens across several adjacent BSs sharing the resources. As shown in Fig. 2, the BS 1 and BS 2 serve for two URLLC networks covered by different BSs, where the two BSs' coverage has overlapping area, and BS 3 serves for the eMBB connections. A mobilizing automatic car is located within the overlapping area of BS 1 and BS 2 which requires stable connection performances to maintain service quality. For handling this bursty condition, BS 2 will forward to the cloud center and request for higher priority in resource allocation. Accordingly, the center cloud will proactively reduce the allocated amount of resources for BS 3, which shares the resources with BS 2 and has lower priority, and schedule larger amount of resources to BS 2.

C. K-Repetition Access and Transmissions

We will denote a generalized concept in the following description by resource block (RB), e.g., a RB includes 12 consecutive subcarriers in LTE and 5G NR system. The BS divides time into consecutive subframes.[1] Each subframe is dedicated

[1] In practical networks, a number of consecutive subframes typically together form a frame.

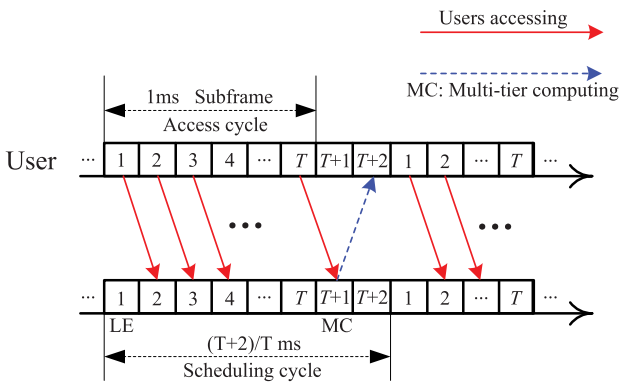


Fig. 3. Illustration of K-repetition, access cycle, and scheduling cycle.

for a GF *access cycle* and is further divided into T slots[2] with equal length. Reference [28] proposed a fundamental slot assignment protocol, and it defined the first slot of every frame as *broadcast slot* for transmitting controlling message and also exchanging information among network nodes. Reference [29] discussed the application of this protocol in dynamic resource allocation and have raised widespread attention [30], [31], [32]. We will follow the setting of broadcast slot mentioned above, and for better description, we present the broadcast slots and other regular transmitting slots separately, as shown in Fig. 3. Between two adjacent subframes, two slots are defined for broadcasting the available resource information for next cycle to users. Accordingly, the term *scheduling cycle* is used to denote an access cycle and the two followed slots. Please note that these definitions are merely logical division, thus the BS can flexibly configure the function of slots, i.e., transmitting or broadcasting, by following the cycle manner. Several GF access approaches have been proposed by researchers, and typical ones include Reactive, K-Repetition, and Proactive [12]. In this paper, we mainly concentrate on the K-repetition scheme, considering its simplicity and well-balanced performances. In the K-repetition scheme, the BS allows every user to repeatedly access and transmit K times in consecutive slots (i.e., *adjacent-occupation*) or arbitrary but different slots (i.e., *Arbitrary-occupation*) of an access cycle. In every access slot, each user is permitted to occupy only one RB. If a RB is occupied by only one user in a slot, this access is successful. Otherwise, if there are two or more users occupying the same RB in a slot, they all happen to collide and fail. Only when all the replicas fail in a cycle, this user needs to retry in the following cycles. Without losing generality, 1 ms will be divided into 8 slots in this paper, and the length of one slot is equal to 0.125 ms [33]. Thus the largest value of K can be set as 8, i.e., a user will be allowed to access in the all slots of an cycle at most.

D. Statistical Feature of URLLC

The general URLLC requirement for one transmission of a packet is 99.999% reliability and latency within 1ms. A typical type of URLLC application is in the scenario of sophisticated industrial production and controlling, in which the machine-type devices, e.g., sensors and controllers, always call for communication services with high reliability and low latency. The industrial internet of things (IIoT) applications with URLLC requirements can be classified into two main types of use cases [34], [35]: motion control and discrete automation. For motion control with continuous and stable data transmission, e.g., automatic production machine tools and 3D printers, this type of services can be regarded as uniform and periodic pattern [36], which has stable access intensity. While for discrete automation, it always has difficulty in accurately predicting when the packets flow will arrive, especially in vehicle to everything (V2X) communications [37].

Fortunately, the internal activation order of a batch of bursty devices can follow some distributions. In particular, the Beta distribution is one of the suitable types to model this process [37], [38], [39], in which all of the devices access to the BS in a deterministic order during a limited period [40]. The Beta distribution in bursty traffic has been defined in [40]:

$$N_i = N \int_{t_{i-1}}^{t_i} p(t)dt, \tag{1}$$

in which N_i denotes the number of users in the ith access cycle of the whole event duration.

In this model, $t_i - t_{i-1}$ is equal to the length of a complete scheduling cycle. Supposing that each subframe is divided into 8 slots in this paper, if the duration of Beta distribution is set as 12.5 ms, we have $i \in \{\mathbb{Z} \mid [1, 10]\}$ considering two scheduling slots in every scheduling cycle.

The $p(t)$ in (1) is derived by:

$$p(t) = \frac{t^{\alpha-1}(T-t)^{\beta-1}}{T^{\alpha+\beta-1}Beta(\alpha,\beta)} \tag{2}$$

in which $Beta(\alpha,\beta) = t^{\alpha-1}(1-t)^{\beta-1}$. Following the common setting in 3GPP release [40], we have $\alpha = 3$ and $\beta = 4$ in this paper. Hereinafter we will use $\mathcal{B}(N,T)$ to represent that there are N users in total to access in T ms with bursty traffic pattern. The parameters used in this paper are summarized in Table I.

III. Network-Load Estimation

The adjacent-occupation K-repetition has strict requirements on the slot selection for single user, while the arbitrary occupation K-repetition allows each user to select any K different slots for access. We propose two estimation schemes for adjacent-occupation K-repetition in Section III-A and III-B, and one estimation scheme for arbitrary-occupation K-repetition in Section III-C, respectively. We list the suitability of each scheme in Table II.

A. Single-Slot Maximum Likelihood With Least Squares Estimation

In the Single-slot Maximum Likelihood with Least Squares Estimation (SS-ML-LS), we focus on the relationship between the numbers of users beginning their first access and total active users, which are denoted by $\phi_r (1 \leqslant r \leqslant T-K+1)$

[2]It is often referred to as mini-slot in 5G. But in this paper, we use the term slot in short for simplicity.

TABLE I
PARAMETERS AND NOTATIONS

Parameters	Descriptions
N_{al}	The total number of users
N_{tr}	The number of active users who attempt to access
$\widehat{N}_{\mathrm{tr}}$	The estimation of N_{tr}
N	The test number of users for hypothesis in ML
E	The prediction of the number of users in next 1ms
$\widehat{n}_t$	The estimation of the number of users who access in the t-th slot
$\mathbf{n}$	The load estimation vector: $\mathbf{n} = (\widehat{n}_1, \widehat{n}_2, \ldots, \widehat{n}_T)^A T$
ϕ_r	The number of users beginning their first access in the r_{th} slot
$\boldsymbol{\phi}$	The starting vector: $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_{T-K+1})^T$
A	The number of RBs each of which is selected by only one user to access, suggesting successful access
B	The number of RBs each of which is selected by two or more users to access, suggesting collision
C	The number of RBs each of which is selected by any users, suggesting idle states
(A, B, C)	The Markov state denoting the number of success RBs, collision RBs and empty RBs, respectively
$\mathbf{P}$	Markovian transition matrix
$\mathbf{S}$	Markovian space for all RBs' states
W	The total number of available RBs
K	The parameter for K-repetition: the number of consecutive slots a user attempt to send access signals
T	The total number of slots within a subframe
α	The parameter of $Beta$ function
β	The parameter of $Beta$ function

TABLE II
SUITABILITY OF NETWORK-LOAD ESTIMATION SCHEMES

K-repetition type	Estimation schemes		
	SS-ML-LS	MS-MLI	MS-MLD
Adjacent-occupation	✓	✓	
Arbitrary-occupation			✓

and $n_t (1 \leqslant t \leqslant T)$, respectively. The n_t denotes the total number of active users in the t-th slot, and the ϕ_r denotes the number of users who begin their first access in the rth slot. Thus we use this equation set below to describe the relationship:

$$n_t = \begin{cases} \sum_{k=1}^{t} \phi_k, & \text{if } 1 \leq t < K; \\ \sum_{k=t-K+1}^{t} \phi_k, & \text{if } K \leq t \leq T-K+1; \\ \sum_{k=t-K+1}^{T-K+1} \phi_k, & \text{if } T-K+1 < t \leq T. \end{cases} \tag{3}$$

Note that users shall initiate the first access before and excluding $T-K+2$ th slot, otherwise they are not able to finish K times in an access cycle.

To obtain the total amounts of users in a cycle, firstly we target at the number n_t of users in every slot. We denote the number of success RBs, collision RBs, and empty RBs observed by the BS in every slot by A, B, and C, respectively. Definitely, here is $A+B+C = W$. Though this is an established fact generated by users' RB choices, it is a new perspective to regard the process as a Markov model with N-steps transition, in which every user's RB choice is related to one step of Markov transition. The detailed derivation of n_t can be found in Appendix A.

We further define $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_{T-K+1})^T$, $\mathbf{n} = (n_1, n_2, \ldots, n_T)^T$, and use $\boldsymbol{\Omega}$ to denote a $(T \times (T-K+1))$ matrix. The detailed expression of $\boldsymbol{\Omega}$ can be found in Appendix B. Then we have an overdetermined equation as follows:

$$\boldsymbol{\Omega}\boldsymbol{\phi} = \mathbf{n}, \tag{4}$$

After obtaining the estimation $\mathbf{n}$ with ML, we can search an approximate solution for (4):

$$\boldsymbol{\phi} = \left(\boldsymbol{\Omega}^T\boldsymbol{\Omega}\right)^{-1}\boldsymbol{\Omega}^T\mathbf{n}. \tag{5}$$

Finally, we obtain the total number of users with ϕ_r as follows:

$$\widehat{N}_{\mathrm{tr}} = \sum_{r=1}^{T-K+1} \phi_r\,. \tag{6}$$

B. Multi-Slot Maximum Likelihood Indirect Estimation for Adjacent-Occupation K-Repetition

In the Multi-Slot Maximum Likelihood Indirect Estimation (MS-MLI), we consider solving this problem in multi-slot which turns out to be more accurate. Firstly, we use (A_t, B_t, C_t) to denote the access states of RBs in the t-th slot, and then we use three vectors to record these states respectively, i.e., $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. For example, $\mathbf{A} = (A_1, A_2, \ldots, A_T)^T$. Then we use $\phi_j (0 \leqslant j \leqslant T-K+1)$ to denote the amount of users beginning their first access in the jth slot. Particularly, we define $\phi_0 = 0$. The vector $\boldsymbol{\phi}$ represents $\{\phi_0, \phi_1, \ldots, \phi_{T-K+1}\}$. Then the hypothesis N, which denotes the total number of users in ML estimation, determines all the possible $\boldsymbol{\phi}$ accordingly. The probability mass function (PMF) of every vector $\boldsymbol{\phi}$ can be calculated by:

$$Pr\left\{\boldsymbol{\phi} \,|\, N\right\} = \frac{1}{(T-K+1)^N} \prod_{r=1}^{T-K+1} \binom{N - \sum_{j=0}^{r-1}\phi_j}{\phi_r}. \tag{7}$$

The $\boldsymbol{n}(\boldsymbol{\phi})$ is utilized to denote the relationship between $\boldsymbol{\phi}$ and $\mathbf{n}$ which has been derived in (3), i.e., a vector $\boldsymbol{\phi}$ will correspond to determined number $n_t (1 \leqslant j \leqslant T)$ of users accessing in each slot. Next, the transition probability from the initial state $(0, 0, W)$ to the terminal state (A_t, B_t, C_t) with n_t steps can be calculated by (25). The difference is that here n_t is definite under the restriction of each $\boldsymbol{\phi}$, thus we will not need to discuss ML for n_t. Finally, we have the probability of hypothesis number N of users with Total Probability Theorem, and the result of ML can be derived by:

$$\begin{aligned} \widehat{N}_{\mathrm{tr}} &= \arg\max_N P\big((\mathbf{A}, \mathbf{B}, \mathbf{C}) \,|\, N\big) \\ &= \arg\max_N \sum_{\boldsymbol{\phi}} Pr\Big\{\big(\mathbf{A}, \mathbf{B}, \mathbf{C}\big) \,|\, \boldsymbol{n}(\boldsymbol{\phi})\Big\}\, Pr\left\{\boldsymbol{\phi} \,|\, N\right\} \\ &= \arg\max_N \sum_{\boldsymbol{\phi}} \left\{\prod_{t=1}^{T} Pr\big((A_t, B_t, C_t) \,|\, n_t\big)\right\} Pr\left\{\boldsymbol{\phi} \,|\, N\right\} \\ &. \end{aligned} \tag{8}$$

Thus we infer this scheme as an indirect estimation in that it utilizes mediate information, i.e., the number of users who begin their first access in every slot.

C. Multi-Slot Maximum Likelihood Direct Estimation for Arbitrary-Occupation K-Repetition

In the Multi-Slot Maximum Likelihood Direct Estimation (MS-MLD), a single user is still considered as a transition step in Markov model, but will cause the states of K RBs to change at one time. The state space $\mathbf{S}$ of Markov model containing all the possible states can be formulated with following described rules:

$$\begin{cases} a+b+c=WT; \\ 0 \leqslant a \leqslant N; \\ 0 \leqslant b \leqslant \frac{N}{2}; \\ 0 \leqslant c \leqslant W, \end{cases} \tag{9}$$

where (a,b,c) represent the possible states, and other a, b and c obeying the rules are not allowed in $\mathbf{S}$.

For an arbitrary state (A,B,C) in $\mathbf{S}$, we use $(\widetilde{A},\widetilde{B},\widetilde{C})$ to denote the possible next state, which must satisfy the following relationships:

$$\begin{cases} A-\min(K,A) \leqslant \tilde{A} \leqslant A+\min(K,C); \\ B \leqslant \tilde{B} \leqslant B+\min(K,A); \\ C-\min(K,C) \leqslant \tilde{C} \leqslant C. \end{cases} \tag{10}$$

Considering the rule of arbitrary-type repetition, we can also have:

$$\begin{cases} x &= \widetilde{B}-B; \\ y &= C-\widetilde{C}; \\ z &= K-(x+y), \end{cases} \tag{11}$$

where $x,y,z \geqslant 0$. It can be thought that the single user chooses x RBs belonging to A success RBs and causes the number of collision RBs to increase from B to $\widetilde{B}$; chooses y RBs belonging to C idle RBs and causes the number of idle RBs to reduce from C to $\widetilde{C}$; chooses z RBs belonging to B collision RBs respectively. Thus for $(\widetilde{A},\widetilde{B},\widetilde{C})$ calculated by (10), the (11) will filter suitable states again.

The transition probability can be formulated as follows:

$$\begin{cases} P_{(A,B,C)\to(\widetilde{A},\widetilde{B},\widetilde{C})} &= \dfrac{\binom{A}{x}\binom{C}{y}\binom{B}{z}}{\binom{WT}{K}}, \\ P_{(A,B,C)\to others} &= 0 \end{cases} \tag{12}$$

in which $x \in [0,A]$, $y \in [0,C]$, and $z \in [0,B]$.

Thus we can calculate the probabilities of transferring from the initial state $(0,0,WT)$ to the final observed state via N-steps, and the estimated number $\widehat{N}_{\mathrm{tr}}$ of active users is related to the maximum one, which can be depicted follows:

$$\begin{aligned} \widehat{N}_{\mathrm{tr}} &= \arg\max_{N} P\big((A,B,C)\,|\,N\big) \\ &= \arg\max_{N} \mathbf{P}^{N}_{(0,0,WT)\to(A,B,C)}. \end{aligned} \tag{13}$$

Thus we refer this scheme as a direct estimation because it directly considers each user's K choices and the corresponding transition of resource states, without benefited by the information of vector $\boldsymbol{\phi}$ which represents the number of users who begin their first access in every slot.

IV. Network-Load Prediction

The first tier of the proposed computing framework achieves network-load estimation by learning from the current resource states. The third tier should calculate the suitable number of allocated resources for the next access cycle, which needs to be based on the knowledge of the future load. Thus the second tier, i.e., network-load prediction, can function as a bridge from the current load to the future load, by fully considering the history load estimated by the schemes of Section III.

Though we have mentioned that it's unrealistic to accurately predict for bursty traffic when the occasional event will happen, the prediction module should have ability to timely response once it detects the beginning, which is the critical basis of resource allocation. There are some common schemes of time series prediction, including simple equal, moving average, exponential smoothing, and machine learning. The simple equal scheme assumes that the next expected value is equal to the current observed value, which always falls behind the real change. The machine learning schemes typically require training process, thus will inevitably raise challenges in model choosing and time complexity.

Then we employ the auto-regressive integrated moving average (ARIMA) model [41], considering it combines simplicity in mathematics and good performances in prediction. In particular, the estimated load derived by the first tier will be added into the history data pool. Then the ARIMA model will predict for the next access cycle based on this history data pool. Obviously, this data pool is updated in real time manner, thus it can provide valid information for the prediction model.

Then we will introduce the mathematical expression of ARIMA model and discuss the selection of parameters p, d, and q. Use n^{raw}_{t-i} to denote the raw data in history pool. The parameter d denotes the number of times the raw data need to be differenced until the sequence becomes stationary. We represent the difference process in this formulation:

$$n_{t-i} = \left(1-B^{d}\right) n^{\mathrm{raw}}_{t-i}, \qquad (\ i\ >\ 0\ ) \tag{14}$$

where the $\left(1-B^{d}\right)$ is a linear operator which represents d-order difference. For instance, when $i=1$ and $d=2$, we have $\left(1-B^{2}\right) n^{\mathrm{raw}}_{t-1} = n^{\mathrm{raw}}_{t-1} - 2n^{\mathrm{raw}}_{t-2} + n^{\mathrm{raw}}_{t-3}$. Then the processed n_{t-i} will be used for predictions directly.

An ARIMA (p,d,q) model can be formulated as follows:

$$\widetilde{n}_t = \underbrace{\sum_{i=1}^{p} \phi_i n_{t-i}}_{\mathrm{AR}(p)} + \underbrace{\sum_{j=1}^{q} \theta_j \varepsilon_{t-j}}_{\mathrm{MA}(q)} + \varepsilon_t. \tag{15}$$

The model predicts the load $\widetilde{n}_t$ with the past data $n_{t-i}\,(1 \leqslant i \leqslant p)$ and the random error $\varepsilon_{t-j}\,(1 \leqslant j \leqslant q)$. The AR and MA are the abbreviations of auto-regressive and moving average. The parameter p in AR(p) model denotes the number of history items n_{t-i} that have been differenced d times. The parameter q in MA(q) model denotes the number

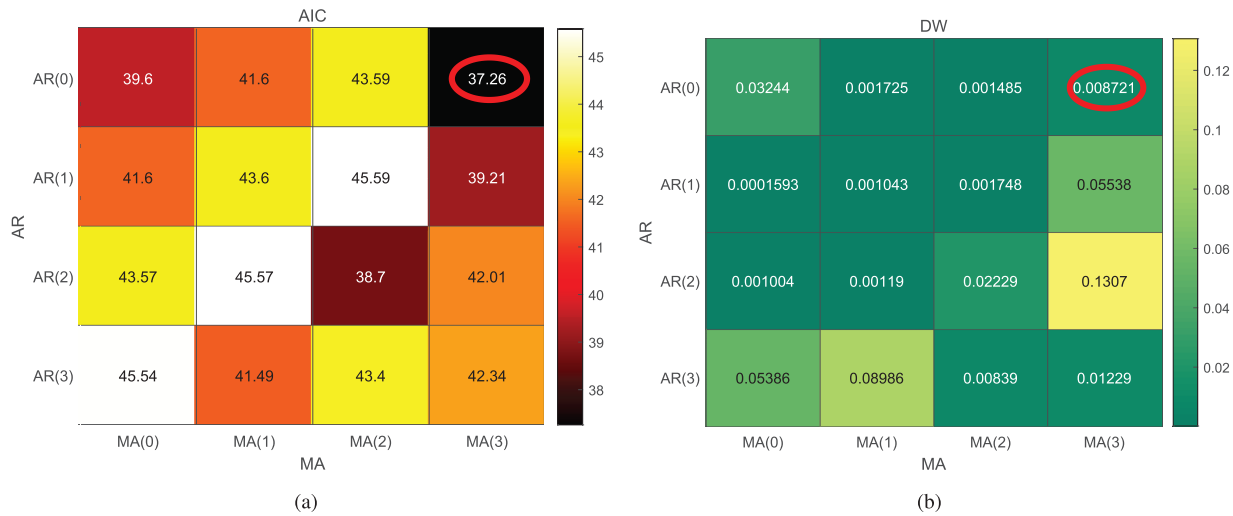


Fig. 4. Prediction performances of ARIMA model using different metrics. (a) uses the AIC metric. (b) uses the DW metric, and all the values have been minus 2 and taken absolute. According to the definitions of AIC and DW, the lower value represents the better performance.

of error items ε_{t-j} that are assumed as following Gaussian distribution with zero mean and constant variance. The ϕ_i and θ_j are coefficients that will be calculated with the history pool in real time manner, and we can use *estimate* function in MATLAB to achieve this process.

For the selection of parameters, we generate the training sequence by $\mathcal{B}(100, 20)$. Note that this setting has no important effect on the parameters, and we also employ different groups of Beta distribution intensity in the following simulation part. We firstly employ Augmented Dickey-Fuller [42] for the stationary test to determine d-order difference. The p-value of raw data, first-order difference data, and second-order difference data are equal to 0.50, 0.19, and 0.001, respectively. Considering the p-value typically needs to be less than 0.05, we set $d = 2$. For determining the values of p and q, we employ two common indicators, i.e., the Akaike information criterion (AIC) [43] and Durbin-Watson (DW) test [44], to evaluate the performances of prediction model. Akaike information criterion (AIC) can measure the goodness of data fitting, and the model with smallest score is viewed as the most accurate. If the valid information of training sequence has be completely utilized to train and fit the model, the results of DW test will approach to 2, which suggests there is no autocorrelation of residuals in regression analysis. The other function of DW test is to verify whether the Gaussian assumption about random errors is reasonable.

As shown in Fig. 4, AR(0) & MA(3) model performs well in both the two metrics, especially its DW score is equal to 2.008721 showing the high efficiency of this prediction model. Thus we finally choose ARIMA model (0,2,3). Moreover, since the uniform traffic is assumed as stable access intensity in this paper, we won't derive additional prediction for this traffic pattern.

V. Adaptive Resource Allocation

As diverse URLLC applications is emerging in the 5G and future communication systems, some services with different QoS requirements will inevitably access and request for the same frequency range simultaneously. This calls for more reasonable resource allocation scheme, which would not only fully consider the current users' QoS requirements but also have prediction capability towards the future pressure of resources. We have derived several network-load estimation and prediction schemes above, and in this section we will discuss the allocation strategy driven by different QoS requirements correspondingly.

A. Access Failure Probability

We derive the access failure probability with W RBs, N users, K repetitions, and T slots in an access cycle, which will provide prospective bases for resource allocation.

1) Adjacent-Type K-Repetition: The number of active users in the t-th slot is represented as n_t, and the number of users beginning their first access in the rth slot is denoted as ϕ_r. The vector $\boldsymbol{\phi}$ denotes $(\phi_1, \phi_2, \ldots, \phi_{T-K+1})^T$. Firstly we target at a randomly chosen user, and at the start of access cycle, it will randomly begin its first attempt at the first $T - K + 1$ slots. The access failure probability of this target user who begin at the rth slot($1 \leqslant r \leqslant T - K + 1$) can be formulated by:

$$Pr\{\text{tar}\} = \prod_{t=r}^{r+K-1} 1 - \frac{\binom{W}{1}(W-1)^{n_t-1}}{W^{n_t}}. \tag{16}$$

We have derived the equation (3) above, which depicts the relationship between n_t and ϕ_r, and here we refer it as $f(\boldsymbol{\phi})$. For every determined $\boldsymbol{\phi}$, we can use $f(\boldsymbol{\phi})$ to derive the exact number n_t of users in each slot.

In the following equation, we formulate the total access failure probability, in which $Pr\{\boldsymbol{\phi} \,|\, N\}$ and $Pr\{\text{tar}\}$ have been derived in (7) and (16) respectively:

$$Pr\{\text{total}\} = \sum_{\boldsymbol{\phi}} \left\{ Pr\{\boldsymbol{\phi} \,|\, N\} \left\{ \sum_r \frac{\phi_r}{N} \, Pr\{\text{tar}\} \right\} \right\}. \tag{17}$$

2) Arbitrary-Type K-Repetition: The expected number e of users in every slot is equal to $\frac{NK}{T}$, and the total access failure probability can be derived by:

$$Pr\{\text{total}\} = \left\{1 - \frac{\binom{W}{1}(W-1)^{e-1}}{W^e}\right\}^K. \quad (18)$$

B. Allocation Strategy

The users usually have flexible QoS requirements for dealing with various resource pressures, and in this paper we use *ideal* and *minimum* to denote the upper and lower boundaries of QoS, respectively. Considering the typical URLLC services require the reliability to achieve 99.999% within 1 ms latency, we still employ the reliability within 1 ms as the quantitative indicator of QoS. On the other hand, 99.999% reliability within 1 ms also means 10^{-5} access failure probability within 1 ms. The priority levels are used to differentiate between QoS flows from different users [45], and we suppose the BS prioritize QoS flows based on their reliability requirements, i.e., the higher reliability requirement will imply higher QoS priority level in resource allocation.

In Section V-A, we have derived the analytical failure access probabilities in (17) and (18), thus with the given requirements of reliability, it can quickly figure out the required number of RBs. The common procedures of our designed allocation strategy are depicted in Fig. 5(a). While several events following grant-free access manner request for resources simultaneously, the BS will calculate the required resources firstly, then organize virtual negotiation with the consideration of QoS priorities. Finally, the allocated RBs for each event will be determined. More detailed descriptions can be found in Fig. 5(b), in which we consider two events as an example.

As shown in Fig. 5 (b), the one named event A (e.g., following bursty traffic pattern) requires the reliability within 1 ms to meet Q_1, and the other named event B (e.g., following uniform traffic pattern) requires the reliability within 1 ms to meet $Q_2^{\min}$ at least, and its ideal reliability is equal to Q_2^{ide}. The number of users is $\widehat{N}_1$ and $\widehat{N}_2$ respectively, which can be calculated by the prediction module. Without loss of generality, $Q_1 > Q_2^{\min}$. Thus the priority of event A is higher than event B. The whole number of available RBs is denoted as W_{all}. If available RBs are sufficient, we can allocate RBs independently according to each QoS. However, we also need to discuss resource negotiation considering the conditions when available RBs are inadequate or the system is under high-loads. Moreover, in order to methodically deal with the sudden access, we also allocate a small number of RBs to event A even though its users are inactive [27].

The aim of resource allocation is to achieve two services' QoS as high as possible, and once the number of available RBs cannot support them simultaneously, the negotiation part would spontaneously sacrifice the uniform service's QoS to satisfy bursty service's QoS. Note that the least permitted QoS of event B should be larger than $Q_2^{\min}$. Thus the maximum number of RBs that can be negotiated is equal to $W_{Q_2^{\text{ide}}} - W_{Q_2^{\min}}$, and we denote it as $W_{\text{neg}}^{\max}$. For event A, we firstly calculate the required number W_{req} of RBs to meet Q_1 according to (17) or (18). The realistic condition is divided according to the following rule:

- Condition 1: $W_{\text{req}} \leqslant W_{\text{all}} - W_{Q_2^{\text{ide}}}$;
- Condition 2: $W_{\text{all}} - W_{Q_2^{\text{ide}}} < W_{\text{req}} \leqslant W_{\text{all}} - W_{Q_2^{\text{ide}}} + W_{\text{neg}}^{\max}$;
- Condition 3: $W_{\text{all}} - W_{Q_2^{\text{ide}}} + W_{\text{neg}}^{\max} < W_{\text{req}}$.

TABLE III
ADVISED ALLOCATION STRATEGY AND EXPECTED QOS

Conditions	W_1	W_2	$\widehat{Q}_1$	$\widehat{Q}_2$
Cond.1	W_{req}	$W_{Q_2^{\text{ide}}}$	Q_1	Q_2^{ide}
Cond.2	W_{req}	$W_{\text{all}} - W_{\text{req}}$	Q_1	$\geqslant Q_2^{\min}$
Cond.3	$W_{\text{all}} - W_{Q_2^{\text{ide}}} + W_{\text{neg}}^{\max}$	$W_{Q_2^{\text{ide}}} - W_{\text{neg}}^{\max}$	$\leqslant Q_1$	$Q_2^{\min}$

The final number of allocated resource for each service is denoted as W_1 and W_2, and the expected QoSs are denoted as $\widehat{Q}_1$ and $\widehat{Q}_2$. We list these parameters under different conditions in Table III. In particular, the system will come to outage in Condition 3, which cannot support the permitted least QoS of each service simultaneously. One possible solution is to meet $Q_2^{\min}$ preferentially and allocate the remaining RBs to event A.

Moreover, the scaled allocation strategy considering more than two events can be depicted in Fig. 6. For more events existing, we always can find one event that has the highest priority. Namely, all of the events will be divided into two types, i.e., the highest priority event (*No.1 event*) and the whole of *other events*, which can be regarded as "Event A" and "Event B" from the generalized prospective. Similar to Fig. 5(b), "Event B" will be calculated the total number of required RBs according to the ideal and minimum QoS requirement of each event accordingly. Thus we can determine the allocated resources for No.1 event firstly, and repeat the division and allocation operations for the remaining events until the last two.

VI. SIMULATION EVALUATIONS

A. Load Estimation

In this part we will evaluate the accuracy of our proposed estimation schemes which have been introduced in Section III. We also employ two outstanding baselines for comparison.

1) Baselines I: Minimum Square Error for Mean Values of Access States Estimation (MSEM)

For comparative analysis, we briefly describe another approach proposed in [17] (for mMTC), and here we adjust it in order to fit the model of URLLC in this paper. Firstly, we use a parameter θ_r^e to denote the total number of users who access in e consecutive slots $(1 < e \leq K)$ starting from rth slot, i.e., $\theta_r^e = n_r + n_{r+1} + \cdots + n_{r+e-1}$. The total number of users in 1ms can be derived by:

$$\widehat{N}_{\text{tr}} = \frac{1}{K^2}\left(\sum_{e=1}^{K-1}\theta_1^e + \sum_{r=1}^{T-K+1}\theta_r^K + \sum_{r=T-K+2}^{T}\theta_r^{T-r+1}\right). \quad (19)$$

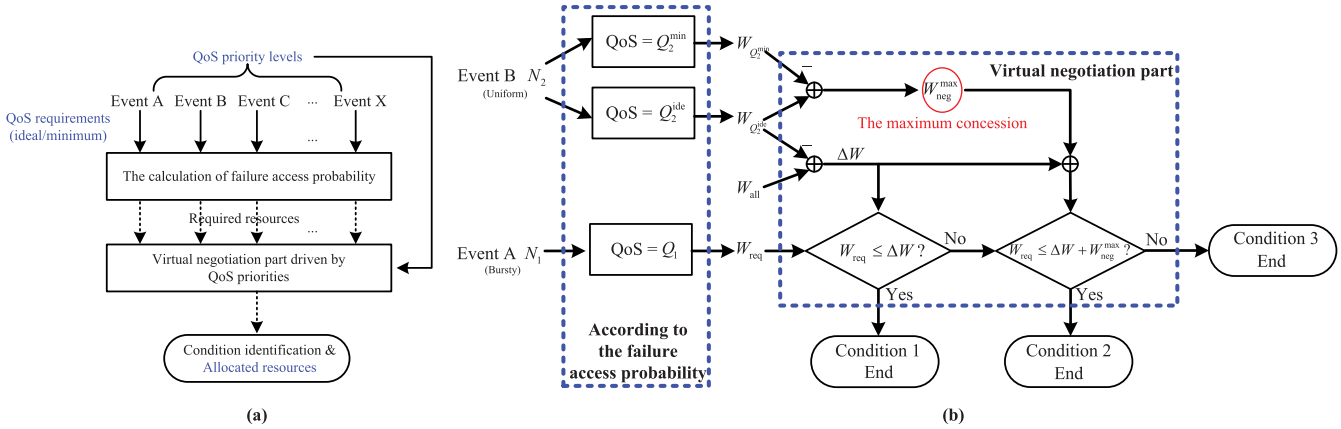


Fig. 5. Flow diagram of resource allocation: (a) the common procedures; (b) taking two events for example.

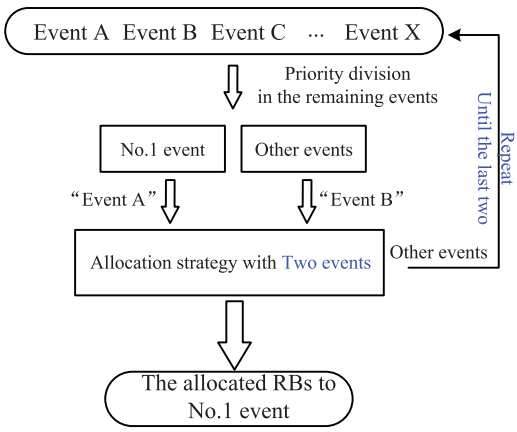


Fig. 6. Scaled allocation strategy considering more than two events.

Next, we use (A, B, C) to describe the total access states of RBs in e consecutive slots. The expectations of them, represented as $\overline{A}$, $\overline{B}$, and $\overline{C}$, can be formulated as follows:

$$\begin{aligned} \overline{A} &= N\left(1-\frac{1}{eW}\right)^{N-1}; \\ \overline{B} &= eW - \overline{A} - \overline{C}; \\ \overline{C} &= eW\left(1-\frac{1}{eW}\right)^{N}, \end{aligned} \tag{20}$$

where N denotes the hypothetical number of total users in e slots. Therefore, the BS can reasonably obtain the estimated θ_r^e by minimizing the Euclidian distance between the theoretical means and the observed access states of RBs:

$$\theta_r^e = \arg\min_N \left[(A-\overline{A})^2 + (B-\overline{B})^2 + (C-\overline{C})^2\right]. \tag{21}$$

2) Baselines II: Idle Resources Counting Estimation (ISCE)

In [46] Oh et al. proposed estimation scheme based on idle resources which can be formulated as follows:

$$n_t = \frac{\log\dfrac{C_i}{W}}{\log\dfrac{W-1}{W}}, \tag{22}$$

in which n_t denotes the number of active users in the t-th slot. According to the system model described in this paper, we know that N_{tr} users with K replicas contribute to the sum of n_t, thus we can obtain:

$$\widehat{N}_{\mathrm{tr}} = \frac{1}{K}\sum_{t=1}^{T} n_t. \tag{23}$$

Note that MSEM is suitable to adjacent-occupation K-repetition, and ISCE is suitable to both of two types. In this subsection, we verify the estimation performances with the results from Monte-Carlo simulations. For adjacent-occupation K-repetition, the SS-ML-LS, MS-MLI, and baseline I (MSEM) are employed; for arbitrary-occupation K-repetition, the MS-MLD and baseline II (ISCE) are employed. Furthermore, in order to accelerate our proposed schemes which significantly utilize Markov model, we generate a state table in advance which can support the quick search of transition probabilities. Thus for every hypothetical N in ML, we only need to query the transition probability with N-steps in this table, rather than calculate $\mathbf{S}$ and $\mathbf{P}$ repeatedly.

The number range of users is from 8 to 18, and we suppose that these users can be served as sufficient resources considering URLLC applications' high QoS requirements. Firstly, we simulate a group of users' choices with MATLAB and count the number of success, collision, and idle RBs respectively. Note that this step has no correlation to certain traffic patterns, and the resource states are only determined by the number of users and available resources. Then all of the estimation schemes will work based on the input $(\mathbf{A}, \mathbf{B}, \mathbf{C})$.

Fig. 7 depicts the accuracy performances of estimation schemes compared with the true values. The most accurate estimation schemes are MS-MLI and MS-MLD with almost no bias, which both consider the overall resource states of a complete access cycle simultaneously and thus avoid introducing errors again caused by quadratic estimation. However, the huge state space of complete access cycle will significantly increase the time complexity searching for the optimal solution, and SS-ML-LS can well offset this problem which estimates in every slots separately. The estimation accuracy of SS-ML-LS is second only to MS-MLI and MS-MLD. The baseline MSEM always has large error, while the baseline ISCE cannot perform stably. In conclusion, if accuracy is a more important factor for estimation, we advice to adopt MS-MLI and MS-MLD;

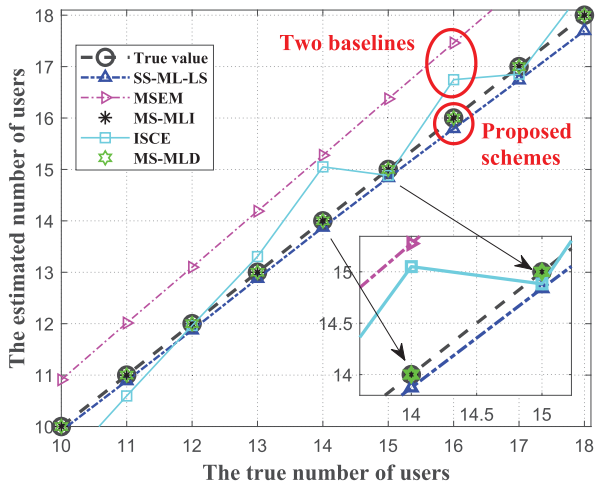


Fig. 7. Estimation performances derived from our proposed schemes (SS-ML-LS, MS-MLI, and MS-MLD) and baselines (MSEM and ISCE).

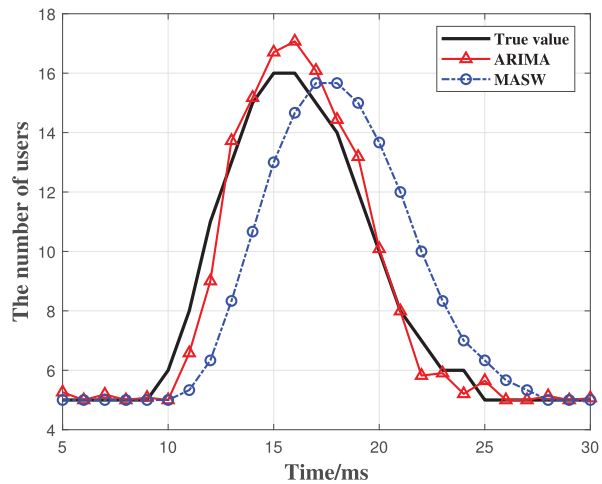


Fig. 8. Prediction performances derived from ARIMA and baseline (MASW).

if operation time is more important, we advice to adopt SS-ML-LS. In the following simulations, we employ the latter.

B. Load Prediction

In this part we will evaluate the prediction model. The parameters of ARIMA model have been discussed in Section IV. We also employ one classic baseline for comparison.

1) Baseline: Moving average with sliding window (MASW)

The t-th expected value is equal to the average of past w observations, which can be formulated as:

$$\widetilde{n}_t = \sum_{i=t-w}^{t-1} n_i, \tag{24}$$

in which n_i $(t-w \leqslant i \leqslant t-1)$ denotes the ith observation selected by the sliding window whose length is equal to w.

In this section we employ ARIMA model (0,2,3) and MASW to achieve single-step prediction. Due to the BS typically has no knowledge on the accurate number of users, the estimated value will be added into the history pool as the realistic observation. As shown in Fig. 8, the uniform event lasts for the whole simulation period, and a bursty event happens over 10 $\sim$ 25 ms which is generated by $\mathcal{B}(80, 15)$. The average prediction error by ARIMA is equal to 6.8%, and that by MASW is equal to 21.9%. From the perspective of global fluctuation, ARIMA can sense and response to the rise and fall tendency of observations in time, while there are always lags between the predictions derived by MASW and realities.

C. Access Failure Probability

In this part we will compare the analytical access failure probability with the simulated results. The analytical formulations have been derived in Section V-A.

In Fig. 9, we compare the analytical results of access failure probability with simulation results and also discuss performances of two types of K-repetition schemes with different K values. The number of active users is set as 10, and the range number of available RBs is from 7 to 33.

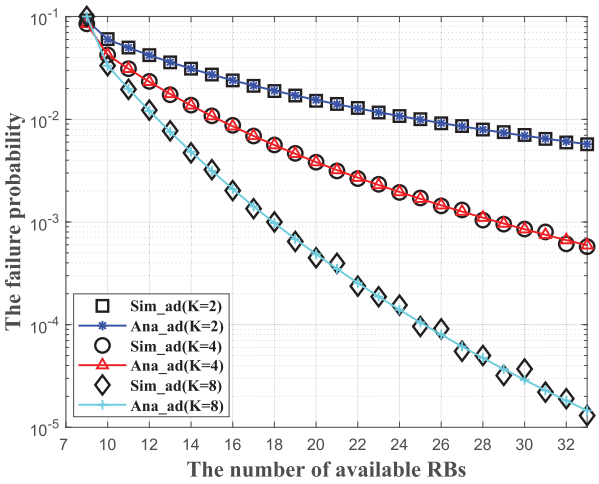
Firstly for adjacent-occupation K-repetition calculated by Eq. (17), analytical results (i.e., Ana_ad) are very close to simulation results (i.e., Sim_ad) with errors ranging in 0.52% ~ 0.74%, as shown in Fig. 9(a). For arbitrary-occupation K-repetition calculated by Eq. (18), as repetition times are closer to the total number of slots in an cycle, the analytical results (i.e., Ana_ar) will be more accurate, as shown in Fig. 9(b).

Moreover, with the same K and repetition times, arbitrary-occupation K-repetition can achieve better access performances compared with adjacent-occupation K-repetition. This is because arbitrary-type permits users to access with higher degrees of freedom and thus reduces collisions. When $K = 8$, i.e., a user will utilize all the available slots of an access cycle, the arbitrary-occupation K-repetition is equivalent to adjacent-occupation.

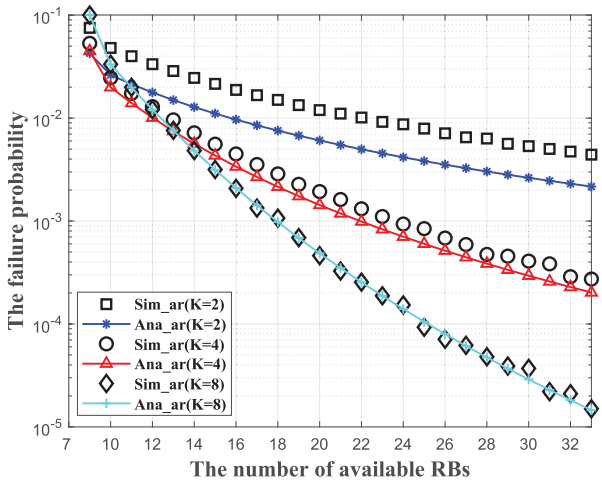
In the K-repetition access scheme, on the one hand, the replicas can enhance success probabilities, while on the other hand, excessive repetitions will also lead to frequent collisions and decline success probabilities instead. Considering the resources allocated to URLLC applications are sufficient, the users can access with relatively high K values without intense collisions. In Fig. 9 we can notice that with the same number of available RBs, either in arbitrary-type or adjacent-type, the larger the repetition times K is, the better the access performances are.

D. Negotiation in Adaptive Resource Allocation

In this part, we illustrate the negotiation part of allocation scheme via Fig. 10, which has been described in Section V-B. Here we consider two applications with different QoS requirements, i.e., event A and event B. Without loss of generality, we assume the event B has higher priority than event A. The W_A and W_B denote the resources allocated to event A and event B originally. If W_B cannot satisfy event B's QoS, the negotiation part will be switched on. For brief description, the parameter δ is used to control negotiation part. In particular, the negotiated number of resources allocated to event A and event B is $W_A * (1-\delta)$ and $W_B + W_A * \delta$. Thus in the actual operation, after we get the predicted number of



(a)



(b)

Fig. 9. Number of available RBs versus the failure probabilities. (a) discusses the adjacent-occupation K-repetition with $K = 2, 4, 8$, and (b) discusses the arbitrary-occupation K-repetition with $K = 2, 4, 8$.

users for the nest cycle, we can calculate the corresponding failure probabilities versus δ, then find the most appropriate δ considering different QoS limitations. For instance, in the Fig. 10, if the minimum reliability requirements of event A and event B are 99% and 99.999% respectively, δ should be 0.55; If the minimum reliability requirements of event A and event B are 99.99% and 99.999% respectively, the total available RBs are insufficient thus it comes to outage.

E. Evaluation of the Overall Performances

We have verified the performances of several important parts above, now we will simulate the overall performances of complete framework proposed in this paper. We set the following methods for comparison: 1) Fairy proportion allocation (FAP): The requirements of various services will be considered fairly rather than according to priorities. The RBs allocated to each service is determined according to the ratio of the number of this service's users to the total number of users. 2) Fixed priority allocation (FIP): The event with lower priority will be allocated with fixed RBs according to its idle QoS ($\mathrm{FIP}\_{\mathrm{ide}}$)

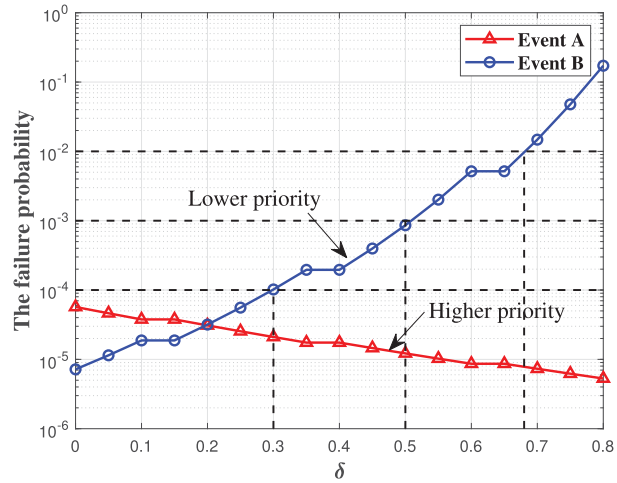
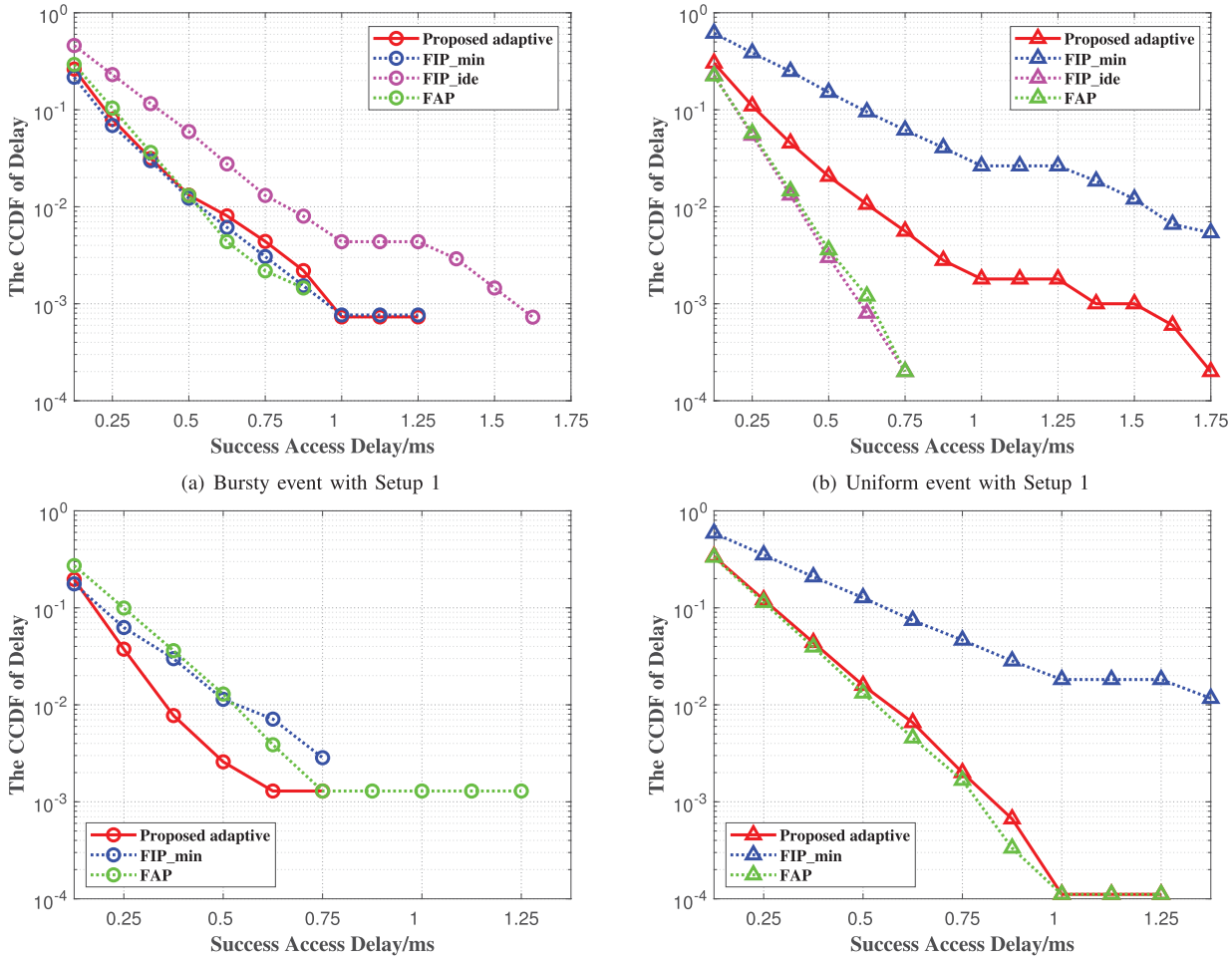


Fig. 10. Parameter δ versus failure probabilities of event A and B. The negotiated number of resources allocated to event A is equal to $W_A * (1-\delta)$, while that to event B is equal to $W_B + W_A * \delta$. Here if reliability requirements of event A and B are 99% and 99.999%, we can set δ as 0.55.

or minimum QoS ($\mathrm{FIP}\_{\mathrm{min}}$), thus the remaining RBs can be reserved for the bursty event with higher priority. Here we suppose the baselines FAP and FIP have the same knowledge of predicted users as our proposed framework.

Specifically, the simulation scenario is composed by two types of services, i.e., uniform and bursty event. We have introduced them in Section II-D and Section V-B. Due to the BS is able to designate RBs to them separately, the access states of each service can be observed independently. Here we set two groups of arrival intensities, as shown in Table IV. The average intensity of beta distribution ranges in $0.5 \sim 6$ users every 1 ms typically [47]. We set the QoS requirement of bursty event and the ideal QoS requirement of uniform event are both equal to 99.999% reliability within 1 ms that is also the typical URLLC standard. Accordingly, the minimum QoS of uniform event is set as 99% reliability. The 1 ms is divided into 8 slots and the length of one slot is equal to 0.125 ms. In Section VI-C, we have verified that when $K = 8$ it can achieve the best performances compared with other K values, thus here we follow this setting. Finally, the total number of available RBs is set as 48, according to [14]. For clear comparison, we employ the empirical complementary cumulative distribution functions (CCDF) of delay as an evaluation indicator. Note that since there are two idle slots between two cycles for the BS to broadcast, the CCDF values in $1 \sim 1.25$ ms remain unchanged.

The simulation results are presented in Fig. 11. In Fig. 11(a) and Fig. 11(b), $\mathrm{FIP}\_{\mathrm{ide}}$ scheme achieves the best QoS for uniform users at cost of serious loss of bursty users' QoS. On the contrary, $\mathrm{FIP}\_{\mathrm{min}}$ scheme enhances bursty users' QoS compared with $\mathrm{FIP}\_{\mathrm{ide}}$, however, its uniform QoS is much lower than other schemes. Our proposed adaptive scheme can achieve 99.999% reliability for bursty users within 1.375 ms, which is similar to $\mathrm{FIP}\_{\mathrm{min}}$ scheme. At the same time, its uniform QoS is also higher than $Q_{\mathrm{u}}^{\mathrm{min}}$. Because when there is no bursty event, the uniform users will have the right to flexibly utilize more RBs than that in $\mathrm{FIP}\_{\mathrm{min}}$ scheme. For FAP scheme, it seems to achieve perfect performances both in



(a) Bursty event with Setup 1

(b) Uniform event with Setup 1

(c) Bursty event with Setup 2

(d) Uniform event with Setup 2

Fig. 11. Success access delay of our proposed adaptive allocation scheme compared with baseline $\mathrm{FIP}_{\_\mathrm{min}}$, $\mathrm{FIP}_{\_\mathrm{ide}}$, and FAP. (a) and (b) are derived with parameter Setup 1, while (c) and (d) are derived with parameter Setup 2. Note that $\mathrm{FIP}_{\_\mathrm{min}}$ means the fixed RBs of uniform users are determined according to their minimum QoS, and $\mathrm{FIP}_{\_\mathrm{ide}}$ means the ideal QoS.

TABLE IV
INTEGRATED SIMULATION PARAMETERS

Parameters	Setup 1	Setup 2
Uniform intensity	10 users	18 users
Bursty intensity	50 users in 10 cycles	25 users in 10 cycles
$Q_{\mathrm{u}}^{\mathrm{min}}$	99% reliability within 1 ms	
$Q_{\mathrm{u}}^{\mathrm{ide}}$	99.999% reliability within 1 ms	
Q_{b}		
W_{all}	48 RBs	
T	8 slots in 1 ms	
K	8 repetitions	

TABLE V
TRAFFIC, IDEAL QOS, AND MINIMUM QOS FROM EVENT A TO EVENT E

	Event A	Event B	Event C	Event D	Event E
Load	10	$\mathcal{B}(30,10)$	$\mathcal{B}(60,15)$	$\mathcal{B}(50,30)$	$\mathcal{B}(40,25)$
Q_{min}	10^{-1}	10^{-3}	10^{-3}	10^{-2}	10^{-2}
Q_{ide}	10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-3}

bursty event and uniform event. However, this is because the ratio $\frac{N_{\mathrm{bur}}}{N_{\mathrm{bur}}+N_{\mathrm{uni}}}$ under parameter group 1 coincidentally makes reasonable divisions of total resources.

In Fig. 11(c) and Fig. 11(d), we won't discuss $\mathrm{FIP}_{\_\mathrm{ide}}$ in that 18 uniform users calling for $Q_{\mathrm{u}}^{\mathrm{ide}}$ will occupy almost all the RBs which causes bursty users' QoS unbearable. Due to the number of uniform users is much larger than bursty users under parameter group 2, most of the available RBs are allocated to the uniform event in FAP scheme. The more unbalanced the ratio is, the more evident this tendency is. Under this condition, our proposed scheme is still better than $\mathrm{FIP}_{\_\mathrm{min}}$ scheme assured by flexible allocation.

F. The Scalability Analysis

When several events with various QoS requirements access to the BS simultaneously, the framework still need to effectively work under these complicated restrictions. In the last paragraph of Section V, we introduced the scaled allocation scheme on the basis of two events. Now we simulate for five events, which traffic patterns and QoS requirements are listed in Table V, to verify the good scalability. Here we use the failure access probability in 1 ms to perform as the indicator

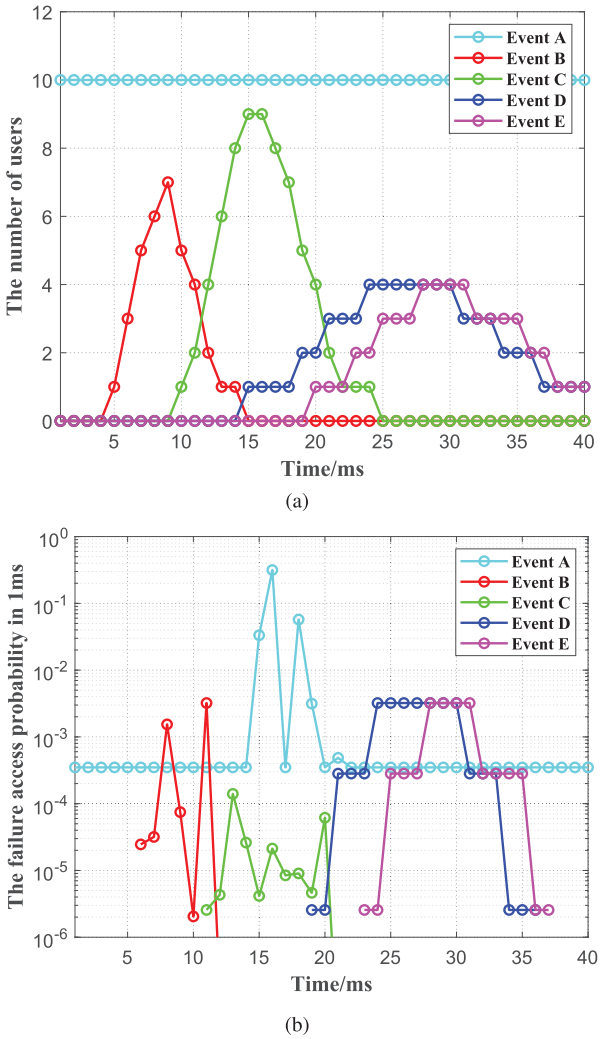


Fig. 12. Scalability performances of our proposed framework considering five different events. (a) shows the loads of each event; (b) shows the failure access probability in 1 ms of each event, i.e., the achieved QoS.

of QoS, and the 10^{-5} failure probability also means 99.999% reliability within 1 ms.

Fig. 12(a) shows the number of users belonging to five events respectively in 40 continuous access cycles. For brief expression, the broadcast slots are not showed in this figure. Especially from 20 ms to 24 ms, there are four events in total accessing to the BS simultaneously with different loads and QoS, which raises a bit challenge to the allocation strategy. Fig. 12(b) depicts the achieved QoS of each event. From 14 ms to 19 ms, due to the strict requirements coming from Event C, the previous resources of Event A are partly allocated to Event C, but the achieved QoS of Event A is still higher than its minimum limit. From 20 ms to 24 ms, the performances of four existing events are all satisfactory. Thus we can believe that even encountering sophisticated conditions, the proposed multi-tier computing framework still has the potential to effectively handle.

VII. Conclusion

In this paper, we considered K-repetition Grant-Free access in URLLC services and proposed a multi-tier-driven computing framework to assure different QoS requirements. In the first tier we designed three network-load estimation schemes, which can estimate the number of current active users based on the resource states (success, collision, and idle). In the second tier we formulated adaptive resource allocation scheme. We employed ARIMA model to predict loads for the next cycle firstly. Moreover, we also derived analytical formulations of access failure probability within 1 ms for K-repetition access. Then the allocation scheme would calculate the reasonable RBs driven by different QoS requirements. Our simulation results showed that MS-MLD and MS-MLI were the most accurate schemes with almost no error, the ARIMA model could achieve accurate and timely predictions with 6.8% relative error, the analytical formulations had relative errors lower than 1% compared with simulation results. Finally, in the integrated simulation, we verified the flexibility, rationality, and scalability of adaptive resource allocation facing the different access intensities and QoS requirements compared with other baselines.

Appendix A
Derivation of n_t

Thus we employ Maximum Likelihood Estimate (ML) to find the most likely N for Markov model, which is represented as $\widehat{N}_{\text{tr}}$:

$$\begin{aligned}\widehat{N}_{\text{tr}} &= \arg\max_{N} P\big((A,B,C)\,|\,N\big)\\ &= \arg\max_{N} \mathbf{P}^{N}_{(0,0,W)\to(A,B,C)}\,,\end{aligned} \tag{25}$$

where $\mathbf{P}$ denotes the transition matrix of Markov model, superscript N denotes the transition steps for hypothesis, and the subscript $(0,0,W) \to (A,B,C)$ shows the initial state and the terminal state in Markov model, respectively.

The state space $\mathbf{S}$ of Markov model containing all the possible states can be derived by the following described rules:

$$\begin{cases} a+b+c=W;\\ 0\leqslant a\leqslant N;\\ 0\leqslant b\leqslant \frac{N}{2};\\ 0\leqslant c\leqslant W,\end{cases} \tag{26}$$

where (a,b,c) represents the possible states, and other values of a, b and c obeying the rules are not allowed. For state (A,B,C) in $\mathbf{S}$, we define the transition probabilities from (A,B,C) to other states as:

$$\begin{cases} P_{(A,B,C)\to(A,B,C)} &= \frac{B}{W};\\ P_{(A,B,C)\to(A+1,B,C-1)} &= \frac{C}{W};\\ P_{(A,B,C)\to(A-1,B+1,C)} &= \frac{A}{W};\\ P_{(A,B,C)\to\text{others}} &= 0.\end{cases} \tag{27}$$

The probabilities that state (A,B,C) transfers into above three states can be described as:

- The new user randomly chooses a RB belonging to the empty RBs with probability $\frac{C}{W}$, thus the number of success RBs increases;

- The new user randomly chooses a RB belonging to the collision RBs with probability $\frac{B}{W}$, and both the number of collision and success RBs do not change;
- The new user randomly chooses a RB belonging to the success RBs exactly with probability $\frac{A}{W}$, causing collision between this new user and the other user who has also chosen this RB, hereafter, the number of collision RBs increases while the number of success RBs decreases.

Substituting (27) into (25), we can calculate the most likely number of users in the t-th slot which is denoted by n_t $(1 \leqslant t \leqslant \mathrm{T})$.

Appendix B
Detailed Expression of Vector $\mathbf{\Omega}$

The vector $\mathbf{\Omega}$ can be derived by:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \\ \mathbf{\Gamma}_3 \end{bmatrix}. \tag{28}$$

The $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, and $\mathbf{\Gamma}_3$ in (28) are equal to:

$$\begin{aligned} \mathbf{\Gamma}_1 &= \begin{bmatrix} \mathbf{1}^{\text{low}}_{(K-1)\times(K-1)} & \mathbf{0}_{(K-1)\times(T-2(K-1))} \end{bmatrix}; \\ \mathbf{\Gamma}_3 &= \begin{bmatrix} \mathbf{0}_{(K-1)\times(T-2(K-1))} & \mathbf{1}^{\text{up}}_{(K-1)\times(K-1)} \end{bmatrix}, \end{aligned} \tag{29}$$

where $\mathbf{1}^{\text{up}}_{(K-1)\times(K-1)}$ denotes a $(K-1)\times(K-1)$ upper triangular matrix in which the main diagonal and all entries above it are equal to 1, $\mathbf{1}^{\text{low}}_{(K-1)\times(K-1)}$ is a $(K-1)\times(K-1)$ lower triangular matrix in which the main diagonal and all entries below the main diagonal are equal to 1, and $\mathbf{0}_{(K-1)\times(T-2(K-1))}$ represents a $(K-1)\times(T-2(K-1))$ matrix in which all entries are equal to 0. $\mathbf{\Gamma}_2$ represents a $(T-2(K-1))\times(T-(K-1))$ which can be described as:

$$\mathbf{\Gamma}_2 = \begin{bmatrix} 1 & \cdots & \underbrace{1}_{K\ \text{times}} & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \underbrace{1}_{K\ \text{times}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & \cdots & \underbrace{1}_{K\ \text{times}} \end{bmatrix}. \tag{30}$$

References

[1] C. She et al., "Deep learning for ultra-reliable and low-latency communications in 6G networks," *IEEE Netw.*, vol. 34, no. 5, pp. 219–225, Sep./Oct. 2020.

[2] M. Afrin, J. Jin, A. Rahman, A. Rahman, J. Wan, and E. Hossain, "Resource allocation and service provisioning in multi-agent cloud robotics: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 842–870, 2nd Quart., 2021.

[3] J. Zhou, Y. Sun, R. Chen, and C. Tellambura, "Rate splitting multiple access for multigroup multicast beamforming in cache-enabled C-RAN," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12758–12770, Dec. 2021.

[4] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document TS 38.913, version 16.0.0., 3GPP, Jul. 2020.

[5] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document TS 36.213, version 16.5.0., 3GPP, Mar. 2021.

[6] W. Cheng, Y. Xiao, S. Zhang, and J. Wang, "Adaptive finite block-length for ultra-low latency in wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4450–4463, Jun. 2022, doi: 10.1109/TWC.2021.3130269.

[7] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Throughput-delay tradeoffs with finite blocklength coding over multiple coherence blocks," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5892–5904, Aug. 2019.

[8] Y. Gu, Q. Cui, Q. Ye, and W. Zhuang, "Game-theoretic optimization for machine-type communications under QoS guarantee," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 790–800, Feb. 2019.

[9] J. Ding, M. Nemati, S. R. Pokhrel, O.-S. Park, J. Choi, and F. Adachi, "Enabling grant-free URLLC: An overview of principle and enhancements by massive MIMO," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 384–400, Jan. 2022.

[10] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Optimization of grant-free NOMA with multiple configured-grants for mURLLC," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1222–1236, Apr. 2022.

[11] T. N. Weerasinghe, V. Casares-Giner, I. A. M. Balapuwaduge, and F. Y. Li, "Priority enabled grant-free access with dynamic slot allocation for heterogeneous mMTC traffic in 5G NR networks," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3192–3206, May 2021.

[12] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, Oulu, Finland, Aug. 2019, pp. 607–612.

[13] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, I. Z. Kovács, and P. Mogensen, "System level analysis of K-repetition for uplink grant-free URLLC in 5G NR," in *Proc. 25th Eur. Wireless Conf. Eur. Wireless*, Aarhus, Denmark, May 2019, pp. 1–5.

[14] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 741–755, Mar. 2021.

[15] H. He, P. Ren, Q. Du, L. Sun, and Y. Wang, "Traffic-aware overload control scheme in 5G ultra-dense M2M networks," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 9, p. e3146, Jan. 2017.

[16] L. Oquendo, J. Bauset, and L. Guijarrro, "Efficient random access channel evaluation and load estimation in LTE-A with massive MTC," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1998–2002, Feb. 2019.

[17] A.-T.-H. Bui, C. T. Nguyen, T. C. Thang, and A. T. Pham, "A comprehensive distributed queue-based random access framework for mMTC in LTE/LTE–A networks with mixed-type traffic," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12107–12120, Dec. 2019.

[18] K. Wang, W. Chen, J. Li, Y. Yang, and L. Hanzo, "Joint task offloading and caching for massive MIMO-aided multi-tier computing networks," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 1820–1833, Mar. 2022.

[19] Y. Yang, T. Zhang, J. Wang, N. Chen, and R. Dirvin, "Fog services and enabling technologies," *IEEE Commun. Mag.*, vol. 57, no. 5, p. 18, May 2019.

[20] S. Aram and B. Jabbari, "Downlink performance of multi-tier wireless networks using punctured Poisson process model," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[21] H. Ma, P. Huang, Z. Zhou, X. Zhang, and X. Chen, "GreenEdge: Joint green energy scheduling and dynamic task offloading in multi-tier edge computing systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4322–4335, Apr. 2022.

[22] Z. Liu, Y. Yang, Y. Chen, K. Li, Z. Li, and X. Luo, "A multi-tier cost model for effective user scheduling in fog computing networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Paris, France, Apr. 2019, pp. 1–6.

[23] J. Xu, K. Ota, and M. Dong, "Saving energy on the edge: In-memory caching for multi-tier heterogeneous networks," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 102–107, May 2018.

[24] E. Baccarelli, M. Scarpiniti, and A. Momenzadeh, "EcoMobiFog–design and dynamic optimization of a 5G mobile-fog-cloud multi-tier ecosystem for the real-time distributed execution of stream applications," *IEEE Access*, vol. 7, pp. 55565–55608, 2019.

[25] A. C. Baktir, C. Tunca, A. Ozgovde, G. Salur, and C. Ersoy, "SDN-based multi-tier computing and communication architecture for pervasive healthcare," *IEEE Access*, vol. 6, pp. 56765–56781, 2018.

[26] W. Lei et al., "Adaptive decision-making framework for federated learning tasks in multi-tier computing," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, New York, NY, USA, May 2022, pp. 1–2.
[27] Z. Zhao, Q. Du, and L. Sun, "Network-load estimation for K-repetition grant-free access enabling adaptive resource allocation towards QoS enhancement," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Helsinki, Finland, Sep. 2021, pp. 1073–1078.
[28] C. D. Young, "USAP: A unifying dynamic distributed multichannel TDMA slot assignment protocol," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, McLean, VA, USA, Oct. 1996, pp. 235–239.
[29] C. D. Young, "USAP multiple access: Dynamic resource allocation for mobile multihop multichannel wireless networking," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Atlantic City, NJ, USA, Nov. 1999, pp. 271–275.
[30] Y. Cao, C. Chen, D. St-Onge, and G. Beltrame, "Distributed TDMA for mobile UWB network localization," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13449–13464, Sep. 2021.
[31] H. Baek and J. Lim, "Design of future UAV-relay tactical data link for reliable UAV control and situational awareness," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 144–150, Oct. 2018.
[32] D. Medina, L. Hu, H. Rosier, and S. Ayaz, "Interference-aware dynamic resource allocation for D2D proximity services with beamforming support," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2014, pp. 1–7.
[33] *Study on New Radio Access Technology Physical Layer Aspects*, document TR 38.802, version 14.0.0., 3GPP, Mar. 2017.
[34] *Service Requirements for Next Generation New Services Markets*, document TS 122.261, version 16.3.0., 3GPP, Mar. 2018.
[35] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42–48, Mar. 2019.
[36] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, "Radio resource allocation and retransmission schemes for URLLC over 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 896–904, Apr. 2019.
[37] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler, and I. Mings, "Cloud-RAN in support of URLLC," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Singapore, Dec. 2017, pp. 1–6.
[38] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Preamble reservation based access for grouped mMTC devices with URLLC requirements," in *Proc. IEEE ICC*, Shanghai, China, May 2019, pp. 1–6.
[39] T. N. Weerasinghe, I. A. M. Balapuwaduge, and F. Y. Li, "Priority-based initial access for URLLC traffic in massive IoT networks: Schemes and performance analysis," *Comput. Netw.*, vol. 178, Sep. 2020, Art. no. 107360.
[40] *Study on RAN Improvements for Machine-type Communications*, document TR 37.868, version 11.0.0., 3GPP, Sep. 2011.
[41] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.
[42] R. Mushtaq. (Aug. 2011). *Augmented Dickey Fuller Test*. [Online]. Available: https://ssrn.com/abstract=1911068
[43] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics*, vol. 81. Dordrecht, The Netherlands: D. Reidel, 1986, p. 26853.
[44] N. E. Savin and K. J. White, "The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors," *Econometrica, J. Econ. Soc.*, vol. 45, no. 8, pp. 1989–1996, Nov. 1977.
[45] S. Ahmadi, *5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*. New York, NY, USA: Academic, 2019.
[46] C.-Y. Oh, D. Hwang, and T.-J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4182–4192, Aug. 2015.
[47] M. Koseoglu, "Lower bounds on the LTE-A average random access delay under massive M2M arrivals," *IEEE Trans. Wireless Commun.*, vol. 64, no. 5, pp. 2104–2115, May 2016.

Zixiao Zhao (Graduate Student Member, IEEE) received the B.S. degree in information engineering from Xi'an Jiaotong University, China, in 2021, where she is currently pursuing the M.S. degree in information and communications engineering.

Her current research interests include ultra-reliable low-latency communications, intrusion detection, and resource allocation.

Qinghe Du (Member, IEEE) received the B.S. degree in information engineering and the M.S. degree in information and communications engineering from Xi'an Jiaotong University, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer engineering from Texas A&M University, College Station, TX, USA, in 2010.

He is currently a Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University. He has authored or coauthored more than 100 technical papers. His research interests include mobile wireless communications and networking with emphasis on security assurance in wireless transmissions, AI-empowered networking technologies, 5G networks and its evolution, cognitive radio networks, industrial internet, blockchain and its applications, and the Internet of Things.

Dr. Du was a recipient of the Best Paper Award from the IEEE GLOBECOM 2007 and IEEE COMCOMAP 2019 and the Best Paper Award of *China Communications* in 2017. He is/was a Technical Program Committee Member of many world-renowned conferences, including IEEE INFOCOM, GLOBECOM, ICC, International Symposium on Personal, Indoor and Mobile Radio Communications, and Vehicular Technology Conference. He was/is the Technical Program Co-Chair of the IEEE International Congress on Cognitive Computing Workshop on Internet of Things (IoT) from 2013 to 2017; the Track Co-Chair of IIKI from 2015 to 2019; and the Publicity Co-Chair of the IEEE International Conference on Communications (ICC) 2015 Workshop on IoT/CPS-Security, IEEE GLOBECOM 2011, International ICST Wireless Internet Conference 2011, and ICST QShine 2010. He was or is an Associate Editor of the IEEE COMMUNICATIONS LETTERS and an Editor of *KSII Transactions on Internet and Information Systems*.

George K. Karagiannidis (Fellow, IEEE) is currently a Professor with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece, and the Head of the Wireless Communications and Information Processing (WCIP) Group. His research interests include wireless communications systems and networks, signal processing, optical wireless communications, wireless power transfer and applications, and communications and signal processing for biomedical engineering.

Prof. Karagiannidis received the two prestigious awards, such as the 2021 IEEE Communications Society Radio Communications (RCC) Committee Technical Recognition Award, for his Outstanding Contributions to *Wireless Systems*, and the 2018 Signal Processing and Communications Electronics (SPCE) Technical Recognition Award of the IEEE ComSoc for his Outstanding Technical Contributions to *Signal Processing for Communications*. He is one of the highly-cited authors across all areas of electrical engineering, recognized by Clarivate Analytics as Web-of-Science Highly-Cited Researcher in the seven consecutive years 2015–2021. He was a past editor of several IEEE journals. From 2012 to 2015, he was the Editor-in-Chief of the IEEE COMMUNICATIONS LETTERS. From September 2018 to June 2022, he has served as the Associate Editor-in-Chief for IEEE OPEN JOURNAL OF COMMUNICATIONS SOCIETY. He is also in the Steering Committee of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKS.