PAPER

RAN Slicing with Inter-Cell Interference Control and Link Adaptation for Reliable Wireless Communications

Yoshinori TANAKA[†a)], *Senior Member and* Takashi DATEKI[†], *Member*

SUMMARY Efficient multiplexing of ultra-reliable and low-latency communications (URLLC) and enhanced mobile broadband (eMBB) traffic, as well as ensuring the various reliability requirements of these traffic types in 5G wireless communications, is becoming increasingly important, particularly for vertical services. Interference management techniques, such as coordinated inter-cell scheduling, can enhance reliability in dense cell deployments. However, tight inter-cell coordination necessitates frequent information exchange between cells, which limits implementation. This paper introduces a novel RAN slicing framework based on centralized frequency-domain interference control per slice and link adaptation optimized for URLLC. The proposed framework does not require tight inter-cell coordination but can fulfill the requirements of both the decoding error probability and the delay violation probability of each packet flow. These controls are based on a power-law estimation of the lower tail distribution of a measured data set with a smaller number of discrete samples. As design guidelines, we derived a theoretical minimum radio resource size of a slice to guarantee the delay violation probability requirement. Simulation results demonstrate that the proposed RAN slicing framework can achieve the reliability targets of the URLLC slice while improving the spectrum efficiency of the eMBB slice in a well-balanced manner compared to other evaluated benchmarks.

key words: *RAN slicing, network slicing, interference control, stochastic network calculus, link adaptation, ultra-reliable low-latency communications, 5G New Radio*

1. Introduction

Utilization of 5G networks in industrial fields holds great potential because 5G networks can meet the quality-of-service (QoS) requirements of various applications [1], which are sometimes specified in the service level agreements (SLAs). Their use in mission-critical applications such as factory automation, motion control, and autonomous vehicle control is particularly promising. For those applications, ultra-reliable and low-latency communications (URLLC) should be supported, where reliability is an important QoS metric in addition to data throughput. An example of a URLLC requirement presented in [2] is a reliability requirement of 99.999% with a user-plane radio latency of 1 ms for a short (32 Bytes) packet transmission. In this paper, reliability is defined as the ratio of successfully delivered packets within the time constraint required by the targeted service to the total number of sent packets. This definition represents a joint probability of meeting related requirements such as maximum latency, packet error rate, and service availability. Network slicing is a key technology for efficiently handling diverse QoS requirements. It allows the network infrastructure to be sliced into logical networks customized to support specific services. An end-to-end (E2E) network slice is composed of network slice subnets configured in different constituent networks, such as core network, transport network, and radio access network (RAN). A slice management system of the entire network determines the requirements of each subnet. Such a hierarchical slicing architecture enables efficient management of E2E slices. A RAN-part slice subnet represents a group of RAN functions and associated resources that support the requirements. However, realizing effective RAN slicing is still challenging due to the time-varying nature of radio propagation environments experienced by users located at different positions. In dense multi-cell deployment scenarios, especially, where inter-cell interference will dominate the communication quality, ensuring the reliability requirements of RAN slices will be difficult without appropriate interference management. This dense deployment within a site's premises is a possible scenario in the industrial arena. Interference management techniques, such as coordinated inter-cell scheduling and coordinated beamforming, can enhance reliability in dense cell deployments. However, tight inter-cell coordination necessitates frequent information exchange between cells, which limits implementation. To satisfy the high reliability requirements of a slice, analysis of the lower tail distributions of the quality-related metrics is necessary because the violation events occur with very low probability. This analysis requires a large amount of data. Our objective is to provide a RAN slicing framework for interference-limited scenarios that efficiently satisfies the stringent reliability requirements of configured RAN slices without necessitating tight inter-cell coordination, utilizing limited measurement data. In terms of reliability-related metrics, we take into account the decoding error probability (DEP) and the delay violation probability (DVP). Consequently, the proposed framework concurrently manages the throughput, maximum latency, and packet error rate for each deployed slice. Throughout the paper, a RAN slice is simply referred to as a slice unless otherwise specified.

1.1 Related Works

Wireless network resource allocation in the context of QoS realization or network slicing in multi-cell environments is a research topic that has attracted considerable attention [3]–[10]. Because inter-cell interference will degrade the DEP

Manuscript received August 4, 2023.
Manuscript revised December 4, 2023.
Manuscript publicized January 30, 2024.

[†]The authors are with Fujitsu Ltd., Kawasaki-shi, 211-8588 Japan.
a) E-mail: yoshi.tanaka@fujitsu.com

DOI: 10.23919/transcom.2023EBP3130

and DVP performances and have a significant impact on the realization of the required reliability in those scenarios, various radio resource allocation schemes to mitigate the impact have been extensively studied. Solutions based on inter-cell coordinated scheduling [3], [11] and centralized resource allocation in cloud RAN (C-RAN) [4], [5] have been well studied, which dynamically allocate the radio resources to individual user equipments (UEs) for spectrally efficient interference control. In [4], [5], QoS-aware resource scheduling algorithms are studied to maximize the sum rate of all users subject to constraints on the acceptable interference power of URLLC and enhanced mobile broadband (eMBB) services. However, these approaches require tightly coordinated scheduling and/or frequent information exchange between cells, which limits implementation. As for solutions without requiring such tight inter-cell coordination, [6] studies a frequency partitioning for coverage enhancement of URLLC communications in interference-limited scenarios, where cell-edge devices use pre-assigned restricted parts of the frequency band so as not to overlap with other neighboring cells. In [7], a slice resource allocation algorithm is proposed to minimize the amount of overlapped radio resources with those allocated to different slices in multi-cell scenarios. In [8], the slice priority and utilization of idle resources are considered to minimize the inter-slice interference, where a pre-determined ratio of radio resource is allocated to each slice, but how to determine the optimum ratio is not studied. In these contributions [6]–[8], the resource allocation algorithms do not take the traffic/slice QoS requirements into account, which could result in inefficient solutions when multiple slices with different QoS requirements are configured. Another approach to cope with the interference is to increase the redundancy instead of proactively reducing the interference. 3rd Generation Partnership Project (3GPP) specified a modulation and coding scheme (MCS) table and a channel quality indicator (CQI) table supporting lower coding rates for packet transmissions with target block error rate (BLER) of 10^{-5} [Table 5.1.3.1-3 of [12]]. The minimum selectable coding rate is 1/4 of that for transmission with target BLER of 10^{-1}. Note that this solution can utilize the increased redundancy more efficiently than a repeated packet transmission, which is another solution taking the same strategy when the maximum number of repetition is 4 or less.

As previously mentioned, it is essential to examine the tail distribution (i.e. rare events) of crucial variables for reliability assurance. Extreme value theory (EVT) [13]–[16] and power-law approximation [17], [18] have been used as tools to model the tail statistics of signal-to-interference-and-noise-ratio (SINR) variables using a limited number of observed data samples. Although only coarsely quantized variables are available at the controller in practical systems, the applicability and the performance of these estimators for such discrete data have not been studied so far.

Even when the minimum SINR of each user is effectively controlled by appropriate inter-cell interference control, rapid changes in the received interference from slot to slot, due to uncoordinated packet scheduling at neighboring cells, degrade the DEP performance. Therefore, to fulfill the stringent DEP requirements of URLLC slices, applying an accurate link adaptation (LA), i.e., the selection of an appropriate MCS for each packet transmission, is essential. An outer loop LA (OLLA) is widely used to compensate for such an MCS selection mismatch at base stations. A popular algorithm of the OLLA employs hybrid automatic repeat request (HARQ) statistics on positive and negative acknowledgments (Ack/Nack) to adjust the MCS values [19]. However, applying this approach to URLLC scenarios is challenging due to slow convergence. For LA algorithms for URLLC, selecting robust MCS values that take into account the worst channel quality reported from UEs is studied in [20]. This scheme necessitates the UEs to reconfigure the CQI calculation rule according to the DEP targets. In [21], the authors propose a modified radio frame structure with additional pilot signals to directly measure the worst-case interference power from each neighboring cells. This approach provides a straightforward solution to capture the worst-case interference; however, it necessitates modifying the standard specifications. In [22], the authors propose an outer-loop CQI correction algorithm for robust MCS selection based on the worst CQI degradation within each observation time window. Although the HARQ statistics are not utilized for fast convergence, the CQI correction without considering the target DEP may result in unnecessarily low MCS selections, consequently degrading the throughput performance.

Meeting the strict E2E latency requirements of delay-sensitive applications/slices in networks is another important issue in the design of network slicing. As tools for analyzing queuing systems with stochastic traffic arrival, departure, and service processes, effective bandwidth [23], effective capacity [24], [25], and stochastic network calculus (SNC) [26] frameworks have been used to obtain bounds on the DVP of the systems [27]. Effective capacity is the dual concept of effective bandwidth. Effective bandwidth (effective capacity) is a large-deviation-type approximation defined as the minimal constant departure rate (arrival rate) needed to serve an arrival process (departure process) under a delay requirement. Mathematical modeling of the departure processes for different typical wireless fading channels have been considered with the effective capacity analysis [24]. However, few attempts have been made on the modelling of interference-limited channels. As remarked in [27], the large-deviation-type approximation will not be appropriate for an analysis dealing with finite packet/queue lengths and very low latencies. In [28], the authors analyze the E2E delay bound and the supportable traffic demands of network slices using the SNC framework with given computation resources in the network. However, no transport properties of the network components (e.g. RAN) are considered here. While arrival processes for a large class of traffic models have been well studied [29], [30], practical modelling of the departure processes for interference-limited channels in multi-cell deployment scenarios has not been studied due to the unpredictable stochastic properties of the inter-cell interference. Several studies have been conducted on modelling the time-varying departure processes of wireless networks for SNC analysis. E2E delay analysis for wireless networks is studied in [31]–[33]. In [31], [32], a two-state (on-off) Markov chain is used to describe the service process of a wireless channel where fading and collisions are present. Such a simple on-off modeling, however, cannot capture the actual departure process of widely used wireless systems applying link adaptation technologies.

1.2 Key Contributions

This paper presents a novel RAN slicing framework designed to efficiently ensure the reliability targets of each slice in interference-limited scenarios. Our key contributions are summarized as follows:

1. We propose a centralized slice-level interference control scheme to fulfill the reliability requirements of all slices, based on the fact that different slices have varying immunities against interference. This approach allocates appropriate frequency resources to each slice, considering lower tail distributions of measured signal-to-interference-ratio (SIR)[†] samples using a power-law estimator applicable to discrete data. We also propose solutions to enhance the estimation accuracy of interference profiles.
2. We introduce a novel SNC framework defined in the radio resource element (RE) domain to analyze wireless systems applying link adaptation, where the traffic arrival process accounts for the tail SIR distributions of the multiplexed traffic flows. Using this framework, we derive performance bounds on the DVP and the required minimum resource size of each slice to fulfill the DVP requirement.
3. We propose a LA algorithm for URLLC to fulfill the target DEP of each scheduled packet flow while suppressing the excessively conservative MCS selections and the resultant throughput degradations. An MCS offset applied for a packet is determined based on lower tail distributions of the CQI variations.

The remainder of this paper is organized as follows. In Sect. 2, we provide a brief description of the system model with RAN slicing. We present the overall configuration of the proposed RAN slicing framework in Sect. 3, and its component technologies, including interference graph generation, resource size determination, slice-aware interference control, and link adaptation for URLLC slices, are described in Sects. 4, 5, 6, and 7, respectively. Section 8 offers the performance evaluation. We conclude the paper in Sect. 9.

2. System Model

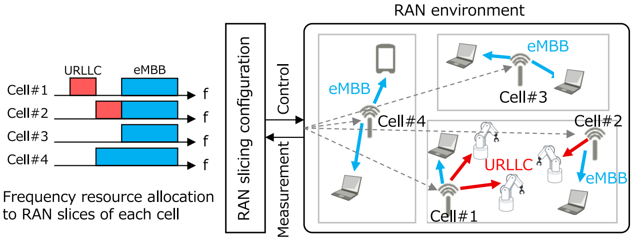


Fig. 1 Dense multi-cell deployment providing eMBB/URLLC services and RAN slicing control.

As shown in Fig. 1, we consider a downlink cellular network system with a set of densely deployed cells C managed by a mobile network operator (MNO), and a set of users U_i served by the cell $i \in C$. Each cell provides multiple services for the users using different slices in a slice set S. In this study, we assume a set S consisting of only two slices (URLLC and eMBB), and each user can associate with only one slice. A RAN slicing model is considered for mitigating the impacts on the slice QoS from inter-cell interference. A central RAN slicing configuration entity (RAN SCE) such as a RAN intelligent controller (RIC) [34] determines radio resources assigned to a slice of a certain cell so as not to overlap with the frequencies used in the neighboring aggressor cells only when the required slice QoS cannot be satisfied, but otherwise the frequency resources can be reused with those of the other cells to improve the spectrum efficiency of the system. As available downlink channel information reported from each UE, infrequent (e.g. every 1 second) beam-level reference signal received power (RSRP) measurements of the serving cell and the neighboring cells, and more frequent (e.g. every 5 ms) CQI reports are considered. Note that the RSRP measurement can be performed for each interfering cell but the CQI can only measure the accumulated interference from all the interfering cells[††].

We assume the RAN SCE has the user's QoS requirements (data rate, acceptable delay, and acceptable packet loss rate) and the prior statistical information of service traffic (e.g. distributions of packet size and packet inter-arrival time). Such information can be obtained as QoS-related parameters allocated for the RAN domain from a network orchestrator to achieve their SLA assurance tasks for industrial users of the system, for example. We assume that a packet loss is caused by the packet decoding error or the delay violation. A delay violation is an event in which the packet delay exceeds a predefined packet delay threshold. The packet delay consists of queuing delay at the packet scheduler and packet retransmission delay. We assume packet retransmission cannot be applied to URLLC due to the stringent requirements of low latency, therefore the packet delay is determined only by the queuing delay. The overall target packet loss probability ϵ^s of slice s can be expressed as follows,

[†] As we focus on interference-limited scenarios, we use SIR instead of SINR to simplify the explanation.

[††] The 5G standard specifies a mechanism to enable the CQI measurement of interference from a particular cell [12]. However, this requires a complex inter-cell coordination of reference signal positions and an increased reference signal overhead. Therefore, we do not consider using this mechanism.

Table 1 Summary of notations.

Symbol	Definition	Symbol	Definition
C	a set of cells	ϕ_i^s	ratio of $\eta_i^{\epsilon_c^s}$ to $\nu_i^{\epsilon_c^s}$
S	a set of RAN slices	$\hat{\xi}_{i,j}^{\epsilon_c^s}$	modified $\xi_{i,j}^{\epsilon_c^s}$
U_i	a set of users served by cell i	N_i^s	number of multiplexed flows for slice s of cell i
ϵ^s	overall target packet loss probability for slice s	$A_{i,k}^s(t_1, t_2)$	cumulative traffic arrivals in (t_1, t_2) for $k \in U_i$
ϵ_c^s	target decoding error probability for slice s	$A_i^s(t_1, t_2)$	cumulative traffic arrivals in (t_1, t_2) of cell i
ϵ_d^s	target delay violation probability for slice s	$a_{i,k}^s(t_1, t_2)$	counting process of $A_{i,k}^s(t_1, t_2)$
$RSRP_{k,n}(i)$	n-th RSRP sample at user $k \in U_i$	d_i^s	RAN-domain steady-state latency for slice s of cell i
$X_{i,j,k}$	RSRP-based pairwise SIR at user $k \in U_i$	τ_s	RAN-domain latency requirement for slice s
$F_{X,ijk}$	distribution function of $X_{i,j,k}$	$D_i^s(t_1, t_2)$	cumulative amount of traffic departure in (t_1, t_2)
$\mu_{i,j,k}$	average of $X_{i,j,k}$	$L_{i,k}^s$	packet size for slice s of user $k \in U_i$
$\xi_{i,j,k}^{\epsilon_c^s}$	ϵ_c^s-quantile pairwise SIR of $F_{X,ijk}$	$\bar{L}$	mean of $L_{i,k}^s$
$\xi_{i,j}^{\epsilon_c^s}$	ϵ_c^s-quantile pairwise SIR of cell i	$B_{i,k}^s$	number of REs for slice s of user $k \in U_i$
G_s	interference graph of slice s	$\rho_{min,i,k}^s$	minimum spectrum efficiency of slice s, user $k \in U_i$
$E_{i,j}^s$	directed edge from vertex j to i of G_s	R_i^s	number of scheduled REs per second
γ_{min}^s	the minimum required SIR for slice s	$\lambda_{i,k}^s$	packet arrival rate for slice s of user $k \in U_i$
γ_{mcs0}^s	required SIR of the slice s when MCS 0 is used	$\bar{b}_{i,k}^s$	mean of exponential distribution model for $B_{i,k}^s$
N_{max}^s	maximum number of retransmission of slice s	R_i^{s*}	required minimum resource size for slice s of cell i
$Z_{i,k}^s$	CQI-based SIR for slice s of user $k \in U_i$	W	required system bandwidth
$Y_{i,k}^s$	RSRP-based cumulative SIR for slice s of user $k \in U_i$	W_{sys}	maximum system bandwidth
$F_{Z,ik}^s$	distribution function of $Z_{i,k}^s$	(f_i^s, w_i^s)	position and size of resource for slice s of cell i
$F_{Y,ik}^s$	distribution function of $Y_{i,k}^s$	$\Delta z_{i,k,\tau}^s(t)$	SIR variation for slice s of user $k \in U_i$
$\eta_{i,k}^{\epsilon_c^s}$	ϵ_c^s-quantile of $F_{Z,ik}^s$	M_k^s	provisional MCS index variable for slice s of user k
$\nu_{i,k}^{\epsilon_c^s}$	ϵ_c^s-quantile of $F_{Y,ik}^s$	$m_k^s(t)$	specific realization of M_k^s
V_i^s	a set of the interferer cells for slice s of cell i	$\tilde{m}_k^s(t)$	final MCS index for slice s of user k
Υ	uniform random variable on $[0, 1]$	$F_{\Delta z_{k,\tau}}^s(z\|M)$	distribution of $\Delta z_{i,k,\tau}^s(t)$ conditioned on M
$\eta_i^{\epsilon_c^s}$	minimum value of $\eta_{i,k}^{\epsilon_c^s}$ within $k \in U_i$	$\Delta_k^s(t)$	required MCS offset for slice s of user k
$\nu_i^{\epsilon_c^s}$	minimum value of $\nu_{i,k}^{\epsilon_c^s}$ within $k \in U_i$		

$$\epsilon^s = 1 - (1 - \epsilon_c^s)(1 - \epsilon_d^s) \approx \epsilon_c^s + \epsilon_d^s, \tag{1}$$

where ϵ_c^s is the target DEP and ϵ_d^s is the target DVP of slice s. When assuming user mobility, another important reliability performance metric is the spatial coverage (availability) of the required QoS. While technical challenges in solving the coverage problem for such scenarios exist, we only consider the cases with no user mobility but design a framework to guarantee the QoS requirements of all the deployed users instead. In each cell, LA is applied using the MCS set and the CQI feedback mechanism defined in 5G New Radio (NR) [12]. We assume there is a constant delay between occasions of a CQI measurement and its subsequent data transmission applying the selected MCS based on the CQI. We apply different LA algorithms for eMBB and URLLC considering the different traffic properties and the different requirements for the DEP. All the notations used throughout the paper are summarized in Table 1.

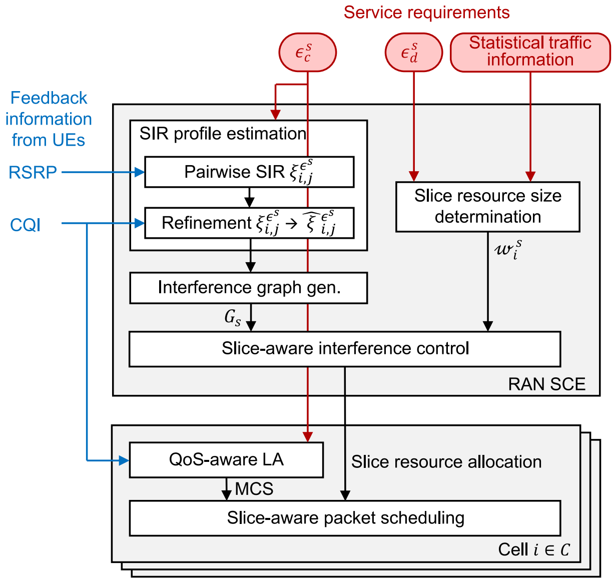

Service requirements
ϵ_c^s
ϵ_d^s
Statistical traffic information
Feedback information from UEs
RSRP
CQI
SIR profile estimation
Pairwise SIR $\xi_{i,j}^{\epsilon_c^s}$
Refinement $\xi_{i,j}^{\epsilon_c^s} \to \hat{\xi}_{i,j}^{\epsilon_c^s}$
Interference graph gen.
G_s
Slice resource size determination
w_i^s
Slice-aware interference control
RAN SCE
QoS-aware LA
MCS
Slice resource allocation
Slice-aware packet scheduling
Cell $i \in C$

Fig. 2 Proposed RAN slicing framework.

3. Proposed RAN Slicing Framework for Reliable Communications

Our proposed RAN slicing framework is shown in Fig. 2. A single RAN SCE cooperatively allocates the frequency resources for each slice of each cell based on per-slice interference graphs (IGs) of the entire network and required resource size of each slice (described in Sect. 6). Each IG is constructed from estimated SIR profiles of all the cells to ensure the resultant interference control can achieve the minimum slice SIR of each cell larger than the required slice SIR. We use a two-step algorithm for this profile estimation to improve the estimation accuracy. The first step provides a coarse SIR profile estimation based on RSRP, and then the second step further refines the outputs more frequently by using the CQI (described in Sect. 4). To determine the required resource size of each slice, an SNC-based estimation is used to achieve the target DVP ϵ_d^s given the statistical information of traffic patterns such as the distribution of packet size and packet inter-arrival time (described in Sect. 5). The traffic properties of each slice and the interference conditions will change on different timescales, and therefore the resource size can be separately updated (e.g. on events of detected traffic pattern changes).

In each cell, a QoS-aware LA is further applied to achieve ϵ_c^s, which controls the probability of selecting insufficient MCS values, due to the rapid interference variations, to be lower than ϵ_c^s (described in Sect. 7).

4. Interference Graph Generation

4.1 RSRP-Based Per-Slice SIR Estimation

For interference control to achieve reliable communications, the network has to accurately estimate an overall SIR profile of the coverage area. In a 5G system, RSRP reported by each user can be used to capture the received interference power at the user from its neighbor cells. A cell periodically broadcasts a set of reference signals transmitted on different beams. Each user can then be configured to report a set of RSRP measurements of these reference signals broadcasted by its neighbor cells. To reduce the amount of measurement work and the amount of feedback signaling, the network can configure each user to report only the highest measured RSRP among measured beams of a limited number of neighbor cells. The following are several problems to consider when using the RSRP for SIR profile estimation:

1. Averaging in time and frequency domain is usually used in the RSRP calculation, and therefore the resultant RSRP will deviate from the instantaneous values.
2. Using the highest measured RSRP among the beam measurements will result in overestimation of the interference for the actual data transmission.
3. The beamforming for the reference signal can use a wider beamwidth than that for user data transmissions to reduce the number of reference signals. Therefore, the estimated SIRs based on such RSRPs will deviate from the actual values of the received user data packets.
4. The actual received interference from a neighbor cell depends on traffic load at the cell. The RSRP-based SIR cannot take into account such traffic-dependent factors.

To improve the estimation accuracy of the SIR distributions based on the available limited channel information, we propose a two-step estimation algorithm as described in the previous section. The first step estimates the pairwise SIR based on the RSRP, which considers only interference from a particular cell. The second step refines these values based on the CQI feedback information to reduce the mismatch between the RSRP-based cumulative SIR and the actual SIR indicated by the CQI. This subsection explains the first step, and details of the second step are described in Sect. 4.3.

We denote the RSRP-based pairwise SIR sequence of user $k \in U_i$ by $X_{i,j,k} = [x_{i,j,k}(0), x_{i,j,k}(1), \ldots]$, which is derived by RSRP as follows,

$$x_{i,j,k}(n) = \frac{RSRP_{k,n}(i)}{RSRP_{k,n}(j)}, \quad k \in U_i, \tag{2}$$

where $RSRP_{k,n}(i)$ and $RSRP_{k,n}(j)$ are the n-th RSRP samples of user $k \in U_i$ measured on reference signals transmitted from cell i and cell j, respectively. This pairwise SIR considers only interference from cell j. The cumulative distribution function (CDF) of $X_{i,j,k}$ is denoted as $F_{X,ijk}$. For URLLC, evaluation of the tail distribution of $F_{X,ijk}$ is important to examine whether the distribution is acceptable to fulfill the required DEP ϵ_c^s of the order of 10^{-5} or less. Using the average SIR values will result in an optimistic evaluation of the interference and it cannot achieve the strict target ϵ_c^s. The system has to estimate the true lower tail distribution of $F_{X,ijk}$, which requires an excessive number of RSRP samples and they are difficult to collect. Considering that the RSRP may not necessarily provide accurate interference information and that the SIR estimation accuracy can be improved by the following second-step refinement, we decided to assume F belongs to a simple parametric model. The small-scale variation of the received reference signal envelope can be modeled by a Rayleigh distribution, therefore we assume F follows an exponential distribution. The CDF can then be given by,

$$F_{X,ijk}(X_{i,j,k}) = 1 - \exp\left(-\frac{X_{i,j,k}}{\mu_{i,j,k}}\right), \tag{3}$$

where $\mu_{i,j,k} = \mathbb{E}[X_{i,j,k}]$ is the average SIR, which can be calculated using fewer RSRP samples. In typical short-range communication scenarios, the channel envelope distribution is more Rician. However, a fixed line-of-site (LOS) component will already be included in the measured RSRP, and therefore the above exponential modelling is still valid.

Given the required DEP ϵ_c^s of slice s, the ϵ_c^s-quantile pairwise SIR $\xi_{i,j,k}^{\epsilon_c^s}$ is defined as, $\xi_{i,j,k}^{\epsilon_c^s} = F_{X,ijk}^{-1}(\epsilon_c^s)$. The ϵ_c^s-quantile pairwise SIR $\xi_{i,j}^{\epsilon_c^s}$ of cell i is represented by the minimum value of $\xi_{i,j,k}^{\epsilon_c^s}$ for $k \in U_i$ as, $\xi_{i,j}^{\epsilon_c^s} = \min_{k \in U_i} \xi_{i,j,k}^{\epsilon_c^s}$. An RSRP-based SIR profile matrix of slice s is defined as, $\{\xi_{i,j}^{\epsilon_c^s} : i, j \in N_{cell}\}$, where $\xi_{i,i}^{\epsilon_c^s}$ is the ϵ_c^s-quantile SNR (i.e. no interference) of the received signal from the serving cell i.

4.2 Graph Representation of Interference Impacts

We then derive a graph representation of the interference impacts in each slice based on the SIR profile. We define an IG G_s of slice s where each vertex represents a cell and the directed edge $E_{i,j}^s$ from vertex j to vertex i represents the existence of a non-negligible level of interference from cell j to cell i at which the slice requirement of ϵ_c^s cannot be guaranteed. We use a binary representation of the label $E_{i,j}^s$ as follows, $E_{i,j}^s = 0$ means there is no edge (i.e., no interference to avoid) and $E_{i,j}^s = 1$ means a directed edge exists. If $E^s_{i,j} = 1$ for any slice $s \in S$, all the resources allocated to cell j should not be overlapped with the resource assigned to slice s of cell i. The directional information of the edges can be utilized for spatial domain interference control where spatial signal processing can realize asymmetric interference control between cells. In this paper, we use a frequency domain interference control, which does not necessarily require such directional information. The following graph formulation can be applied to both types of graphs.

The edge of graph G_s can be derived by solving the following integer optimization problem,

$$\text{minimize} \quad \sum_{i \in C} \sum_{j \in C} E^s_{i,j} \tag{4a}$$

$$\text{s.t.} \quad \left(\sum_{j \in C} \left(1 - E^s_{i,j}\right) \left(\xi^{\epsilon^s_c}_{i,j}\right)^{-1} \right)^{-1} > \gamma^s_{min}, \quad \forall i \in C, \tag{4b}$$

$$E^s_{i,j} = \{0, 1\}, \quad \forall i \in C, \forall j \in C, \tag{4c}$$

$$E^s_{i,i} = 0, \quad \forall i \in C, \tag{4d}$$

where γ^s_{min} is the minimum required SIR to achieve the DEP below ϵ^s_c assuming the lowest MCS (i.e., MCS 0) is used (it is also denoted as γ^s_{mcs0}). The value of γ^s_{min} will further include an additional HARQ gain if it is applicable under the delay constraint. Assuming that Chase combining is used and the interference has no correlation with the desired signals as a simple example case, $\gamma^s_{min} = \gamma^s_{mcs0} / N^s_{max}$ where N^s_{max} is the maximum number of retransmissions for slice s. A solution of the problem (4) minimizes the total number of edges with label equal to 1 while guaranteeing the required SIR, which results in maximizing the spectrum efficiency of the system. The lower the target value of ϵ^s_c, the lower the spectrum efficiency. A URLLC slice, therefore, usually consumes more spectrum resources than the other slices.

4.3 Refinement of IG

In the second step of the algorithm, we introduce an adaptive correction of $\xi^{\epsilon^s_c}_{i,j}$ to mitigate the overestimation of the interference while maintaining the required QoS. Under the current resource allocation of all cells in the network, each user can measure the SIR of the received channel state information reference signal (CSI-RS) transmitted on each slice and report them to the serving cell as CQI feedback. We denote the SIR sequence obtained from the CQI of slice s at user $k \in U_i$ as $Z^s_{i,k} = [z^s_{i,k}(0), z^s_{i,k}(1), \ldots]$, which contains all the interference contributions from neighbor cells unlike the pairwise SIR. An ϵ^s_c-quantile cumulative SIR $\eta^{\epsilon^s_c}_{i,k}$ of slice s at user $k \in U_i$ can be obtained by the empirical distribution function $F^s_{Z,ik}$ of $Z^s_{i,k}$,

$$\eta^{\epsilon^s_c}_{i,k} = {F^s_{Z,ik}}^{-1}\left(\epsilon^s_c\right). \tag{5}$$

The RSRP-based cumulative SIR can be calculated using the pairwise SIR $X_{i,j,k}$ as follows,

$$Y^s_{i,k} = \left(\sum_{j \in V^s_i} \left(X_{i,j,k}\right)^{-1} \right)^{-1} \tag{6}$$

where V^s_i is a set of the cells $\{j | E^s_{i,j} = 0\}$ (which use overlapping resources with those of cell i for the slice s). Equation (6) uses the fact that the contribution of the desired signal power is common for all the SIR $X_{i,j,k}$ for $j \in V^s_i$ at cell i. An ϵ^s_c-quantile cumulative SIR $\nu^{\epsilon^s_c}_{i,k}$ of slice s at user $k \in U_i$ can be obtained by the distribution function $F^s_{Y,ik}$ of $Y^s_{i,k}$,

$$\nu^{\epsilon^s_c}_{i,k} = {F^s_{Y,ik}}^{-1}\left(\epsilon^s_c\right). \tag{7}$$

As explained in Sect. 4.1, the accurate estimation of $F^s_{Y,ik}$ is difficult due to the limited number of available RSRP samples. Instead of using measured RSRP samples in (6), $X_{i,j,k}$ can be sampled from the estimated distributions of the pairwise SIR (3) by using the inverse transformation method [35] as follows,

$$X_{i,j,k} = F^{-1}_{X,ijk}(\Upsilon) = -\lambda_{i,j,k} \ln \Upsilon, \tag{8}$$

where Υ is a uniform random variable on $[0, 1]$.

As our interference control policy is to fulfill all the user requirements of each slice, we define the minimum values of $\eta^{\epsilon^s_c}_{i,k}$ and $\nu^{\epsilon^s_c}_{i,k}$ for $k \in U_i$ which correspond to the worst-quality user, expressed as,

$$\eta^{\epsilon^s_c}_{i} = \min_{k \in U_i} \eta^{\epsilon^s_c}_{i,k}, \tag{9}$$

$$\nu^{\epsilon^s_c}_{i} = \min_{k \in U_i} \nu^{\epsilon^s_c}_{i,k}. \tag{10}$$

When the ratio $\phi^s_i = \eta^{\epsilon^s_c}_i / \nu^{\epsilon^s_c}_i$ is higher than 1, it can be interpreted that there is an overestimation of the interference power in $\nu^{\epsilon^s_c}_i$. Based on the ϕ^s_i, we modify the overestimated pairwise SIR $\xi^{\epsilon^s_c}_{i,j}$ as follows,

$$\hat{\xi}^{\epsilon^s_c}_{i,j} = \phi^s_i \xi^{\epsilon^s_c}_{i,j}, \quad j \in V^s_i. \tag{11}$$

Note that the $\xi^{\epsilon^s_c}_{i,j}$ for $j \notin V^s_i$ remains unchanged because the $Z^s_{i,k}$ measurements have not evaluated the interference from these cells yet. Therefore, applying the above modification to these parameters might result in an excessive relaxation of the evaluated interference from these cells (cell $j \notin V^s_i$). The aggregated interference contribution from all the cells in V^s_i can be corrected by (11). Note that it cannot separately correct the contribution of the individual cell $j \in V^s_i$, but such a fine refinement is not necessary for the relaxation of the overestimated interference. By solving the optimization problem (4) using $\hat{\xi}^{\epsilon^s_c}_{i,j}$ instead of $\xi^{\epsilon^s_c}_{i,j}$, a modified IG can be derived, which can improve the resource utilization efficiency in the following resource allocation.

4.4 Power Law Approximation of Quantized SIR Tail

The proposed IG generation process requires estimating ϵ^s_c-quantile SIR values from the measured data samples. To make a reliable estimation without assuming any parametric models for the data distribution, the data set size of the order $1/\epsilon_c^s$ will be required, which is enormous for practical applications. In [17], power law approximation is used to approximate the lower tail distribution of received power. The tail distribution of logarithmic data $Z = \{\log(X_i)\}_{i=1}^{\mathrm{N}}$, where X_i is a continuous variable, is expressed as $F_Z(z) \approx \alpha e^{z/\kappa}$, where $\kappa = \frac{1}{l}\sum_{i=1}^{l}(z_{(l)} - z_{(i)})$ and $\alpha = \frac{l}{N}e^{-z_{(l)}/\kappa}$. Only the $l = \lceil \beta N \rceil$ smallest order statistics $z_{(1)}, \cdots z_{(l)}$ are used for the estimation of α and κ, given a small constant β. However, this cannot appropriately work when the variable Z is discrete, because the probability of taking the lowest value includes all cases where the value before quantization is less than this value and the selection of the parameter l without considering the quantization boundaries will result in incorrect estimation. To solve these problems, we modified the formulation for Q-level discrete data as follows,

$$\kappa = \frac{1}{\psi(l)}\sum_{i=1}^{\psi(l)}(z_{(\psi(l))} - z_{(i)} - \Delta/2), \tag{12a}$$

$$\alpha = \frac{\psi(l)}{N}e^{-z_{(\psi(l))}/\kappa} \tag{12b}$$

$$\psi(l) = \begin{cases} |\{z : z \leq q_{k_{min}+l-1}\}| & (\psi(1) = 0) \\ |\{z : q_2 \leq z \leq q_{k_{min}+l}\}| & (\psi(1) \neq 0) \end{cases} \tag{12c}$$

where Δ is the quantization step, q_k is the k-th quantized level, k_{min} is the minimum quantization index of the sample set, $l = \lceil \beta Q \rceil$, and $|X|$ denotes the cardinality of a set X. Only the $\psi(l)$ smallest order statistics are used for the estimation. Note that when the probability $\psi(1)/N$ is found to exceed the ϵ_c^s for $\psi(1) \neq 0$, the quantile value can be upper bounded by it and therefore the power law approximation is no longer necessary. To improve the estimation accuracy, it is desirable to set the quantization range to preserve the lower tail distribution well. We applied this estimator for 4-bit quantized (15-level) SIR data for interference control and its differential data (29-level) for link adaptation.

5. Size of Resources Required for Each Slice

In this section, we determine the minimum resource size required for each slice to achieve the target DVP ϵ_d^s. The traffic requirements for slice s of cell i are denoted as $\{A_i^s(t_1, t_2), \tau_s, \epsilon_c^s, \epsilon_d^s\}$, where $A_i^s(t_1, t_2)$ represents the cumulative traffic arrivals in the time period (t_1, t_2), and τ_s denotes the RAN-domain latency requirement. We assume that $A_i^s(t_1, t_2)$ is an aggregate process of the N_i^s statistically independent arrival processes $A_{i,k}^s(t_1, t_2)$ multiplexed in the slice as follows,

$$A_i^s(t_1, t_2) = \sum_{k=1}^{N_i^s} A_{i,k}^s(t_1, t_2). \tag{13}$$

To reduce the loss of frequency utilization efficiency and control complexity, we consider only two slices (URLLC and eMBB) in the proposed RAN slicing scheme. Multiple data flows with similar reliability requirements will be multiplexed in a slice. SNC is a framework to analyze the end-to-end performance of networks with multiple data flows, for deriving performance bounds such as latency and backlog. It can be used to analyze systems with various kinds of arrival processes. We assume that packets are arrived and processed at discrete time $[0, T, 2T, \ldots]$ at a scheduler. Based on the SNC framework, we can upper bound the DVP that the RAN-domain steady-state packet latency d_i^s exceeds τ_s (i.e. latency bound) as follows,

$$\begin{aligned} P\left(d_i^s > \tau_s\right) &< \inf_\theta \left\{\sum_{n=1}^{\infty} \mathbb{E}\left[e^{-\theta D_i^s(0, \tau_s + nT)}\right] \mathbb{E}\left[e^{\theta A_i^s(0, nT)}\right]\right\} \\ &= \inf_\theta \left\{\sum_{n=1}^{\infty} \mathbb{E}\left[e^{-\theta D_i^s(0, \tau_s + nT)}\right] \prod_{k=1}^{N_i^s} \mathbb{E}\left[e^{\theta A_{i,k}^s(0, nT)}\right]\right\}, \end{aligned} \tag{14}$$

where $D_i^s(0, nT)$ represents the cumulative traffic departure from a scheduler in the time period $(0, nT)$, and $\theta > 0$ is a free parameter to be optimized. In (14), $\mathbb{E}[e^{\theta A_i^s(0, nT)}]$ is the moment generating function (MGF) of $A_i^s(0, nT)$ and can be simply expressed as a product of the MGFs of statistically independent $A_{i,k}^s(0, nT)$ for $1 \leq k \leq N_i^s$. This upper bound does not consider the additional traffic for the HARQ packet retransmissions. When there is no correlation between the packet arrival process and the packet sizes, $E[e^{\theta A_{i,k}^s(0, nT)}]$ can be expressed as,

$$\mathbb{E}\left[e^{\theta A_{i,k}^s(0, nT)}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\theta L_{i,k}^s}\right]^{a_{i,k}^s(0, nT)}\right], \tag{15}$$

where $\mathbb{E}\left[e^{\theta L_{i,k}^s}\right]$ is the MGF of the packet size variable $L_{i,k}^s$, and $a_{i,k}^s(0, nT)$ is the counting process of the arrival packets (i.e. the number of arrival packets in a time period $(0, nT)$). For example, if $L_{i,k}^s$ follows an exponential distribution with mean $\bar{L}$, then $\mathbb{E}\left[e^{\theta L_{i,k}^s}\right] = 1/(1 - \theta \bar{L})$.

As briefly reviewed above, the SNC framework can analyze the latency bound with a given arrival and departure process. However, modeling the departure process for wireless communication systems with LA is difficult, because the departure process of a slice is an aggregation of statistically independent scheduling processes of the multiplexed flows, and therefore the cumulative amount of the transmitted bits will be highly dependent on the radio channel condition of each flow. To simplify the application of SNC for RANs with LA, we developed a modified formulation of the latency bound. To determine the minimum radio resource size of each slice, we define the arrival processes and departure processes as the cumulative amount of the radio resources instead of information bits. We use the number of RE as the radio resource amount†. An RE corresponds to one orthogonal frequency-division multiplexing (OFDM) symbol resource of a single subcarrier. The departure process in

†Another unit, such as a resource block, can be used depending on the resource allocation policy of the scheduler.

the RE domain is now a constant rate process unless the allocated bandwidth for the slice changes. To define the arrival processes in the RE domain, we convert the packet size $L_{i,k}^s$ to the number of REs $B_{i,k}^s$ using the minimum spectrum efficiency $\rho_{min,i,k}^s$ of the flow as follows,

$$B_{i,k}^s = \left\lceil \frac{L_{i,k}^s}{\rho_{min,i,k}^s} \right\rceil. \tag{16}$$

The value of $\rho_{min,i,k}^s$ can be determined from the minimum MCS applicable to $\eta_{i,k}^{\epsilon_c^s}$ of (5). If no information about $\rho_{min,i,k}^s$ is available, e.g. at the beginning of the operation, a value corresponding to the lowest MCS (i.e. MCS 0) can be used for the initial value. $B_{i,k}^s$ represents the minimum resource size required for a packet transmission to guarantee the target DEP ϵ_c^s. Note that the distribution property of $B_{i,k}^s$ is the same as that of $L_{i,k}^s$ unless $\rho_{min,i,k}^s$ changes. We can then upper bound $P(d_i^s > \tau_s)$ in the RE domain as follows,

$$\begin{aligned} &P\left(d_i^s > \tau_s\right) \\ &\le \inf_{\theta} \left\{ \sum_{n=1}^{\infty} e^{-\theta(\tau_s+nT)R_i^s} \prod_{k=1}^{N_i^s} \mathbb{E}\left[e^{\mathbb{E}\left[\theta B_{i,k}^s\right] a_{i,k}^s(0,nT)} \right] \right\} \end{aligned} \tag{17}$$

where R_i^s represents the constant number of scheduled REs per second for the slice. For example, suppose we have an aggregated traffic of N_i^s flows, where each flow has a Poisson counting process with a packet arrival rate $\lambda_{i,k}^s$, and the packet resource sizes $B_{i,k}^s$ are exponential with mean $\bar{b}_{i,k}^s$. Based on the formulas that the MGF of the Poisson counting process $a(0, nT)$ is $M_a(\theta) = \exp\left(\lambda nT(e^{\theta} - 1)\right)$ and the MGF of the exponential variables B is $M_B(\theta) = 1/(1 - \theta\bar{b})$, we can upper bound $P(d_i^s > \tau_s)$ as follows,

$$\begin{aligned} P\left(d_i^s > \tau_s\right) &\le \inf_{\theta} \left\{ \sum_{n=1}^{\infty} e^{-\theta(\tau_s+nT)R_i^s} \prod_{k=1}^{N_i^s} e^{\lambda_{i,k}^s nT\left(e^{\theta \bar{b}_{i,k}^s}-1\right)} \right\} \\ &\le \inf_{\theta} \left\{ \frac{e^{-\theta\tau_s R_i^s}}{\theta T\left(R_i^s - \lambda_i^s(\theta)\right)} \right\}, \end{aligned} \tag{18}$$

where,

$$\lambda_i^s(\theta) = \sum_{k=1}^{N_i^s} \frac{\lambda_{i,k}^s \bar{b}_{i,k}^s}{1 - \theta\bar{b}_{i,k}^s}. \tag{19}$$

The required minimum radio resource size w_i^s of the slice to satisfy the DVP less than ϵ_d^s can be obtained by equating (18) to ϵ_d^s as follows,

$$\begin{aligned} w_i^s(\epsilon_d^s) &= \min\left(R_i^s \mid P\left(d_i^s > \tau_s\right) < \epsilon_d^s\right) \\ &= \inf_{\theta}\left(-\frac{1}{\theta R_i^s} \ln\left(\epsilon_d^s \theta T\left(R_i^s - \lambda_i^s(\theta)\right)\right)\right). \end{aligned} \tag{20}$$

This theoretical bound (20) provides useful guidelines to determine the actual slice resource sizes. For practical operations, the following factors should be further considered; 1) unexpected traffic variations and channel variations in every control period, 2) the resource control latency, and 3) the measurement inaccuracy. Further study is needed to develop the solution that takes these factors into account.

6. Slice-Aware Interference Control

We allocate radio resources to slices in units of resource block (RB), which is defined as 12 consecutive subcarriers in the frequency domain and 1 slot length in the time domain. This is the minimum scheduling unit for transmitting user traffic [36]. Based on the derived interference graph G_s and the slice resource size set $\{w_i^s\}$, a frequency resource of the size w_i^s is allocated for slice s of cell i, denoted as (f_i^s, w_i^s), to satisfy all the reliability requirements of the slice, where f_i^s denotes the lowest RB position of the assigned resource. Each cell transmits the user packets of a slice on the allocated radio resources for it. The optimum resource allocation can be obtained by solving the following optimization problem,

$$\text{minimize } W \tag{21a}$$

$$\text{subject to:}$$

$$0 \le f_i^s \le W - w_i^s, \quad i \in C,\ s \in S, \tag{21b}$$

$$f_i^s + w_i^s \le f_j^{s'} + K\left(2 - z_{i,j,s,s'}^l - E_{j,i}^{s'} - H_{i,j}^{s,s'}\right), \quad i,j \in C,\ s,s' \in S, \tag{21c}$$

$$f_j^{s'} + w_j^{s'} \le f_i^s + K\left(2 - z_{i,j,s,s'}^h - E_{j,i}^{s'} - H_{i,j}^{s,s'}\right), \quad i,j \in C,\ s,s' \in S, \tag{21d}$$

$$z_{i,j,s,s'}^l + z_{i,j,s,s'}^h = 1, \quad i,j \in C,\ s,s' \in S, \tag{21e}$$

$$z_{i,j,s,s'}^l, z_{i,j,s,s'}^h \in 0, 1, \quad i,j \in C,\ s,s' \in S, \tag{21f}$$

$$W \le W_{sys}, \tag{21g}$$

where f_i^s is a parameter to be optimized, $E_{j,i}^s \in \{0, 1\}$ denotes the directed edge (cell i to cell j) of G_s, W denotes the required system bandwidth, W_{sys} is the maximum system bandwidth, and K is a sufficiently large number. $H_{i,j}^{s,s'}$ is 1 if $i = j$ and $s \neq s'$, otherwise it is 0. The binary variables $z_{i,j,s,s'}^l$ and $z_{i,j,s,s'}^h$ are 1 if the resource (f_i^s, w_i^s) is located lower than or higher than the resource $(f_j^{s'}, w_j^{s'})$ for $i,j \in C$, $s,s' \in S$, respectively, otherwise, these are 0. The problem (21) is an integer programming problem with disjunctive constraints (21c) (21d) and the spectrum efficiency can be maximized by minimization of W. When $W < W_{sys}$, the unallocated resource with the total size of $W_{sys} - W$ can be further allocated to the best-effort type slices with lower QoS requirements to improve their throughput.

The problem (21) belongs to the class of strip packing problems [37], which are NP-hard problems, making it difficult to efficiently find the solution for large networks. Instead of directly solving problem (21), we derived a simple but efficient approximation algorithm based on the fact that a resource allocation is sufficient as long as the conditions specified in G_s are fulfilled. The proposed algorithm

Algorithm 1 Sequential slice resource allocation

```
Input : $G_s, w_i^s \quad (i \in C, s \in S)$
Output : $f_i^s \quad (i \in C, s \in S)$
$i \leftarrow 0, s \leftarrow 0$
while $s < |S|$ do
    while $i < |C|$ do
        $J = \{k \mid E_{k,i}^{s'} = 1 \cup H_{i,k}^{s,s'} = 1,\ \ k \in C,\ \ s' \leq s\}$
        $f_i^s = \min\{f \mid (f, w_i^s) \perp (f_j, w_j^{s'}), j \in J, s' \leq s\}$
        $i \leftarrow i + 1$
    end while
    $s \leftarrow s + 1$
end while
```

is shown in Algorithm 1, where $A \perp B$ means resource A and resource B do not overlap. This algorithm allocates resources sequentially for each cell in ascending order of the slice label s. We can further assume the slice labels s are assigned from 0 to $S - 1$ in ascending order of ϵ_d^s. In this case, the resource allocation starts with the slice that is most vulnerable to interference, and the last allocated slice which is usually a best-effort type slice can easily get the remaining unallocated resources as a continuous block.

7. QoS-Aware LA for URLLC Slice

In this section, we address another critical problem to ensure the required stringent reliability for URLLC slices. In interference-limited scenarios, the received interference will rapidly change on a slot-by-slot basis due to the independent packet scheduling at the neighboring cells and the high beamforming gain[†] applied for those packet transmissions. In such environments, the radio channel condition changes between the time of CQI measurement and the time of subsequent data transmission. Therefore, the selected MCS $m_k(t)$ for a transmission at time t based on a CQI measured at $t' = t - \tau$ is not appropriate, where τ is the CQI reporting delay. Note that τ is usually not constant because the CQI feedback occasions do not usually align with the packet scheduling occasions. The LA algorithm for URLLC requiring stringent DEP ϵ_c^s (e.g., less than 10^{-5}) should be robust to any change in channel conditions. As described in Sect. 1.1, applying OLLA is necessary to reduce the decoding errors due to such MCS mismatch, but the popular OLLA algorithm based on HARQ feedback information is difficult to apply for URLLC due to slow convergence.

We enhanced the approach of [22] for application to our RAN slicing framework and propose a novel LA algorithm which has following new features: 1) QoS-aware control of the MCS offsets, and 2) utilization of the conditional distributions of the SIR variations for it. We define the SIR variation $\Delta z_{i,k,\tau}^s(t)$ of user $k \in U_i$ as follows,

$$\Delta z_{i,k,\tau}^s(t) = z_{i,k}^s(t) - z_{i,k}^s(t - \tau). \tag{22}$$

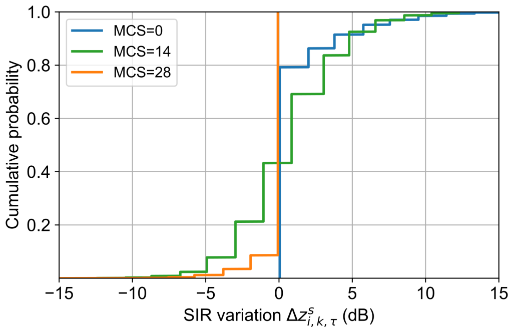


Fig. 3 Example distributions of SIR variations conditioned on the selected MCS of 0, 14, 28 before applying MCS offsets.

The distribution function of $\Delta z_{i,k,\tau}^s(t)$ conditioned on M_k^s is denoted as $F_{\Delta z_{k,\tau}}^s(z|M_k^s)$, where M_k^s is the provisional MCS index variable before applying MCS offsets. An example of $F_{\Delta z_{k,\tau}}^s(z|M_k^s)$ for $M_k^s = \{0, 14, 28\}$ is shown in Fig. 3. We use the MCS table specified by 3GPP [Table 5.1.3.1-1 of [12]] where $M_k^s \in \{0, 1, \ldots, 28\}$. This figure shows the distributions have smaller variances at both ends of the possible MCS index range (i.e., $M_k^s = 0$ and $M_k^s = 28$) than for $M_k^s = 14$. Based on the observation, we derive the required MCS offset $\Delta_k^s(t)$ for slice s depending on the value of M_k^s and calculate the final MCS index $\tilde{m}_k^s(t)$ by applying the offset to the provisional MCS index $m_k^s(t)$ as follows,

$$\Delta_k^s(t) = \left\lfloor - \max(z | F_{\Delta z_{k,\tau}}^s(z | M_k^s = m_k^s(t)) \leq \epsilon_c^s) \right\rfloor, \tag{23a}$$

$$\tilde{m}_k^s(t) = m_k^s(t) - \Delta_k^s(t), \tag{23b}$$

where $m_k^s(t)$ denotes specific realizations of M_k^s at time t. For ϵ_c^s-quantile estimation of $F_{\Delta z_{k,\tau}}^s$ in (23a), we use the power law approximation of (12). This algorithm allows the final MCS index $\tilde{m}_k^s(t)$ to fall almost within the range of the observed SIR distribution. Therefore, it can prevent the use of unnecessarily low MCS indices, thereby enhancing spectrum efficiency and DVP performance. The target DEP ϵ_c^s of the slice can also be satisfied by applying this MCS offset.

For the calculation of (23a), we need to estimate multiple conditional distributions. We present an alternative simpler method to calculate $\Delta_k^s(t)$, which uses only an unconditional distribution $F_{\Delta z_{k,\tau}}^s(z)$. The reason the distribution $F_{\Delta z_{k,\tau}}^s(z|M_k^s)$ differs depending on M_k^s is that the received SIR distribution has a limited range. Based on this, we can constrain the values of $\tilde{m}_k(t)$ whose required SIR to be no less than the minimum observed SIR of the user. We use $\eta_{i,k}^{\epsilon_c^s}$ obtained in (5) as the lowest SIR, and therefore the alternative LA algorithm is given by,

$$\Delta_k^s(t) = \left\lfloor - \max(z | F_{\Delta z_{k,\tau}}^s(z) \leq \epsilon_c^s) \right\rfloor, \tag{24a}$$

$$\tilde{m}_k^s(t) = \max\left(m_k^s(t) - \Delta_k^s(t), f\left(\eta_{i,k}^{\epsilon_c^s}\right)\right), \tag{24b}$$

where the function $f(x)$ returns the required minimum MCS index for a given SIR x. It is expected that the LA given

[†] For 5G and beyond-5G systems, high-gain antennas tend to be used to compensate for the larger path loss and to suppress the interference.

Table 2 Simulation assumptions.

Parameter	Value
Carrier configuration	20 MHz bandwidth at 3.8 GHz; DL FDD
PHY configuration	30 kHz subcarrier spacing; 0.5 ms slot length
Number of RBs	51
Channel model	3GPP TS38.901 InH model [38]
Network layout	12 cells / 120 m × 50 m × 3 m; 20 m inter-site distance
UE distribution	120 UEs uniformly distributed; speed 3 km/h 5 URLLC UEs + 5 eMBB UEs per cell
Traffic model	URLLC: FTP Model 3 with 50 Bytes payload eMBB: Full buffer
Antenna configuration	Cell: 2x2 and 2 polarizations; Single stream UE: 1x1 and 2 polarizations
Transmission power	24 dBm/cell
Control channel	Error free
RSRP	Report only maximum beam/cell every 1 s
MCS table	Table 5.1.3.1-1 of [12]
CQI	Report every 5 ms; 4 ms processing delay
HARQ	URLLC: Disabled eMBB: Max. 5 retransmissions
Slice	URLLC: $\epsilon_c = 10^{-3}$ or 10^{-5}; $\tau_s = 1$ ms; Fixed slice size of 6 RBs/cell eMBB: $\epsilon_c = 10^{-1}$

by (23) and the LA given by (24) have comparable performance but the latter has less complexity.

8. Performance Evaluation

We evaluate the downlink performance of the proposed RAN slicing scheme by conducting a system-level simulation of a multi-cell, multi-user scenario. The default simulation assumptions are summarized in Table 2. We assume a 20 MHz system bandwidth with 51 RBs. We use an indoor deployment model with 120 m × 50 m × 3 m room size, assuming a factory site. Twelve cells are uniformly located within the area with a minimum inter-site distance of 20 m. Each cell has a 2 × 2 downward-facing antenna panel with four fixed beams mounted on the ceiling and transmits a user data packet using the best of the four beams for the user. There are 120 UEs uniformly dropped in the area with stationary positions, while a 3 km/h moving speed is assumed for the fast fading modeling. The 3GPP indoor channel model described in [38] is used. Additional overheads of control signals and reference signals are not included. HARQ is not used for URLLC due to the strict latency constraint of, for example, 1 ms or less, but is used for eMBB with up to 5 retransmissions. For URLLC traffic of each UE, a payload size of 50 Bytes is generated in the downlink following a Poisson distribution with an predefined arrival rate. To simplify the latency evaluation, we assumed a frequency division duplex (FDD) system where there is no buffering delay caused by the alternate uplink/downlink scheduling that must be considered in a time division duplex (TDD) system. We use Table 5.1.3.1-1 of [12] for the MCS table and Table 5.2.2.1-2 of [12] for the CQI table.

8.1 IG-Based Interference Control

We first evaluate the performance of the IG-based interference control. Fig. 4 shows the distributions of the per-UE ϵ_c^s-quantile SIR values of URLLC slices with different interference control schemes. The target DEP is $\epsilon_c^{\mathrm{URLLC}} = 10^{-5}$, and the traffic arrival rate is 50 packets per second (pps). To achieve the target, the ϵ_c^s-quantile SIR must be larger than −4.8 dB, which is the required minimum SIR for the minimum MCS at the ϵ_c^s. This figure indicates that without the interference control, sufficient SIR to achieve the target cannot be attained for three UEs, representing 2.5% of the deployed 120 UEs. By applying the interference control based on the RSRP-based IG, all UEs can achieve sufficiently high SIR to fulfill the target. However, the excessively conservative protection against interference may degrade the spectrum efficiency. Furthermore, the ϵ_c^s-quantile SIR of approximately 90% of UEs becomes larger than 19.8 dB, which is the required minimum SIR for the maximum MCS. Consequently, the spectrum efficiency improvement achieved by LA may be limited because the MCS cannot be increased any further beyond the maximum. Applying the CQI-based refinement to the RSRP-based IG alleviates the overestimated SIRs while satisfying the target ϵ_c^s and reduces the probability that the selected MCS is clipped to its maximum value.

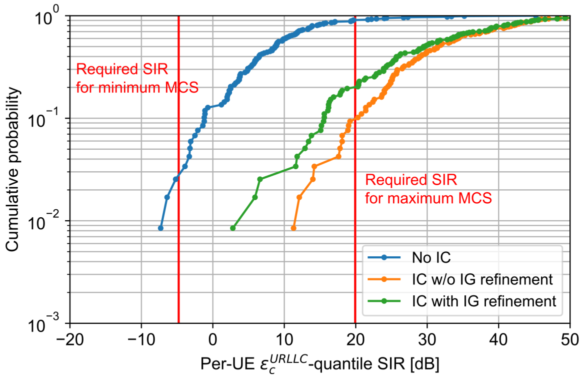
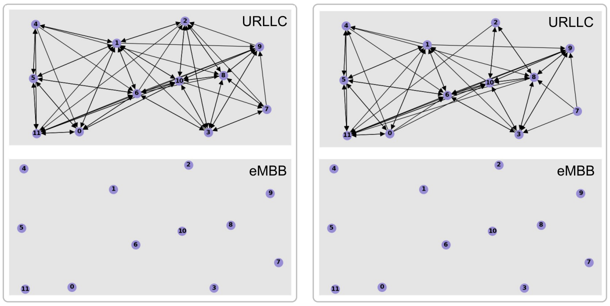


Fig. 4 Distributions of per-UE $\epsilon_c^{\mathrm{URLLC}}$-quantile SIR with different interference control schemes.

The resulting example IGs are shown in Fig. 5, and the performance gains obtained by the IG refinement for the URLLC slice are summarized in Table 3. By applying the CQI-based IG refinement, the number of edges in the URLLC IG is reduced from 73 to 61 (16.4% reduction), which improves the spectrum efficiency by 12.5% without any average throughput degradation. For the eMBB slice, no edge is required in the IG due to the low requirement of ϵ_c^s. The remaining resources after allocation to the URLLC slice can be allocated to the eMBB slice. The CQI-based IG refinement can increase the resource for eMBB from 3 RBs to 9 RBs, which raises the eMBB average cell throughput from 0.67 Mbps to 2 Mbps. Note that the increase rate of the eMBB resources depends on the system bandwidth.

In the 5G system, the available SIR information at each cell consists of coarsely quantized CQI reports fed back from the served UEs. We assume that only 4-bit (15-level) wideband CQIs are available as the CQI reports. For the proposed IG generation described in Sect. 4, it is important to



Without IG refinement		With IG refinement

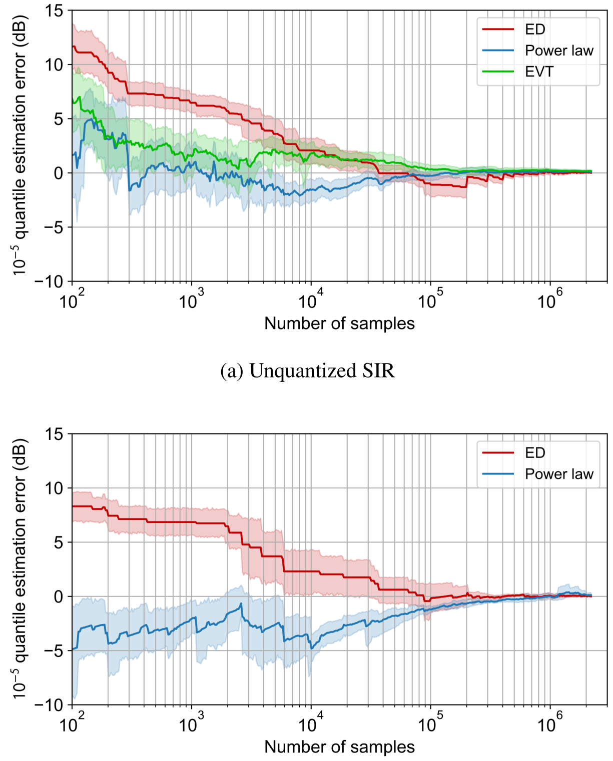
Fig. 5 Examples of estimated interference graphs.

Table 3 Performance gain of IG refinement for URLLC slice.

	w/o refinement	w. refinement
The number of edges in IG	73	61 (-16.4%)
URLLC: Required RBs	48	42 (-12.5%)
URLLC: Ave. user throughput	20 kbps	20 kbps
eMBB: Available resources	3 RBs	9 RBs
eMBB: Ave. throughput	0.67 Mbps	2.0 Mbps

efficiently estimate the ϵ_c^s-quantile SIR from a limited set of coarsely quantized SIR samples. We evaluate three kinds of the estimators: a simple empirical distribution (ED) estimation (i.e., histogram calculation), the power-law approximation, and the EVT-based estimation [39]. For the ED estimation, the minimum value among the sample set is selected as the ϵ_c^s-quantile when the number of samples is less than $1/\epsilon_c^s$. For the power-law approximation, the proposed estimator of (12) is used only for quantized data. In the EVT-based estimation, the CDF of excess samples below a threshold u is modeled as the generalized Pareto distribution (GPD), and the knee/elbow detection algorithm is used to select the threshold u [40]. Note that the EVT-based analysis is theoretically not suitable for a data sequence with dependency, such as the quantized data sequence [39], and therefore it is evaluated only for unquantized data as a reference. The quantile value of the ED estimator is discrete (i.e., one of the quantization levels), whereas those of the power-law estimator and the EVT-based estimator can take continuous values. For each UE, the ED estimation value using 1×10^7 evaluation samples is used as the ground truth. The estimation error is evaluated as the deviation from the ground truth for each UE. We evaluate the performance of 10 randomly selected UEs for $\epsilon_c^s = 10^{-5}$.

As shown in Fig. 6(a), we first evaluate the ϵ_c^s-quantile estimation performances for unquantized SIR as a reference. The solid lines correspond to the mean estimation error, while the shaded curves show the 95% confidence interval, as a function of the number of collected samples. The positive error indicates the estimated values are larger than the ground truth, and the resultant control may not fulfill the requirements. It is observed that the power-law estimator and the EVT-based estimator have faster convergence properties than the ED estimator. The power-law estimation values tend to be below the ground truth because it linearly extrapolates the inaccurate lower tail distribution in the log domain, causing the estimated quantile point to be lower than the actual point. This is a beneficial feature from a risk minimization perspective because the resultant control will be conservative, thus achieving a lower outage risk. In this respect, the power-law estimator is superior to the EVT-based estimator for our purposes. The power-law estimator requires 10^3 samples for convergence to the ground truth within ±3 dB.



(a) Unquantized SIR

(b) Quantized SIR

Fig. 6 ϵ_c^s-quantile SIR estimation ($\epsilon_c^s = 10^{-5}$).

Next, we evaluate the ϵ_c^s-quantile estimation performances for 15-level quantized SIR samples, as shown in Fig. 6(b). The EVT-based estimator was not evaluated, as explained previously. The proposed power-law estimator of (12) exhibits much faster initial convergence properties than the ED estimator and achieves mostly negative estimation errors. The deviations are larger than 3 dB when using fewer than 3×10^4 samples, but these estimation results can be used without increasing the outage risk if the resultant efficiency loss is acceptable because the estimation errors are negative. When the ED estimator is used, about 5×10^4 samples are required to reduce the positive errors below 3 dB. Based on the evaluations, we used the power-law estimation of (12) with 10^3 SIR samples for the initial acquisition in our evaluations. The IG can be updated every 5 s when CQIs are reported every 5 ms.

In general, URLLC encompasses various types of applications and transmissions, such as short-duration applications, transmission scheduled with semi-persistent resources, and sporadic transmissions. For applications with durations shorter than 5 s, the power-law estimation using fewer samples can still be used if the resultant spectrum efficiency loss is acceptable. A UE can be configured for frequent CQI reporting, regardless of the data traffic pattern, to improve communication reliability for semi-persistent scheduling and sporadic transmissions. There is a trade-off between the UE power consumption and acceptable communication reliability/spectrum efficiency. Fine-tuning these trade-offs is a matter of operational policy based on the SLA.

8.2 LA for URLLC

The performance of the proposed periodic control of the MCS offset depends on the number of CQI samples to capture the sufficiently accurate lower tail distribution of the SIR variations. We evaluate the required number of CQI samples to estimate the ϵ_c^s-quantile of $F^s_{\Delta z_{k,\tau}}$ for $\epsilon_c^s = 10^{-5}$. The SIR variation $\Delta z^s_{i,k,\tau}(t)$ is a discrete (29-level) variable (it is the difference between two 15-level quantized SIR values). We compare the simple ED estimator and the proposed power-law estimator of (12) in the same way as the evaluation described in the previous subsection. For each UE, the ED estimation value using 1×10^7 SIR variation samples is used as the ground truth. We evaluate the performance of 10 randomly selected UEs. The ϵ_c^s-quantile estimation performances are shown in Fig. 7. We consider convergence when the confidence limit is within the range of the CQI minimum quantization step of 2 dB from the ground truth. The proposed power-law estimator shows faster initial convergence properties than that of the ED estimator and mostly achieves negative estimation errors. For the ED estimator, about 4×10^4 samples are required to reduce the positive errors below 2 dB, whereas about 10^3 samples can be used for the power-law estimation without increasing the outage risk. Based on the evaluations, we used the proposed power-law estimation of (12) using 10^3 SIR variation samples for the initial estimation in our evaluations. It requires 5 s when CQIs are reported every 5ms. After the initial setting, the latest more samples can be used to improve the throughput.

Figure 8 shows the DEP and the DVP of the system with the proposed LA (24) without the MCS offset restriction (i.e., (24a)+(23b)), the proposed LA (24), and the proposed LA (23), for $\epsilon_c^s = \epsilon_d^s = 10^{-3}$ and 10^{-5}. We evaluated LA (24a)+(23b) to demonstrate the performance limitation when the MCS offset restriction (24b) is not applied in LA (24) as a reference. As a benchmark, the LA scheme proposed in [22], denoted as Min⊿SIR, was also evaluated, where 4×10^4 samples were used for the estimation based on the performance of the ED estimator shown in Fig. 7. The vertical lines indicate the maximum outlier among the evaluated UEs. The RB size allocated to a URLLC slice of each cell is set to 6, which is the minimum size required to fulfill the DVP target of $\epsilon_d^s = 10^{-5}$ when both the IG-based interference control and the proposed LA are used.

For $\epsilon_c^s = 10^{-5}$, all the schemes can meet the DEP requirement, as these schemes take into account the lower tail distribution of $F^s_{\Delta z_{k,\tau}}$ for the MCS selection. Note that the 10^{-5}-quantile value and the minimum value of $F^s_{\Delta z_{k,\tau}}$ are very close, therefore the DEP performances of all schemes are nearly identical. Regarding the DVP performance, Min⊿SIR and LA (24a)+(23b) cannot meet the DVP target because they select MCS indices that are unnecessarily lower than those required for the actual SIR. Although it is a conservative approach to achieving the target DEP, it increases the latency. On the other hand, both the proposed LA (23) and LA (24) can achieve the target DVP by effectively avoiding such unreasonably low MCS selection while still meeting the target DEP, thereby improving both the spectrum efficiency and the DVP performance. By comparing the performances of LA (24a)+(23b) and LA (24), we observed that simply setting a lower limit on selectable MCS indices (24b) is effective in achieving performance comparable to that of LA (23), even though LA (24) is less complex than LA (23).

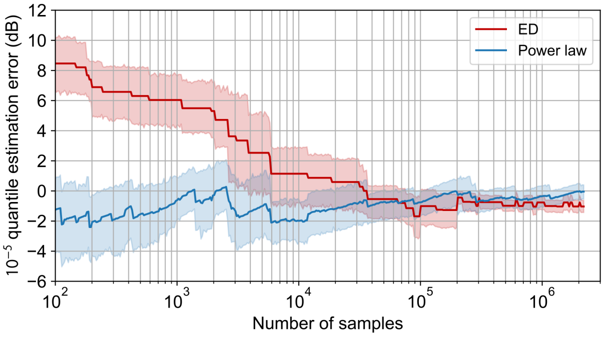


Fig. 7 ϵ_c^s-quantile Δz estimation ($\epsilon_c^s = 10^{-5}$).

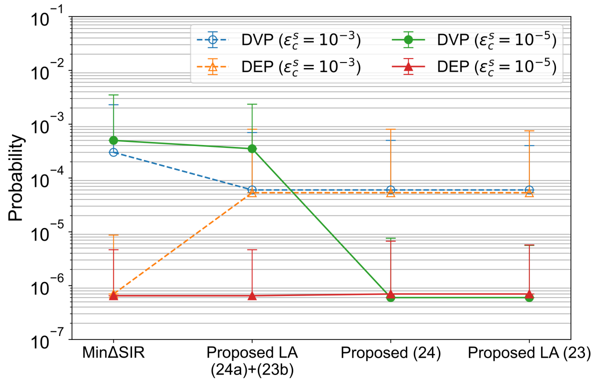


Fig. 8 Reliability performances of different LA algorithms.

In the case of $\epsilon_c^s = 10^{-3}$, all the proposed LA schemes control the DEP to be just below the target. However, the Min⊿SIR results in an unnecessarily small DEP, which consumes extra radio resources and therefore the DVP upper deviation value exceeds 10^{-3}. This is due to Min⊿SIRs inability to properly adjust the amount of MCS offset in line with the relaxation of the DEP target. From the above results, we confirmed that the proposed LA algorithms can achieve better spectrum efficiency by adjusting the DEP for a given target ϵ_c^s.

8.3 Performance of RAN Slicing

We evaluate the achieved reliability performance, user throughput, and spectrum efficiency of our proposed RAN slicing framework (referred to as **Per-slice interference control (PSIC)**), where URLLC slices of all cells have the same fixed resource size of 6 RBs. We also evaluate the following benchmark schemes that do not require tight scheduling coordination between cells, similar to the proposed framework.

- **HigherRedundancy**: MCS and CQI tables supporting lower coding rates for packet transmissions with target DEP of $\epsilon_c^s = 10^{-5}$ specified by 3GPP [Table 5.1.3.1-3 of [12]] are used. No proactive interference control is employed. The entire RBs are shared by all slices.
- **Single slice IC (SSIC)**: The same interference control policy is applied to all slices. This is a conventional slice-agnostic interference control, where a block of frequency resources multiplexing all types of traffic for each cell is allocated to protect traffic with the most stringent QoS requirements of the cell (i.e., URLLC traffic). Partially overlapping of frequency resources between cells is not allowed in the evaluation, similar to the proposed method for ease of comparison. The resource allocation is based on the IG for URLLC traffic of the proposed scheme, and the resultant resource size allocated to each cell is 7.3 RBs (=51/7).

In our evaluation, we assume that the traffic arrival rate of each UE is the same, and therefore the same resource size can be allocated to each slice. Instead of changing the slice resource size, we evaluate the performances for the fixed slice resource size at different traffic arrival rates. Figure 9 and Fig. 10 show the reliability performances (DVP with a 1ms target delay and DEP) for different traffic arrival rates with the target of $\epsilon_c^{\text{URLLC}} = 10^{-5}$ and $\epsilon_c^{\text{URLLC}} = 10^{-3}$, respectively. The vertical lines indicate the maximum outlier among the evaluated UEs. The multiplexed 5 URLLC users in each cell have the same packet arrival rate, and each user data transmission is subject to independent channel variations.

For $\epsilon_c^{\text{URLLC}} = 10^{-5}$, the DEPs are well controlled around the target by combination of the proposed LA and the IG-based interference control. On the other hand, when the packet arrival rate increases, the number of packets waiting to be scheduled will increase, and therefore the DVP is increased. To fulfill the DVP target $\epsilon_d^{\text{URLLC}} = 10^{-5}$, the packet arrival rate should be less than 50 pps for PSIC and SSIC. The DVP performance of HigherRedundancy is much worse than those of PSIC/SSIC because the lower coding rates are selected for the packet transmissions, which require more radio resources. We verified that, for HigherRedundancy, the resource allocation has to be increased from 51 RBs (the current setting) to 140 RBs to accommodate the traffic load of 50 pps for ensuring the DVP below 10^{-5}. Note that the lower the SIR and the lower the target DEP, the more RBs are required to achieve the target DVP due to the resultant smaller MCS selection. Therefore, the required resource size of a slice depends on the interference condition and the target DEP values of UEs multiplexed in the slice.

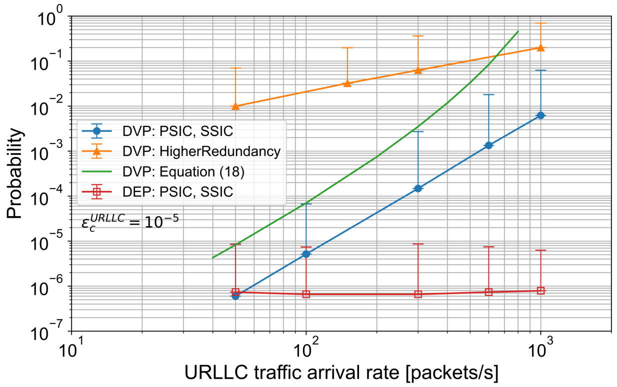


Fig. 9 Reliability performances of URLLC slice applying the proposed RAN slicing framework for $\epsilon_c^{\text{URLLC}} = 10^{-5}$.

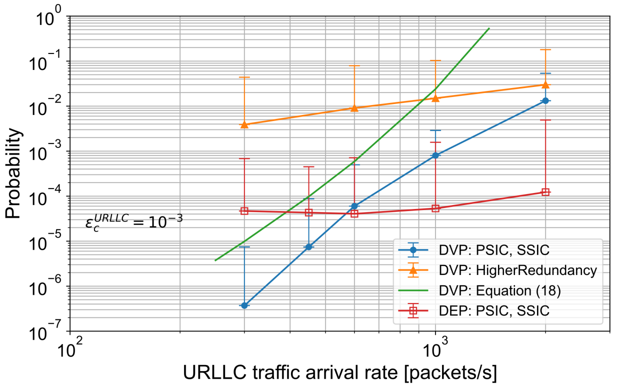


Fig. 10 Reliability performances of URLLC slice applying the proposed RAN slicing framework for $\epsilon_c^{\text{URLLC}} = 10^{-3}$.

For $\epsilon_c^{\text{URLLC}} = 10^{-3}$, the DEPs are well controlled around the target when the packet arrival rate is below 10^3 pps; otherwise, they are slightly increased due to the increased interference. To fulfill the DVP target $\epsilon_d^{\text{URLLC}}$, the packet arrival rate should be less than 700 pps for PSIC and SSIC. The DVP performance of HigherRedundancy is much worse than those of PSIC/SSIC, as for $\epsilon_c^{\text{URLLC}} = 10^{-5}$.

The DVP bounds derived by (18) are in fairly tight agreement with the simulation results. For the calculation of (18), we use the $\eta_{i,k}^{\epsilon_c^s}$ of (5) estimated from simulation to determine $\rho_{min,i,k}^s$. This indicates that (18) can provide useful guidelines for selecting the slice resource sizes. The traffic arrival rate that can be accommodated for $\epsilon_c^{\text{URLLC}} = \epsilon_d^{\text{URLLC}} = 10^{-3}$ is almost 14 times that for $\epsilon_c^{\text{URLLC}} = \epsilon_d^{\text{URLLC}} = 10^{-5}$. By relaxing the reliability requirements, the throughput performance can be greatly improved. Such flexibility arises because the proposed framework can appropriately control the communication reliability according to given target values with the minimum necessary resources. It is reasonable to set ϵ_c^s and ϵ_d^s to the same value for a slice as expected from (1); however, latency requirements τ_s can take various values. In the evaluations of

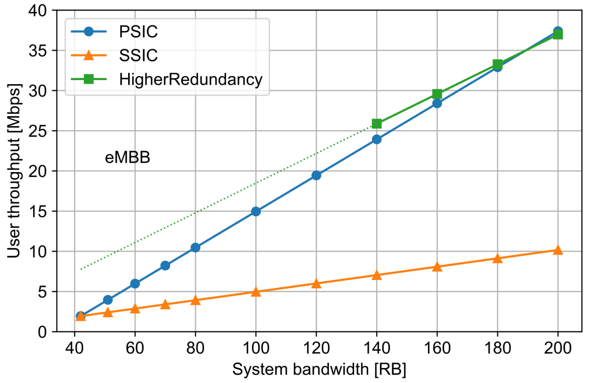


Fig. 11 User throughput of eMBB slice for $\epsilon_c^{\mathrm{URLLC}} = \epsilon_d^{\mathrm{URLLC}} = 10^{-5}$.

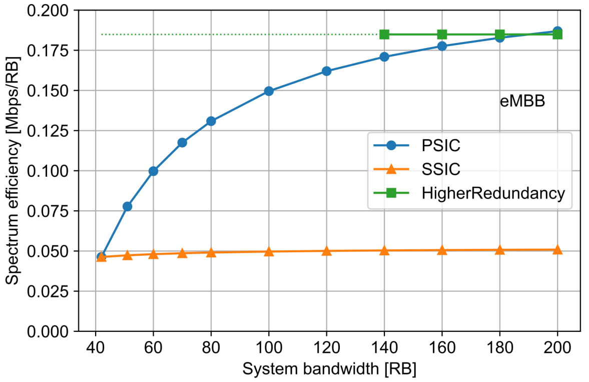


Fig. 12 Spectrum efficiency of eMBB slice for $\epsilon_c^{\mathrm{URLLC}} = \epsilon_d^{\mathrm{URLLC}} = 10^{-5}$.

Fig. 9 and Fig. 10, the same URLLC latency requirement of $\tau_{\mathrm{URLLC}} = 1$ms is used. When longer delays are acceptable for a slice with the same $\epsilon_d^{\mathrm{URLLC}}$, this slice can accommodate more traffic. The proposed framework can provide appropriate control for any τ_{URLLC}, and (18) can properly calculate the latency bound as well.

Next, we evaluate the performance of the eMBB slice when concurrent URLLC transmissions can fulfill the targets of $\epsilon_d^{\mathrm{URLLC}} = \epsilon_c^{\mathrm{URLLC}} = 10^{-5}$ at a traffic arrival rate of 50 pps. For PSIC, in the resources assigned for URLLC slice, the unused resources after allocating the URLLC traffic are used for the eMBB slice to improve spectrum efficiency. The mean user throughput and spectrum efficiency for different system bandwidths are shown in Fig. 11 and Fig. 12, respectively. It should be noted that in these evaluations, the resource size allocated to each cell increases with the system bandwidth for the SSIC and the HigherRedundancy, whereas only the resource size of the eMBB slice grows and the URLLC resource size remains 6 RBs for the PSIC. Although the HigherRedundancy performs slightly better than the PSIC, it cannot fulfill the DVP requirement of the URLLC slice when the system bandwidth is less than 140 RBs (represented by a dotted line). Therefore, the PSIC is superior to the HigherRedundancy considering the higher spectrum efficiency of the URLLC slice. Although the SSIC fulfills the DVP requirement of the URLLC slice for all the evaluated system bandwidths, the performance of the eMBB slice is much lower than that of the PSIC because the available eMBB resources are smaller than those of the PSIC.

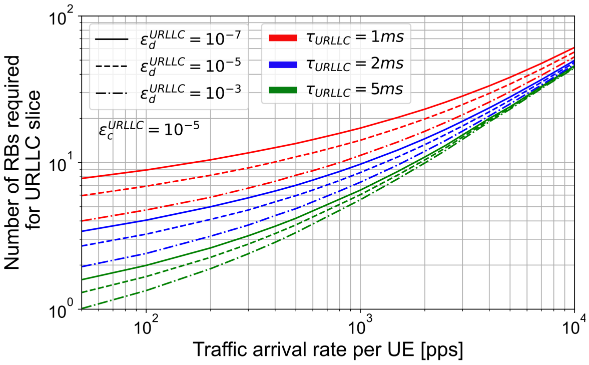


Fig. 13 Number of RBs required for URLLC slice ($\epsilon_c^{\mathrm{URLLC}} = 10^{-5}$).

The derived constraints of the traffic arrival rate (50 pps for $\epsilon_d^{\mathrm{URLLC}} = 10^{-5}$ and 700 pps for $\epsilon_d^{\mathrm{URLLC}} = 10^{-3}$) may be acceptable for certain low-rate delay-sensitive communications, such as mobile robots and motion control use cases in factory automation scenarios as shown in [2]. However, the requirements of the various URLLC use cases are highly diverse, and the traffic constraint for a UE actually depends on several factors, including the number of active UEs connected to its serving cell, interference conditions, system bandwidth, and the antenna configuration deployed at each cell. Although the resource size of each URLLC slice was fixed at 6 RBs in the previous evaluations, the selection of the system bandwidth to accommodate expected traffic demands will be a crucial network design consideration. Figure 13 shows the required resource sizes for a URLLC slice, calculated based on formula (20), to guarantee several target DVP values $\epsilon_d^{\mathrm{URLLC}}$ and the latency requirements τ_{URLLC} across a wider range of traffic arrival rates (from 50 pps (20 kbps) to 10^4 pps (4 Mbps)). The target DEP $\epsilon_c^{\mathrm{URLLC}}$ is fixed to 10^{-5}, meaning the same interference condition is assumed. All other conditions remain the same as those for the previous evaluations. The closer τ_{URLLC} is to the slot length of 0.5 ms, the more resources are required to achieve the target DVP, as more packets need to be scheduled into a slot immediately after their arrival.

The above results show that the proposed RAN slicing framework can realize both the achievement of the reliability targets of the URLLC slice and the improved spectrum efficiency of the eMBB slice in a well-balanced manner compared to the other benchmarks evaluated.

9. Conclusion

In this paper, we proposed a RAN slicing framework for interference-limited scenarios to ensure the reliability targets of each slice based on a novel IG-based per-slice interference control and a novel QoS-aware link adaptation for each packet flow. Our motivation is to develop solutions to efficiently achieve the different reliability targets in mixed traffic scenarios and to avoid tight scheduling/beamforming collaboration between cells. An important component supporting our framework is the quantile estimation of lower tail distributions, and we presented a power-law estimator for discrete data which requires only about 10^3 samples for 10^{-5}-quantile estimation. The proposed SNC-based formulation of the required slice resource size provides useful guidelines to determine the appropriate resource sizes to achieve the target DVP of each slice. Simulation results show that the proposed RAN slicing framework can realize both the achievement of the reliability targets of the URLLC slice and the improved spectrum efficiency of the eMBB slice in a well-balanced manner compared to other evaluated benchmarks. This framework is useful for network service providers to support reliable wireless communications in a variety of industrial applications with different reliability requirements.

The presented solution does not sufficiently demonstrate adaptability to changes in the environment, including interference conditions, traffic volume, UE mobility, and the number of active UEs, required to ensure meeting the slice requirements. Future research will focus on investigating dynamic solutions to effectively handle these changes. Frequent reconfiguration results in increased control overhead and complexity; thus, it is essential to achieve rapid adaptation to such changes with minimized reconfiguration costs.

Acknowledgments

This paper includes a part of results of "The research and development project for 5G Shared Radio Unit (JPJ000254)" commissioned by The Ministry of Internal Affairs and Communications, Japan.

References

[1] 3GPP, "Study on scenarios and requirements for next generation access technologies," Technical Specification TR38.913 v17.0.0, 2022.

[2] 3GPP, "Service requirements for cyber-physical control applications in vertical domains," Technical Specification TS22.104 v18.3.0, 2021.

[3] M.S.J. Solaija, H. Salman, A.B. Kihero, M.I. Sağlam, and H. Arslan, "Generalized coordinated multipoint framework for 5G and beyond," IEEE Access, vol.9, pp.72499–72515, 2021.

[4] M. Setayesh, S. Bahrami, and V.W. Wong, "Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN," GLOBECOM 2020 - 2020 IEEE Global Communications Conference, pp.1–6, 2020.

[5] M.K. Motalleb, V. Shah-Mansouri, S. Parsaeefard, and O.L.A. López, "Resource allocation in an open RAN system using network slicing," IEEE Trans. Netw. Service Manag., vol.20, no.1, pp.471–485, 2023.

[6] N. Brahmi, O.N.C. Yilmaz, K.W. Helmersson, S.A. Ashraf, and J. Torsner, "Deployment strategies for ultra-reliable and low-latency communication in factory automation," 2015 IEEE Globecom Workshops (GC Wkshps), pp.1–6, 2015.

[7] S. D'Oro, L. Bonati, F. Restuccia, and T. Melodia, "Coordinated 5G network slicing: How constructive interference can boost network throughput," IEEE/ACM Trans. Netw., vol.29, no.4, pp.1881–1894, 2021.

[8] H. Li, H. Li, X. Wen, L. Wang, Z. Lu, W. Jing, and Y. Chen, "An interference minimization-based RAN slicing strategy in 5G systems," 2021 17th International Symposium on Wireless Communication Systems (ISWCS), pp.1–6, 2021.

[9] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," IEEE Trans. Wireless Commun., vol.17, no.10, pp.6419–6432, 2018.

[10] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, "Network slicing games: Enabling customization in multi-tenant mobile networks," IEEE/ACM Trans. Netw., vol.27, no.2, pp.662–675, 2019.

[11] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," IEEE Commun. Mag., vol.49, no.2, pp.102–111, 2011.

[12] 3GPP, "NR; physical layer procedures for data," Technical Specification TS38.214 v17.3.0, 2022.

[13] E. Castillo, A.S. Hadi, N. Balakrishnan, and J.M. Sarabia, Extreme Value and Related Models With Applications in Engineering and Science, Wiley, 2004.

[14] Y.H. Al-Badarneh, C.N. Georghiades, and M.S. Alouini, "Asymptotic performance analysis of generalized user selection for interference-limited multiuser secondary networks," IEEE Trans. Cogn. Commun. Netw., vol.5, no.1, pp.82–92, 2019.

[15] A. Subhash, M. Srinivasan, and S. Kalyani, "Asymptotic maximum order statistic for SIR in κ-μ shadowed fading," IEEE Trans. Commun., vol.67, no.9, pp.6512–6526, 2019.

[16] A. Subhash, M. Srinivasan, S. Kalyani, and L. Hanzo, "Transmit power policy and ergodic multicast rate analysis of cognitive radio networks in generalized fading," IEEE Trans. Commun., vol.68, no.6, pp.3311–3325, 2020.

[17] M. Angjelichinoski, K.F. Trillingsgaard, and P. Popovski, "A statistical learning approach to ultra-reliable low latency communication," IEEE Trans. Commun., vol.67, no.7, pp.5153–5166, 2019.

[18] P.C.F. Eggers, M. Angjelichinoski, and P. Popovski, "Wireless channel modeling perspectives for ultra-reliable communications," IEEE Trans. Wireless Commun., vol.18, no.4, pp.2229–2243, 2019.

[19] 3GPP, "NR and NG-RAN overall description; stage-2," Technical Specification TS38.300 v17.2.0, 2022.

[20] G. Pocovi, A.A. Esswie, and K.I. Pedersen, "Channel quality feedback enhancements for accurate URLLC link adaptation in 5G systems," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), pp.1–6, 2020.

[21] D.T. Phan-Huy, P. Chauveau, A. Galindo-Serrano, and M. Deghel, "High data rate ultra reliable and low latency communications in bursty interference," 2018 25th International Conference on Telecommunications (ICT), pp.175–180, 2018.

[22] A. Belogaev, E. Khorov, A. Krasilov, D. Shmelkin, and S. Tang, "Conservative link adaptation for ultra reliable low latency communications," 2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp.1–5, 2019.

[23] C.S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," Proc. INFOCOM'95, vol.3, pp.1001–1009, 1995.

[24] M. Amjad, L. Musavian, and M.H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," IEEE Commun. Surveys Tuts., vol.21, no.4, pp.3007–3038, 2019.

[25] J. Choi, "An effective capacity-based approach to multi-channel low-latency wireless communications," IEEE Trans. Commun., vol.67, no.3, pp.2476–2486, 2019.

[26] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," IEEE Commun. Surveys Tuts., vol.17, no.1, pp.92–105, 2015.

[27] M. Bennis, M. Debbah, and H.V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," Proc. IEEE, vol.106, no.10, pp.1834–1853, 2018.

[28] Q. Xu, J. Wang, and K. Wu, "Learning-based dynamic resource pro-

visioning for network slicing with ensured end-to-end performance bound," IEEE Trans. Netw. Sci. Eng., vol.7, no.1, pp.28–41, 2020.

[29] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," IEEE/ACM Trans. Netw., vol.15, no.6, pp.1442–1453, 2007.

[30] M.A. Beck, S.A. Henningsen, S.B. Birnbach, and J.B. Schmitt, "Towards a statistical network calculus — Dealing with uncertainty in arrivals," IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, pp.2382–2390, 2014.

[31] O. Azuaje and A. Aguiar, "End-to-end delay analysis of a wireless sensor network using stochastic network calculus," 2019 Wireless Days (WD), pp.1–8, 2019.

[32] L. Zhang, X. Chen, X. Xiang, and J. Wan, "A stochastic network calculus approach for the end-to-end delay analysis of LTE networks," 2011 International Conference on Selected Topics in Mobile and Wireless Networking (iCOST), pp.30–35, 2011.

[33] S. Ma, X. Chen, Z. Li, and Y. Chen, "Performance evaluation of URLLC in 5G based on stochastic network calculus," Mobile Netw. Appl., vol.26, no.3, pp.1182–1194, 2021.

[34] O-RAN Alliance, "O-RAN Architecture Description - v07.00," Technical Specification, 2022.

[35] S.M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, 5th ed., Elsevier, 2014.

[36] 3GPP, "NR; physical channels and modulation," Technical Specification TS38.211 v17.3.0, 2022.

[37] B.S. Baker, E.G. Coffman, Jr., and R.L. Rivest, "Orthogonal packings in two dimensions," SIAM J. Comput., vol.9, no.4, pp.846–855, 1980.

[38] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," Technical Specification TR38.901 v16.1.0, 2020.

[39] N. Mehrnia and S. Coleri, "Wireless channel modeling based on extreme value theory for ultra-reliable communications," IEEE Trans. Wireless Commun., vol.21, no.2, pp.1064–1076, 2022.

[40] A. Gomes, J. Kibiłda, and L.A. DaSilva, "Capturing rare network conditions to dimension resources for ultra-reliable communication," IEEE Commun. Lett., vol.26, no.11, pp.2789–2793, 2022.

Yoshinori Tanaka received the B.S. degree in electrical engineering from Yokohama National University, Japan, in 1983, and the Ph.D degree in electrical engineering from Keio University, Japan, in 2008. Since 1983, he has been working in various research and management positions at Fujitsu Laboratories Ltd. From 2003 to 2014, he was a part-time lecturer with Graduate School of Global Information and Telecommunication Studies of Waseda University, Japan. From 2017 to 2018, he was a part-time lecturer of Tokyo Institute of Technology, Japan. He is currently a research expert with Fujitsu Ltd. His current research interests include wireless communications, reliable and low-latency communications and statistical signal processing. He is a senior member of IEICE and IEEE.

Takashi Dateki received the M.S. degree in theoretical physics from Nagoya University, Japan, in 1996. Since 1997, he has been working on research and development of mobile communication systems at Fujitsu Laboratories Ltd. He is currently a senior manager with Fujitsu Ltd. His current research interests include wireless technologies for 5G and 6G mobile communication systems. He is a director of international coordination and publicity of the Institute of Electronics, Information, and Communication Engineers (IEICE) of Japan.