Joint Admission Control and Slice Dimensioning Based on Symbol-Level Resource Allocation in 5G+

Valentin Thomas Haider, Fidan Mehmeti, Wolfgang Kellerer
Chair of Communication Networks, Technical University of Munich, Germany
E-mail: {valentin.haider, fidan.mehmeti, wolfgang.kellerer}@tum.de

***Abstract*—With the development of 5G, network slicing was proposed to enable service provisioning for diverse sets of Ultra-Reliable Low-Latency Communications (URLLC), enhanced Mobile Broadband (eMBB), and massive Machine-Type Communications (mMTC) users which are characterized by different Quality of Service (QoS) demands. Within network slicing, Radio Access Network (RAN) slicing plays a central role for efficient resource management. In addition, user admission control poses a major challenge. In this context, the problem of *joint* slice dimensioning and user admission control is investigated in this paper. To this end, an optimization problem based on *symbol*-level resource allocation with the objective of maximizing an operator's revenue while fulfilling the traffic requirements of all users is formulated. Afterward, the optimization problem is reduced to a knapsack problem and integrated into a Long-Term Revenue Maximization (LTRM) algorithm. Using data from real-world 5G measurements, the efficiency of the LTRM algorithm is verified, and the impact of various resource granularities in the time domain (*symbol* vs. *slot*) and frequency domain (varying Resource Block Group (RBG) sizes) is investigated. The revenue gain of the proposed *joint* algorithm over a sequential slice dimensioning and user admission control scheme is 24%, while *symbol*-level resource allocation offers at least 13% gain over a *slot*-based allocation for specific slices.**

***Index Terms*—5G and Beyond, RAN Slicing, User Admission Control, Optimization.**

I. Introduction

Network slicing is a new concept that was introduced in 5G to enable network operators to support diverse services in an economically sustainable manner [1]. The idea is to virtually separate physical resources in the Radio Access Network (RAN) and core network to provide End-to-End (E2E) Quality of Service (QoS) guarantees for diverse users [2]. These users can be grouped into three categories: Ultra-Reliable Low-Latency Communications (URLLC), like autonomous vehicles [3], enhanced Mobile Broadband (eMBB), corresponding to services like eXtended Reality (XR) [4], and massive Machine-Type Communications (mMTC), comprising a large number of sensors and actuators [5]. Within each of these categories, users can further be subdivided by specific QoS demands based on their use case. This is done using Service Level Agreements (SLAs), which classify the various slices within a traffic category.

An integral part of network slicing is *RAN slicing*, which deals with the problem of allocating radio resources to slices. With the resources available in RANs being finite, it is important to employ efficient allocation algorithms according to objectives like energy or resource efficiency, or operator revenue. Given the amount of resources available per slice, the issue of *user admission control* arises. This problem consists of selecting the users to admit to the slice/network such that all admitted users' guaranteed QoS are not violated.

So far, the problems of slice dimensioning and user admission control while guaranteeing a certain QoS have only been tackled sequentially or without QoS guarantees [6], [7]. In a sequential approach, the RAN resources are first distributed (statically or dynamically) among all slices, and afterward, users are admitted to these slices based on the available resources per slice.

In this work, however, the network performance is optimized by considering these two processes *jointly*. This is done by considering all users' channel conditions and traffic requirements. In this way, a new degree of freedom in making decisions is provided to the operators, as the slices can be tuned based on the users requesting a service at a certain point in time. The objective in this work is the maximization of the network operator's long-term profit while satisfying the QoS requirements of all admitted users at all times. The profit a user generates for the operator is defined by the type of traffic and the specific SLA. Generally, the stricter the traffic demands are, the higher the price for the specific service.

Besides frequency, the time domain is the other resource allocation dimension in current cellular systems. Works dealing with resource allocation in RANs so far have considered slots or Transmission Time Intervals (TTIs) as the granularity level of resources in the time domain [5], [8]. However, sometimes, this leads to wasting valuable network resources. Such an example arises when considering Hybrid Automatic Repeat Request (HARQ) processes [9]. Retransmissions might be delayed in case NACK packets and retransmissions are scheduled on a per-*slot* basis, although the previously received undecodable data packet or the NACK packet were processed much faster than the slot duration. Based on the 5G standardization, however, resource allocation in the time domain is possible on a lower scale, i.e., on the *symbol*-level, where 14 symbols comprise one slot [10]. This allows for a much faster (re)transmission of NACK and data packets. Hence, this is the considered granularity level in the time domain in this work. In the frequency domain, the considered resource granularity is a Resource Block Group (RBG), which consists of multiple Physical Resource Blocks (PRBs) [11].

In the sense of these elaborations, the important questions that arise for a network operator are: (i) How to dynamically dimension the network slices and which users to admit to each slice at certain points in time to maximize the long-term revenue? (ii) How does the resource granularity, both in the time and frequency domain, impact the achieved revenue?

To answer these questions, this paper investigates the problem of *joint* slice dimensioning and user admission control based on *symbol*-level resource allocation. The traffic requirements of URLLC, eMBB, and mMTC users are mathematically modeled in a generic way such that the specific SLAs for different network slices of the same traffic category can be specified based on various parameters. Leveraging these formulations, an optimization problem for joint slice dimensioning and user admission control is formulated. This problem is then analyzed and reduced to a knapsack problem, enabling an efficient solution. Finally, the devised knapsack problem is embedded into an algorithm for revenue maximization, which ensures the QoS of both previously and newly accepted users at all times. The system model introduced in this work is tightly coupled to the current 5G standards, e.g., the channel conditions of all users are described with the Channel Quality Indicator (CQI), or the different resource granularities used in the time and frequency domain [10]. Nevertheless, reasonable deviations from the standard allow for interesting insights relevant to the design process of future mobile communication systems. The simulation results, which are based on real-world 5G measurements, verify the effectiveness and practicality of the presented approach and show that it outperforms other revenue maximization solutions. Thus, the results are particularly relevant for network operators. Moreover, the findings are valuable for gaining insights into the impact of different resource granularities on network performance, i.e., *symbols* vs. *slots* in the time domain and different *RBG sizes* in the frequency domain. Summarizing, the main contributions are:

- Formulating an optimization problem for joint slice dimensioning and user admission control based on symbol-level resource allocation (Section II);
- An in-depth analysis of the formulated optimization problem resulting in the reduction of the problem to a 0-1 knapsack problem (Section III);
- Integrating the 0-1 knapsack problem into a Long-Term Revenue Maximization (LTRM) algorithm (Section III);
- An extensive simulation-based evaluation leveraging a real-world 5G dataset, emphasizing the performance of the proposed LTRM algorithm (Section IV).

II. Problem Formulation

In this section, the system model is introduced. Subsequently, the joint slice dimensioning and user admission control optimization problem is mathematically formulated. All variables are listed in Table I.

A. System Model

Slices and User Sets: In this work, a set of users $\mathcal{U}$ is in the coverage area of a single 5G Base Station (BS) (gNodeB) during a so-called Slice Dimensioning Interval (SDI) with duration T^{slice} (see Fig. 1). Each user can be mapped to a single slice from the URLLC, eMBB, or mMTC slice sets $\mathcal{S}^U$, $\mathcal{S}^E$, or $\mathcal{S}^M$, and moreover, either belongs to the set of users that are currently being served and will still be present in the upcoming SDI ($\mathcal{U}^s$), or to the set of waiting users that are trying to get served in the upcoming SDI(s) ($\mathcal{U}^w$). Since there exist multiple slice instances for the same traffic type, diverse QoS demands for specific services can be fulfilled. The BS has a total of N^{tot} PRBs to serve all users, from which N^{av} PRBs are available to admit users from the set $\mathcal{U}^w$ in the upcoming SDI, as the remaining resources are reserved for the users that were already accepted and are still getting served, i.e., the set $\mathcal{U}^s$. The newly accepted users per slice s are then grouped in the set $\mathcal{U}_s^{a,\{U,E,M\}}$. The admittance decision is taken with the objective of maximizing the network operator's revenue, which also leads to a re-dimensioning of all slices for the upcoming SDI. Note that once a user is admitted, it is served until it no longer requests its service, and that a user's service duration is not known at the time of admittance.

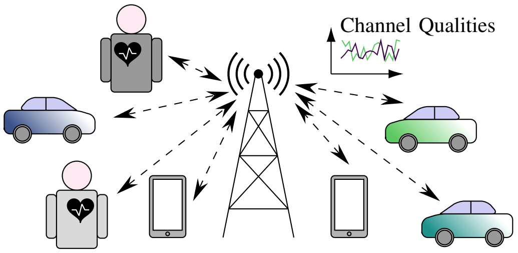

Fig. 1: Illustration of the system model.

Traffic Characterization: In general, URLLC users require their packets to be delivered within a maximum time D_s^{max} with high reliability $1-\epsilon_s^l$, where ϵ_s^l is the maximum allowed outage probability [3]. Hence, for the transmission of these packets, retransmissions, i.e., HARQ, are also considered. A packet is assumed to be successfully transmitted with probability $1-\delta$, i.e., the packet error rate is δ. Moreover, to enable faster retransmissions in order to comply with the strict delay requirements of URLLC users, *symbol*-level resource allocation is performed in the time domain. This allows for a 14 times more granular resource allocation and thus faster packet rescheduling than usual *slot*-based allocation. The eMBB users require a constant high data rate R_s^{min} [4], which requires efficient resource allocation of RBGs in the frequency domain. Lastly, mMTC users periodically send small packets [12] with periodicity λ_s, measured in packets per SDI, which enables multiplexing them in the time domain.

Channel Conditions and Modeling: Finally, each user experiences varying channel conditions over time. It is assumed that these conditions are constant over all PRBs, i.e., a user only reports one CQI value per Channel State Information (CSI) report period, which corresponds to wideband CQI reporting in 5G [10]. Based on the chosen Modulation and Coding Scheme (MCS) table and slot duration, the per-PRB rate can then be determined. Given the dynamic channel conditions and various CQI values of the users, the slice dimensioning is influenced by resource allocations performed in the *time* and *frequency* domains. For the optimization of slice dimensioning and user admission control, the CQI values experienced by a user are characterized by a probability

TABLE I: List of Symbols

$\mathcal{U}$	set of all users	$\mathcal{U}^{\{s,w\}}$	set of served/wait. users
$\mathcal{U}_s^{w,\{U,E,M\}}$	set of wait. users in URLLC/eMBB/mMTC slice s	$\mathcal{U}_s^{a,\{U,E,M\}}$	set of newly acc. users in URLLC/eMBB/mMTC slice s
N^{tot}	total no. of available RBGs	N^{av}	available RBGs in a specific SDI
T^{slice}	dur. of an SDI	$T_s^{\{slot,sym\}}$	dur. of a slot/symbol in slice s
$T^{p,\{BS,UE\}}$	proc. dur. of the BS/of user u	$T^{N/ACK}$	dur. of an ACK/NACK pkt.
$\mathcal{S}^{\{U,E,M\}}$	set of URLLC/eMBB/mMTC slices	v_u	value/rev. of user u
I_u, $\mathbf{I}$	var. spec. whether a user is acc. to the netw., vec. of vars. I_u	N_u, $\mathbf{N}$	no. of RBGs req. by user u, vec. of vars. N_u
$N_u^{d,t}$	no. of RBGs req. by user u, time-disc. res. alloc.	N_u^d	no. of RBGs req. by user u, time- and freq.-disc. res. alloc.
$S_u^{\{PRB,RBG\}}$	size of a PRB/RBG of user u in bits	T_u^{max}	max. wait. time of user u
D_u^{pkt}	eperienced pkt. lat. for URLLC user u	D_s^{max}	max. allowed lat. for a user's URLLC pkt. in slice s
ϵ_s^l	outage prob. for a URLLC pkt. of slice s	δ	pkt. error ratio
I_u^{pkt}	var. spec. whether a user has a pkt. to send	p_u	prob. for URLLC user u to have a pkt.
X_u	trans. attempts of a pkt. of URLLC user u (rand. var.)	Δ_u	pkt. size of user u
R_u	data rate experienced by user u	R_s^{min}	min. required data rate for every user of slice s
λ_s	mMTC arrival rate of slice s in pkts. per SDI	N_s^{sym}	no. of symbols in one mMTC period of slice s
N^M	no. of RBGs occupied by an mMTC user group	$N_u^{d,t,sym}$	no. of symbols req. by mMTC user u, time-disc. res. alloc.
$\mathcal{U}_{i,s}^{w,M}$	wait. user group i in mMTC slice s	$v_{i,s}^M$	value/rev. of mMTC user group i in mMTC slice s

distribution known to the operator or transmitted by each user together with its service request.

B. Optimization Problem Formulation

The user admission control problem that is solved for every SDI serves two purposes: Firstly, it determines which waiting users should be admitted to the network to maximize the operator's total revenue. Secondly, it specifies how many resources, i.e., RBGs, should be allocated to each network slice to optimally use the available resources during an SDI. The corresponding optimization problem is formulated as

$$\mathcal{P}_1\colon \max_{\mathbf{I}} \quad \sum_{u\in\mathcal{U}^w} v_u I_u \tag{1a}$$

$$\text{s.t.} \quad P\left(D_u^{pkt} \le D_s^{max}\right) \ge 1-\epsilon_s^l, \quad \forall u \in \mathcal{U}_s^{a,U},\ s\in\mathcal{S}^U, \tag{1b}$$

$$R_u \ge R_s^{min},\ \forall u \in \mathcal{U}_s^{a,E},\ s\in\mathcal{S}^E, \tag{1c}$$

$$N_u \ge \frac{\Delta_u\cdot\lambda_s}{S_u^{RBG}},\ \forall u\in\mathcal{U}_s^{a,M},\ s\in\mathcal{S}^M, \tag{1d}$$

$$\sum_{u\in\mathcal{U}^w} N_u\cdot I_u \le N^{av}, \tag{1e}$$

$$I_u\in\{0,1\},\ \forall u\in\mathcal{U}^w, \tag{1f}$$

where

$$D_u^{pkt} = \Big(X_u\left(\Delta_u/R_u + T^{p,UE}\right) + (X_u-1)\Big(T^{N/ACK}+T^{p,BS}\Big)\Big)\cdot I_u^{pkt} \tag{2}$$

denotes the latency experienced by a packet based on the number of transmissions X_u, and

$$R_u = N_u\cdot S_u^{RBG}/T_s^{slot} \tag{3}$$

is the data rate experienced by user u. It is calculated by multiplying the number of allocated RBGs N_u with the size of an RBG S_u^{RBG} and dividing by the duration of a slot T_s^{slot}. Thereby, S_u^{RBG} depends on the channel conditions of user u and T_s^{slot} depends on the configured numerology μ employed in slice s. The allocations N_u of all waiting users are summarized in the vector $\mathbf{N}$ ($|\mathcal{U}^w|\times 1$).

The first term of the summation in (2) comprises the transmission duration, which is computed by dividing the packet size Δ_u by the experienced data rate R_u, and the processing duration of the packet at the User Equipment (UE), $T^{p,UE}$. In case of a successful reception, the packet transmission is finished. If the packet cannot be correctly decoded, a NACK packet is sent to the BS ($T^{N/ACK}$), and a retransmission is triggered ($T^{p,BS}$), which is captured by the second term in (2). Lastly, the variable I_u^{pkt} indicates whether a user has a packet to send, i.e., it takes the value 1 with probability p_u and the value 0 with probability $1-p_u$.[1]

The objective (1a) of $\mathcal{P}_1$ is to maximize an operator's revenue by accepting these waiting users to its network that generate the highest profits and require the least amount of resources. Hence, the binary variable I_u indicates whether user u, which generates value v_u for the operator, is accepted in the network. The variables I_u are aggregated in the vector $\mathbf{I}$ ($|\mathcal{U}^w|\times 1$). If a user from one of the waiting user sets $\mathcal{U}_s^{w,U}$, $\mathcal{U}_s^{w,E}$, or $\mathcal{U}_s^{w,M}$ is accepted, i.e., $I_u = 1$, it is added to the respective accepted user sets $\mathcal{U}_s^{a,U}$, $\mathcal{U}_s^{a,E}$, or $\mathcal{U}_s^{a,M}$.

The users can have the different traffic types URLLC, eMBB, or mMTC with different QoS specifications, which are captured by (1b), (1c), or (1d), respectively. Eq. (1b) determines the resources required in the frequency domain to fulfill the latency requirement with reliability $1-\epsilon_s^l$, where the constant ϵ_s^l represents the maximum allowed outage probability for a URLLC packet in slice s. This allocation is expanded over the entire SDI to account for the probability p_u of having a packet to send. Since the rate requirement of an eMBB user needs to be fulfilled at any time, (1c) specifies the amount of resources reserved in the frequency domain for the entire SDI. Lastly, (1d) establishes the minimum required resources to transmit all arriving packets of a single mMTC

[1] Note that (2) captures the packet delay for downlink transmission. The packet delay for uplink transmission can be modeled similarly by exchanging the processing durations of the UE and the BS and by replacing the duration of the ACK packet with the duration of a new transmission grant. Moreover, note that all other constraints are agnostic to the transmission direction.

user during an SDI. Since there is no latency requirement, the needed resources for a single packet transmission can be allocated in the time or frequency domain. A network slice is thus characterized by one of the Eqs. (1b) to (1d).[2] Given the previously accepted users, there is only a limited amount of resources, i.e., RBGs, available, which is captured by (1e). Lastly, (1f) merely states that the decision variables are binary.

III. Analysis

In the following, $\mathcal{P}_1$ is reduced into a 0-1 knapsack problem. Afterward, the knapsack problem is integrated into the LTRM algorithm. To this end, first, the number of required RBGs per user are determined based on continuous resource allocation. Then, these values are employed to determine the required resources for a discrete resource allocation. Ultimately, these values are used as weights for the knapsack problem. To guarantee all SLAs of the different users, the data rates R_u are calculated based on a user's lowest possible CQI value.

A. Determination of the Number of Required Resources

Hereafter, all constraints related to the various traffic types are rewritten with the aim of determining the number of required resources per user N_u such that the slice-specific QoS guarantees can be fulfilled.

URLLC Users: To determine the required number of RBGs per URLLC user u, (1b) is rewritten as

$$\begin{aligned} &P\left(D_u^{pkt} \le D_s^{max} \mid I_u^{pkt} = 0\right) \cdot P\left(I_u^{pkt} = 0\right) + \\ &P\left(D_u^{pkt} \le D_s^{max} \mid I_u^{pkt} = 1\right) \cdot P\left(I_u^{pkt} = 1\right) \ge 1 - \epsilon_s^l, \end{aligned} \tag{4}$$

by conditioning on having a packet. Next, (4) is solved to

$$P\left(D_u^{pkt} \le D_s^{max} \mid I_u^{pkt} = 1\right) \ge 1 - \epsilon_s^l / p_u,$$

by substituting the probabilities of having a packet or not. Replacing the experienced packet delay D_u^{pkt} by (2) leads to

$$\begin{aligned} P\Big(X_u \left(\Delta_u / R_u + T^{p,UE} + T^{N/ACK} + T^{p,BS}\right) \\ - T^{N/ACK} - T^{p,BS} \le D_s^{max}\Big) \ge 1 - \epsilon_s^l / p_u, \end{aligned}$$

which is rewritten as

$$\begin{aligned} P\left(X_u \le \frac{D_s^{max} + T^{N/ACK} + T^{p,BS}}{\Delta_u / R_u + T^{p,UE} + T^{N/ACK} + T^{p,BS}}\right) \\ \ge 1 - \epsilon_s^l / p_u, \end{aligned} \tag{5}$$

by solving the argument of the probability function for the random variable X_u. The left-hand side (LHS) of (5) is now equal to the Cumulative Distribution Function (CDF) of the random variable X_u at the point

$$x = \frac{D_s^{max} + T^{N/ACK} + T^{p,BS}}{\Delta_u / R_u + T^{p,UE} + T^{N/ACK} + T^{p,BS}}. \tag{6}$$

Hence, (5) can be expressed as

$$F_{X_u}(x) \ge 1 - \epsilon_s^l / p_u. \tag{7}$$

Recalling the assumption that each packet is successfully transmitted with probability $1 - \delta$, the variable X_u is geometrically distributed, i.e., $X_u \sim \mathcal{G}(1 - \delta)$. With the CDF of a geometric distribution given as $1 - (1 - p)^{\lfloor x \rfloor}$ [13], where p is the success probability, (7) is rewritten as

$$1 - (1 - (1 - \delta))^{\lfloor x \rfloor} \ge 1 - \epsilon_s^l / p_u.$$

This term can be reformulated as

$$\delta^{\lfloor x \rfloor} \le \epsilon_s^l / p_u.$$

Taking the logarithm of both sides, then taking $\lfloor x \rfloor$ out of the logarithm and dividing by $\log \delta$ results in

$$\log\left(\epsilon_s^l / p_u\right) / \log \delta \le \lfloor x \rfloor, \tag{8}$$

since $\log \delta$ is a negative number ($\delta < 1$).

Result 1. *The statements*

$$x \le \lfloor y \rfloor, \tag{9a}$$

$$\lceil x \rceil \le y \tag{9b}$$

are equivalent for $x, y \in \mathbb{R}_0^+$.

Proof. Define, $x = x' - e_1$ and $y = y' + e_2$, where $x', y' \in \mathbb{N}_0$ and $e_1, e_2 \in [0, 1)$. Then, (9b) can be reformulated as

$$\lceil x \rceil = x' = x + e_1 \le y \Leftrightarrow e_1 \le y - x. \tag{10}$$

Substituting the equations for x, y into (10) leads to

$$e_1 \le y' - x' + e_2 + e_1 \Leftrightarrow -e_2 \le y' - x'. \tag{11}$$

Since $y', x' \in \mathbb{N}_0$ and $e_1, e_2 \in [0, 1)$, it follows from (11) that

$$0 \le y' - x', \tag{12}$$

but also that

$$-e_1 \le y' - x'. \tag{13}$$

Adding e_2 on both sides of (13) and rearranging yields

$$e_2 \le y' - x' + e_2 + e_1 \Leftrightarrow e_2 \le y - x. \tag{14}$$

Reordering (14) and substituting y' for $y - e_2$ results in

$$x \le y - e_2 = y' \Leftrightarrow x \le \lfloor y \rfloor, \tag{15}$$

which concludes the proof. □

Using Result 1, (8) is written as

$$\left\lceil \log\left(\epsilon_s^l / p_u\right) / \log \delta \right\rceil \le x. \tag{16}$$

Substituting (6) into (16) and replacing R_u by (3) leads to

$$\left\lceil \frac{\log\left(\epsilon_s^l / p_u\right)}{\log \delta} \right\rceil \le \frac{D_s^{max} + T^{N/ACK} + T^{p,BS}}{\frac{\Delta_u \cdot T_s^{slot}}{N_u \cdot S_u^{RBG}} + T^{p,UE} + T^{N/ACK} + T^{p,BS}}, \tag{17}$$

which, after some calculus, can be solved for N_u as

$$\begin{aligned} \frac{\Delta_u \cdot T_s^{slot}}{S_u^{RBG}} \Bigg(\frac{D_s^{max} + T^{N/ACK} + T^{p,BS}}{\left\lceil \log\left(\epsilon_s^l / p_u\right) / \log \delta \right\rceil} - T^{p,UE} \\ - T^{N/ACK} - T^{p,BS} \Bigg)^{-1} \le N_u. \end{aligned} \tag{18}$$

eMBB Users: Since the data rate constraint (1c) does not include any probabilities, it can easily be solved for the number of required RBGs

$$N_u \ge \left(R_s^{min} \cdot T_s^{slot}\right) / S_u^{RBG}, \tag{19}$$

when substituting (3) for the experienced data rate R_u.

mMTC Users: Lastly, (1d) already determines the number of required RBGs of an mMTC user.

B. Discretizing the Number of Required Resources

The established bounds (1d), (18), (19) for the required RBGs are deduced for a continuous allocation in the time and

[2] Note that the allowed outage probability ϵ_s^l, as well as the minimum rate R_s^{min} can also be set individually per user instead of specifying them per slice. Nevertheless, this does not influence the subsequent analysis.

frequency domain. To comply with the real-world discrete allocation units, these numbers are adapted in the following.

URLLC Users: As the processing times $T^{p,BS}$ and $T^{p,UE}$, and $T^{N/ACK}$ are constant and expressed as a multiple of the symbol duration T_s^{sym}, the only variable time duration in the experienced delay of a packet is the transmission time. To fulfill the URLLC guarantee, the transmission time dependent on a time-discrete allocation must be smaller than or equal to the transmission time dependent on the continuous allocation. Thus, discretizing the transmission time is achieved by setting

$$\frac{\Delta_u \cdot T_s^{slot}}{N_u^{d,t} \cdot S_u^{RBG}} = \left\lfloor \frac{\frac{\Delta_u \cdot T_s^{slot}}{N_u \cdot S_u^{RBG}}}{T_s^{sym}} \right\rfloor \cdot T_s^{sym},$$

which can be solved for the required RBGs depending on a time-discrete allocation $N_u^{d,t}$ as

$$N_u^{d,t} = \left\lfloor \frac{\frac{\Delta_u \cdot T_s^{slot}}{N_u \cdot S_u^{RBG}}}{T_s^{sym}} \right\rfloor^{-1} \cdot \frac{\Delta_u \cdot T_s^{slot}}{S_u^{RBG} \cdot T_s^{sym}}.$$

As the final step for discretizing the allocation also in the frequency domain, the required RBGs are determined as

$$N_u^d = \left\lceil N_u^{d,t} \right\rceil. \tag{20}$$

eMBB Users: Since eMBB users require a constant data rate, the resources reserved for such a user are allocated for an entire SDI. Hence, no discretization in the time domain is necessary for (19). To discretize the resource allocation in the frequency domain, (20) can be applied when replacing $N_u^{d,t}$ by N_u of an eMBB user (19).

mMTC Users: Lastly, mMTC users only require a few resources periodically, which allows for multiplexing the users in the time domain and creating mMTC groups. To discretize the number of required resources of a single user, the number of symbols required by a user during one mMTC period, defined as T^{slice}/λ_s, is computed as

$$N_u^{d,t,sym} = \left\lceil \left(\Delta_u / S_u^{RBG}\right) \cdot 14 \right\rceil. \tag{21}$$

Thereby, the factor 14 results from the fact that one time slot in 5G consists of 14 symbols [11]. Next, the total amount of available symbols in one mMTC period is determined as

$$N_s^{sym} = \left\lfloor \frac{14 \cdot T^{slice}}{\lambda_s \cdot T_s^{slot}} \right\rfloor \cdot N^M, \tag{22}$$

where N^M denotes the number of RBGs that are occupied by a single mMTC user group. As opposed to (21) where the number of *required* symbols was determined, the floor function is used in (22) to determine the number of *available* symbols since no half symbol can be allocated to any user. To finally create the mMTC user groups, the 0-1 knapsack problem

$$\begin{aligned} \mathcal{P}_2\colon \max_{\mathbf{I}^M} & \sum_{u \in \mathcal{U}_s^{w,M}} v_u I_u \\ \text{s.t.} & \sum_{u \in \mathcal{U}_s^{w,M}} N_u^{d,t,sym} \cdot I_u \leq N_s^{sym}, \\ & I_u \in \{0,1\},\ \forall u \in \mathcal{U}_s^{w,M} \end{aligned}$$

is solved multiple times for every mMTC slice s with the specific slice arrival rate λ_s, where the users that are already served are removed from the set $\mathcal{U}_s^{w,M}$ in every next iteration.

Algorithm 1 Creation of mMTC user groups

```
Input: $v_u$, $N_u^{d,t,sym}$, $\forall u \in \mathcal{U}_s^{w,M}, s \in \mathcal{S}^M$; $N_s^{sym}$, $\forall s \in \mathcal{S}^M$
Output: $\mathcal{U}_{i,s}^{w,M}$ assoc. with $v_{i,s}^M$, $\forall\, i$, $s \in \mathcal{S}^M$
1: for all $s \in \mathcal{S}^M$ do
2:     $i = 0$
3:     while $\mathcal{U}_s^{w,M} \neq \emptyset$ do
4:         Solve $\mathcal{P}_2$
5:         $\mathcal{U}_{i,s}^{w,M} = \{u \in \mathcal{U}_s^{w,M} \mid I_u = 1\}$
6:         $v_{i,s}^M = \sum_{u \in \mathcal{U}_{i,s}^{w,M}} v_u$
7:         $\mathcal{U}_s^{w,M} = \mathcal{U}_s^{w,M} \setminus \mathcal{U}_{i,s}^{w,M}$, $i = i + 1$
```

This procedure is summarized in Algorithm 1. In the worst case, no multiplexing is possible, which implies that the mMTC user group sets $\mathcal{U}_{i,s}^{w,M}$ only contain one user. Hence, only one user would be removed from the set of all mMTC users of slice s in line 7. This means that $\mathcal{P}_2$ is solved at most $\left|\mathcal{U}_s^{w,M}\right|$ times. Therefore, the time complexity of Algorithm 1 is $\mathcal{O}(\left|\mathcal{U}_s^{w,M}\right| \left|\mathcal{S}^M\right| T)$, where T determines the complexity of the algorithm employed to solve $\mathcal{P}_2$.

Based on the solutions to $\mathcal{P}_2$, the mMTC user groups $\mathcal{U}_{i,s}^{w,M}$ with value $v_{i,s}^M$ can now be regarded as users attempting to enter mMTC slice s. The amount of resources this group requires is equal to N^M RBGs.

C. Transformation of $\mathcal{P}_1$ to a 0-1 Knapsack Problem

With the preceding discretization of the number of required resources both in the time and frequency domain, every user or user group attempting to enter one of the eMBB, URLLC, or mMTC slices can now be associated with a value v_u and a weight equal to the required number of RBGs N_u^d. This allows for transforming $\mathcal{P}_1$ to the 0-1 knapsack problem

$$\begin{aligned} \mathcal{P}_3\colon \max_{\mathbf{I}} & \sum_{u \in \mathcal{U}^w} v_u I_u \\ \text{s.t.} & \sum_{u \in \mathcal{U}^w} N_u^d \cdot I_u \leq N^{av}, \\ & I_u \in \{0,1\},\ \forall u \in \mathcal{U}^w. \end{aligned}$$

By solving $\mathcal{P}_3$, the set of users accepted in each slice is determined. Based on the required number of resources per user contained in the set of accepted users per slice, every slice can eventually be dimensioned in the frequency domain.

D. Long-Term Revenue Maximization Algorithm

The final subsection presents the LTRM algorithm. For each SDI, new users are arriving and are added to the set $\mathcal{U}^w$. They are associated with a maximum waiting time T_u^{max}, value v_u, and weight N_u^d. At the end of each SDI, the slices are dimensioned in the frequency domain, and admission control for the set $\mathcal{U}^w$ is performed for the upcoming SDI. First, the set of users that will still be present in the upcoming SDI $\mathcal{U}^s$ is updated. Based on these updates, the remaining available RBGs N^{av} are determined. Next, the joint slice dimensioning and user admission control optimization problem $\mathcal{P}_3$ is solved. Afterward, the newly distributed resources are added to the current slice configurations based on the required RBGs of accepted users, and the served and waiting user sets $\mathcal{U}^s$ and $\mathcal{U}^w$

Algorithm 2 Long-Term Revenue Maximization (LTRM)

```
Input: $v_u, N_u^d, T_u^{max} \ \forall u \in \mathcal{U}^w$
for every SDI do
    Update $\mathcal{U}^s$ based on service request terminations
    $N^{av} = N^{tot} - \sum_{u \in \mathcal{U}^s} N_u^d$
    Solve $\mathcal{P}_3$ given $N^{av}$, $\mathcal{U}^w$
    Update slice configurations,
    add/remove newly accepted users to $\mathcal{U}^s$/from $\mathcal{U}^w$
    for all $u \in \mathcal{U}^w$ do
        if waiting time $> T_u^{max}$ then
            Remove $u$ from $\mathcal{U}^w$
```

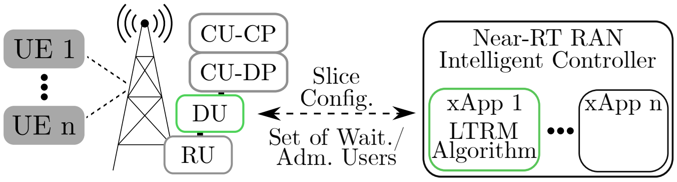

Fig. 2: Placement of the LTRM Algorithm as an xApp within the O-RAN architecture.

are updated based on the newly accepted users. Lastly, if a user exceeds its maximum waiting time, it is finally rejected and removed from the waiting list. The procedure is summarized in Algorithm 2. Its complexity per SDI is $\mathcal{O}\left(|\mathcal{U}^s| + |\mathcal{U}^w| + T\right)$, where T is the complexity of the algorithm used to solve $\mathcal{P}_3$.[3]

Deployment Considerations: Due to the polynomial time complexity of the LTRM algorithm, it could, e.g., be implemented as an xApp [16] within the O-RAN [17] architecture. This xApp then performs dynamic slice resource management as well as user admission control. An architectural overview of such a network is depicted in Fig. 2.

IV. Performance Evaluation

In this section, first, the simulation setup is described and the benchmarks are introduced. Then, the performance of the LTRM algorithm is evaluated, and insights into the evolution of slice dimensions are given. Finally, the impact of various resource granularities is assessed.

A. Simulation Setup

To obtain a CQI probability distribution for each user, data from a 5G measurement campaign was used [18]. These measurements were realized during different time periods for a single user that was either static (23 traces) or moving around (59 traces). For each trace, a probability distribution is calculated based on the measured CQI values. A statistical analysis of the CQI distributions is provided in [19].

To calculate the size of an RBG in bits, the reported CQI value is first mapped to an MCS index by using the mapping table from [20]. Afterward, depending on the type of traffic, the suitable MCS table is chosen from [10] to map the MCS index to a modulation order and a code rate. Finally, assuming that 156 Resource Elements (REs) are available for data transmission per Resource Block (RB) [10], the PRB size in bit is calculated according to the standardized formulas [10].

For the simulations, seven different slices were created. Three URLLC slices that correspond to the use cases of an intelligent transport system [21], a smart grid millisecond-level precise load control [22], and mobile robot machine control [23]. These use cases and their QoS requirements are defined by the European Telecommunications Standards Institute (ETSI) [24], [25]. It was assumed that roughly one-third of the E2E latency can be spent on the radio transmission to define the maximum allowed latency. Besides, three eMBB slices are specified, where the first two correspond to mobile users, e.g., pedestrians walking around or users driving in a car, while the third corresponds to static users, e.g., office workers. The required minimum data rates were chosen according to [3]. Lastly, the mMTC slice corresponds to the use case of Internet of Things (IoT) devices with downlink traffic [5], [24], [26]. For each slice, the number of arriving users per SDI is modeled as a Poisson process with various arrival rates. Service times are assumed to be uniformly distributed within predefined intervals per slice. The user values v_u are determined per slice, i.e., per SLA, and based on the average resources required for fulfilling an SLA. For each slice, the user's CQI probability distribution is randomly sampled from the data traces without replacement according to the mobility type of the user. The processing durations of the UE and the BS as well as the duration of a (N)ACK packet are calculated as a multiple of the symbol duration based on the values in [27]. Since all packet sizes of the URLLC traffic are smaller than the maximum code block size [28], a packet is transmitted using a single code block. Hence, a target Block Error Rate (BLER) of 10^{-4} [29] can be taken as the packet error rate δ. The slot duration for all slices is $0.5\,\text{ms}$, which corresponds to numerology $\mu = 1$. For this numerology, the maximum bandwidth is $100\,\text{MHz}$, which corresponds to 273 PRBs [30]. Depending on the active Bandwidth Part (BWP) size, the corresponding minimum number of PRBs comprising a single RBG is in the set $\mathcal{P} = \{2, 4, 8, 16\}$ [10]. However, the value 1 is also used to show the benefits of a smaller resource granularity in the frequency domain. All parameter values used in the simulations are summarized in Table II[4]. In total, 25 discrete-time simulation runs with a duration of 1200 SDIs were conducted for every RBG size P. To solve the knapsack problems, DP was used. The simulations were conducted in MATLAB R2024a on a computer with an Apple M1 processor using 16 GB of RAM, running macOS 14.2.1.

B. Benchmark Algorithms

The solutions obtained with the LTRM algorithm are compared against three benchmarks. For the first one, $\mathcal{P}_3$ is solved with the exact same solution approach. However, resource allocation in the time domain is done on a per-*slot* basis,

[3] To solve 0-1 knapsack problems, there exist very efficient algorithms like Dynamic Programming (DP) (pseudo-polynomial time), a Fully Polynomial Time Approximation Scheme (FPTAS) (polynomial time with performance guarantee), or a Greedy algorithm [14], [15].

[4] Wherever possible, references are given for simulation parameters. In all other cases, reasonable values were chosen based on use case descriptions.

TABLE II: Simulation Parameters [3], [5], [10], [24]–[27], [29]

Slice Number	1 [24]	2 [25]	3 [25]	4 [3]	5 [3]	6 [3]	7 [5], [24], [26]
Traffic/Mobility Type	URLLC/Static	URLLC/Static	URLLC/Mobile	eMBB/Mobile	eMBB/Mobile	eMBB/Static	mMTC/Mobile
User Value v_u	6	5	7	20	34	48	1
User Arrival Rate [user arrivals/SDI]	1	3	2	2	2	1	15
Max. Queue Waiting Time [SDI]	30	1	20	5	5	5	20
Min./Max. Service Durations [SDI]	{60, …, 144}	{1, …, 3}	{30, …, 90}	{30, …, 120}	{30, …, 120}	{30, …, 120}	{60, …, 180}
Packet Size Δ_u [bit]	2208	960	1320	—	—	—	8160
Max. Delay D_s^{max} [ms]	10	20	3	—	—	—	—
Reliability (1-ϵ_s^l) [%]	99.999	99.9	99.99	—	—	—	—
Packet Probability p_u [%]	20	80	33.3	—	—	—	—
Minimum Data Rates R_s^{min} [Mbps]	—	—	—	3	5	10	—
Packet Arrival Rate λ_s [pkts./SDI]	—	—	—	—	—	—	20
Avail. RBGs N^M per P [RBG]	—	—	—	—	—	—	{4, 2, 1, 1, 1}

Proc. Dur. at the BS plus (N)ACK Dur. $T^{p,BS}$ + $T^{N/ACK}$ [ms] [27]	0.2143	Packet Error Rate δ [29]	10^{-4}
Proc. Dur. at the UE $T^{p,UE}$ [ms] [27]	0.1429	PRBs per RBG P [10]	{1, 2, 4, 8, 16}
Numerology μ; Slot Dur. [ms]; Number of PRBs [30]	1; 0.5; 273	Slice Dimensioning Interval T^{slice} [ms]	1000

TABLE III: Bandwidths [RBGs] for Benchmark 2/3

P \ Slice	1	2	3	4	5	6	7
1	39/25	39/6	39/50	39/25	39/51	39/102	39/13
2	19/13	20/3	19/25	20/13	19/25	20/51	19/6
4	10/6	10/2	9/13	10/6	10/13	9/25	10/3
8	5/3	5/1	5/6	4/3	5/6	5/13	5/2
16	2/2	3/1	2/3	3/2	2/3	3/5	2/1

which is the resource granularity in [31], [32]. By comparing to this benchmark, called Benchmark 1 (BM1), the benefit of exploiting the presented *symbol*-based allocation is demonstrated. Furthermore, the second and third benchmarks follow an approach where slice dimensioning and admission control are not performed jointly. This implies that the slices are first dimensioned in terms of the allocated RBGs using a slice dimensioning algorithm, e.g., [8], [33], and afterward, the admission control is performed depending on the number of available RBGs in each slice. For user admission control, on the one hand, the Partially Adaptive GrEedy (PAGE) algorithm [34], referred to as Benchmark 2 (BM2), is employed. Note that only one user class is considered in this work, and hence, the PAGE algorithm is adapted to support only best-effort users. On the other hand, user admission control is performed using a Greedy algorithm in the sense of [34], where users are ordered based on their values and resource demands. This algorithm is called Benchmark 3 (BM3). Two different constant slice allocations are used for BM2 and BM3. For the first, referred to as BM2.1 or BM3.1, the available RBGs are shared equally among all slices. For the second, called BM2.2 or BM3.2, the RBGs are distributed to the slices proportionally to the users' average resource demand per slice. The different numbers of allocated RBGs per dimensioning approach and per RBG size P are summarized in Table III.[5] In general, the complexity of the sequential approach depends on whether the slice dimensioning is performed statically or dynamically and on the complexity of the admission control algorithm, which is $\mathcal{O}\left(|\mathcal{U}^w| \log |\mathcal{U}^w|\right)$ for the PAGE algorithm.

C. Simulation Results

First, the performance of the LTRM algorithm is compared to all benchmarks in terms of the achieved operator revenue. To this end, the average revenue per SDI was calculated for every simulation run based on 1000 data points. The first 200 SDIs were not considered for evaluation as these data points represent the initialization phase of the system. Afterward, the mean value across all 25 average revenues was calculated together with the 99% confidence interval. The results are depicted in Fig. 3. Several conclusions can be drawn:

Firstly, the average revenue decreases with an increasing RBG size P. The reason is that a larger P leads to a larger resource granularity in the frequency domain, and thus, the resources can be split less efficiently according to the users' demands. The revenue gain from the largest RBG size $P = 16$ to the smallest size $P = 1$ is 40.19%. Therefore, *for future 6G systems, it is highly desirable to enable smaller RBG sizes for large BWP sizes.*

Secondly, for a specific P, the joint slice dimensioning and user admission control approach based on *slot*-level allocation (BM1) outperforms all other benchmarks. This proves the superiority of the joint approach over the disjoint solution methods. Furthermore, BM3 achieves a higher average revenue for a given slice configuration. The reason for this observation is that the PAGE algorithm (BM2) prioritizes users whose maximum waiting time is almost reached. Thus, users are getting served although they generate less revenue given the resources they require compared to other waiting users.

Thirdly and most importantly, for all RBG sizes P, the proposed LTRM algorithm outperforms both BM2 and BM3 for both slice configurations. Moreover, the revenues based on *symbol*-level resource allocation outperform those achieved in case *slot*-level resource allocation was conducted. Hence, it can be concluded that the LTRM algorithm optimally exploits the small resource granularity in the time domain, i.e., symbols, and perfectly dimensions the slice widths for every SDI according to the resources the users require and the revenue they generate. The performance gains achieved when using the LTRM algorithm instead of BM1 or the best performing BM2 or BM3, respectively, are added in Fig. 3.

[5] Considering dynamic resource allocation approaches for disjoint slice dimensioning and user admission control is deferred to future work, as this poses the challenge of guaranteeing previously accepted users' SLAs when redimensioning slices.

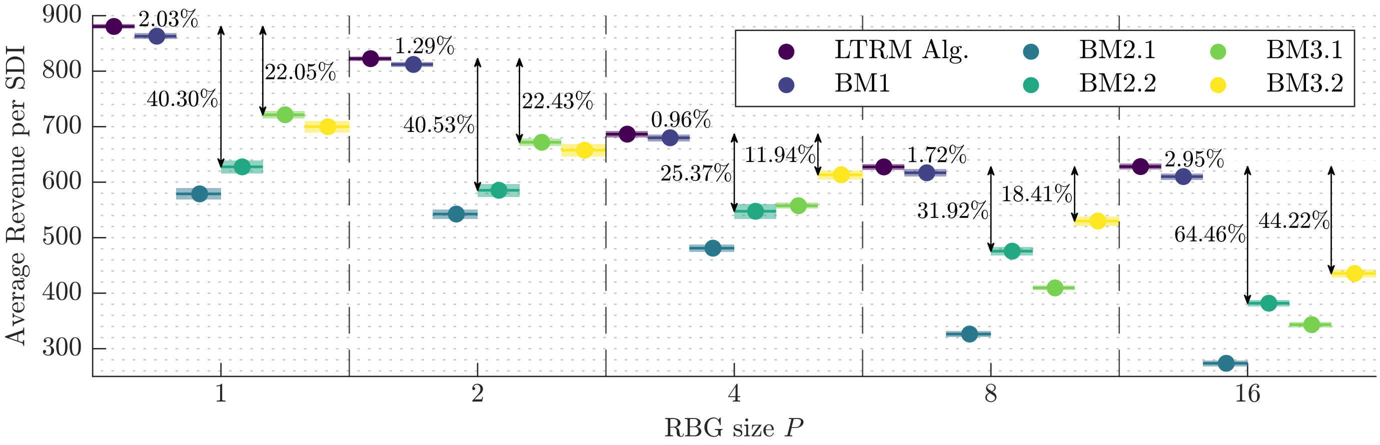

Fig. 3: Average revenues per SDI including the 99% confidence interval of the LTRM alg., BM1, BM2.1, BM2.2, BM3.1, and BM3.2 for $P \in \mathcal{P}$ and the performance gains (in %) of the LTRM alg. over BM1 and the best-performing BM2 and BM3.

The average revenue gain across all RBG sizes P over BM1 is 1.79%, while it is 40.52% over the best performing BM2 and 23.81% over the best performing BM3.

Next, the user acceptance ratios, defined as the number of accepted users divided by all users that were either accepted or rejected to/from the system, are investigated. The user acceptance ratios of selected slices for $P = \{1, 2\}$ achieved with the LTRM algorithm and the PAGE algorithm with slice widths based on the users' average resource demand (BM2.2) are depicted in Fig. 4. It is discernible that the user acceptance ratio highly depends on the specific service type and different resource granularities in the frequency domain, i.e., the RBG size P. For all slices but the mMTC slice, the acceptance ratios are higher for smaller RBG sizes P. This trend was also perceivable for higher P and can especially be observed for the URLLC slices, as these users generally require a small amount of resources, and hence their profit-to-resource consumption factor decreases with increasing P. The user acceptance ratios for the eMBB slice are quite low due to the high resource demand of these users and the comparably high service durations. The acceptance ratios for the other eMBB slices were in the same range as the depicted values. Generally, the acceptance ratios achieved with the LTRM algorithm are higher than the ratios obtained with the PAGE algorithm, as the slice widths can be adjusted perfectly to the demands of the waiting users. Only for the eMBB users the acceptance ratios are higher for BM2.2, as there is a fixed amount of resources reserved for these users, whereas resources are reallocated to other slices with the LTRM algorithm in case there are waiting users in other slices that have a higher profit-to-resource consumption factor. The acceptance ratios for the mMTC slice are close to zero for the PAGE algorithm, as there is a very high number of arriving users and the available resources in an SDI cannot be adjusted to the demands. Similar conclusions can be drawn when comparing the LTRM algorithm to the other benchmarks. Finally, it must be noted that the acceptance ratios and thus the fairness among the users can be strongly influenced by adjusting the prices for a specific service, which influences a user's profit-to-resource consumption factor.[6]

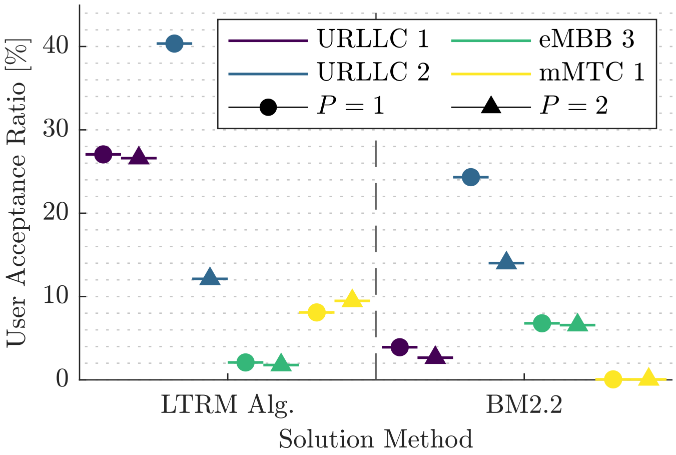

Fig. 4: User acceptance ratios over 1000 SDIs for the LTRM algorithm and BM2.2 for $P = 1$ and $P = 2$ for selected slices.

To verify the slice width adaptations realized by the LTRM algorithm and to compare them among the different slices and for various RBG sizes P, the coefficient of variation (c_v) for the slice widths in RBGs is given in Table IV. If no value is shown, these slices were not allocated any resources for this RBG size since only users from other slices were accepted in the network due to better profit-to-resource consumption factors. The c_v is higher the smaller the service durations are, i.e., the higher the user fluctuation is. No clear trend is apparent when comparing c_v across different P. However, for large P and for high consuming users, i.e., eMBB slices, c_v tends to be larger, as single users with a very high resource demand that were accepted or are leaving the system have a large influence on the slice width in a given SDI.

To emphasize the impact of *symbol*-level resource allocation in the time domain, the accumulated revenues of all URLLC

[6] To ensure a specific minimum fairness among all users, minimum acceptance rates per slice can be added as constraints to the optimization problem, or resources can be reserved for certain slices with low acceptance ratios. This task is beyond the scope of this paper and is deferred to future work.

TABLE IV: Coefficient of variation (c_v) for the slice widths in RBGs for the LTRM algorithm over 1000 SDIs

Slice / P	1	2	3	4	5	6	7
1	0.145	0.522	0.197	0.433	0.588	0.660	0.143
2	0.151	1.541	0.216	0.436	0.559	0.751	0.129
4	—	—	0.430	0.431	0.566	0.663	0.058
8	—	—	—	0.350	0.837	0.750	0.106
16	—	—	—	1.930	0.438	0.490	0.167

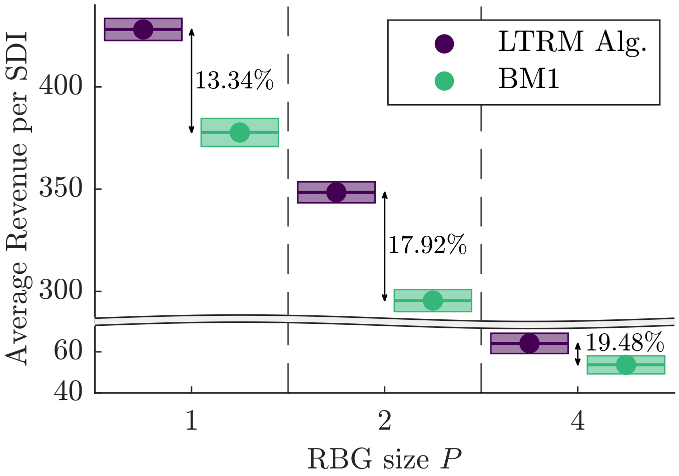

Fig. 5: Accumulated average URLLC revenues per SDI for the LTRM algorithm and BM1 for $P = 1$, $P = 2$, $P = 4$.

slices achieved with the LTRM algorithm and with BM1 are depicted for $P = \{1, 2, 4\}$ in Fig. 5. Since eMBB users constantly require resources, the resource granularity in the time domain does not influence them. However, URLLC and mMTC users are positively affected by smaller resource granularities in the time domain due to faster retransmissions and better-fitted resource allocations. Hence, there are large revenue gains of 13.34% to 19.48% observable that are achieved when only considering URLLC slices. Since the revenue gains shown in Fig. 3 are based on all slices and the individual value of eMBB users is quite high, the observed revenue gains over BM1 are much smaller.

Finally, the average run times over 1000 SDIs and 25 simulation runs of the DP algorithm employed to solve $\mathcal{P}_3$, which is the dominating time-consuming process within the LTRM algorithm, were analyzed. Although the simulations were conducted on a personal computer, the average runtimes are below 1 ms for all RBG sizes P. These values emphasize the practicality and scalability of the proposed algorithm and show that SDI durations as small as 10 ms are possible. Smaller intervals are not considered useful as the slice reconfiguration overhead would be excessive.

V. Related Work

There exist several works on resource allocation to various network slices, mainly with the objective of revenue maximization [8], [32], [33], [35]–[39]. Due to the complexity of allocating radio resources while fulfilling SLAs, many works rely on Reinforcement Learning (RL) based solutions [32], [33], [37], [38]. Since the RL agents require the SLA satisfaction rate as feedback to take proper actions, these solutions have to cope with the disadvantage of being reactive, i.e., no QoS guarantees can be given. In [35], an auction-based approach is taken, where service and network providers place bids by requesting and offering resources. The GREET [36] algorithm allocates radio resources across a set of BSs to various network slices. Enough resources are assumed to be available to fulfill all slices' minimum resource demands while GREET distributes all remaining resources. In [8], the authors leverage different numerologies and, thus, multiplex slices in the time and frequency domain for efficient radio resource usage. The authors of [39] develop a network slicing framework including algorithms for network slice pricing and resource allocation with the objective of maximizing a service provider's profit and overall resource utilization. Since the approach relies on a resource demand predictor, QoS guarantees cannot be ensured.

To efficiently manage the network load, there exists a large body of slice and user admission control schemes [31], [34], [40]–[42]. The works [40], [41] deal with slice admission control for heterogeneous slices based on queueing systems. The objective of [41] is to maximize the network slice providers' long-term revenue, while the algorithm in [40] is designed for general utility functions. In [31] and [42], user admission control algorithms are presented. While the authors of [31] propose both an offline and an online algorithm for admission control of users' service requests, an algorithm with the objective of maximizing revenue and resource utilization is devised in [42]. In [34], a user admission control heuristic called PAGE algorithm is introduced. The authors consider two different willingness-to-pay user classes. Admission control is then done using a Greedy algorithm, where users are ordered based on their profit, resource demand, and queueing time.

The works [6], [7], [43] deal with both user admission control/slice association and dimensioning. Although the authors of [6] design a joint optimization problem, they solve the problem sequentially by first performing user admission control and then conducting slice association and bandwidth allocation. The authors of [43] develop a RAN slicing framework that enables to perform slice admission and dimensioning, and UE to slice association. The slice dimensioning is done periodically based on the SLA statuses. However, no specific slice admission control policy or resource allocation mechanism has been developed. A slice dimensioning framework based on SLA fulfillment, slice configuration, and channel quality monitoring is devised in [7]. User admission control is also performed, but due to necessary data acquisition processes, the presented approach has a response time in the order of seconds, which is two orders of magnitude greater than the potential SDIs in the present work.

VI. Conclusion

In this paper, a joint slice dimensioning and user admission control optimization problem was formulated and integrated into an algorithm for long-term revenue maximization. The novel joint approach offers considerable performance gains.

The maximum observed gain between the average per-SDI revenues across all RBG sizes P is 64.46%. Besides, conducting RAN resource allocation on the *symbol*-level offers significant performance gains for URLLC and mMTC slices. An operator's revenue can be increased further by decreasing the RBG size for resource allocation in the frequency domain. The increase from $P = 16$ to $P = 1$ is 40.19%. Finally, the time complexity analysis and the run time measurements prove the applicability of the proposed algorithm in real networks. Considering fairness among users and including mobility management in a multi-BS setup is seen as possible future work.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research of Germany (BMBF) under the projects "6G-Life" and "6G-ANNA", with project identification numbers 16KISK002 and 16KISK107.

References

[1] S. Zhang, "An overview of network slicing for 5G," *IEEE Wirel. Commun.*, vol. 26, no. 3, 2019.

[2] H. Babbar, S. Rani, A. A. AlZubi, A. Singh, N. Nasser, and A. Ali, "Role of network slicing in software defined networking for 5G: Use cases and future directions," *IEEE Wirel. Commun.*, vol. 29, no. 1, 2022.

[3] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE ACM Trans. Netw.*, vol. 28, no. 2, 2020.

[4] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, 2018.

[5] F. Mehmeti and T. F. La Porta, "Modeling and analysis of mMTC traffic in 5G base stations," in *Proc. IEEE CCNC*, 2022.

[6] Y. Sun, S. Qin, G. Feng, L. Zhang, and M. A. Imran, "Service provisioning framework for RAN slicing: User admissibility, slice association and bandwidth allocation," *IEEE Trans. Mob. Comput.*, vol. 20, no. 12, 2021.

[7] M. Maule, J. Vardakas, and C. Verikoukis, "5G RAN slicing: Dynamic single tenant radio resource orchestration for eMBB traffic within a multi-slice scenario," *IEEE Commun. Mag.*, vol. 59, no. 3, 2021.

[8] C.-Y. Chang, N. Nikaein, and T. Spyropoulos, "Radio access network resource slicing for flexible service execution," in *Proc. IEEE Infocom WKSHPS*, 2018.

[9] S. R. Khosravirad, G. Berardinelli, K. I. Pedersen, and F. Frederiksen, "Enhanced HARQ design for 5G wide area technology," in *Proc. IEEE VTC Spring*, 2016.

[10] ETSI, "5G; NR; phyiscal layer procedures for data: 3GPP TS 38.214 version 17.5.0 release 17." www.etsi.org, 2023. Technical Specification.

[11] ETSI, "5G; NR; physical channels and modulation: 3GPP TS 38.211 version 17.4.0 release 17." www.etsi.org, 2023. Technical Specification.

[12] N. K. Pratas, S. Pattathil, C. Stefanovic, and P. Popovski, "Massive machine-type communication (mMTC) access with integrated authentication," in *Proc. IEEE ICC*, 2017.

[13] R. Chattamvelli and R. Shanmugam, *Discrete Distributions in Engineering and the Applied Sciences*. Springer Cham, 2020.

[14] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011.

[15] H. Kellerer, U. Pferschy, and D. Pisinger, *Knapsack Problems*. Springer, 2004.

[16] M. Hoffmann, S. Janji, A. Samorzewski, L. Kułacz, C. Adamczyk, M. Dryjański, P. Kryszkiewicz, A. Kliks, and H. Bogucka, "Open RAN xApps design and evaluation: Lessons learnt and identified challenges," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 2, 2024.

[17] ORAN ALLIANCE, "O-RAN Specifications." https://www.o-ran.org/specifications. Accessed: 2024-07-30.

[18] D. Raca, D. Leahy, C. J. Sreenan, and J. J. Quinlan, "Beyond throughput, the next generation: A 5G dataset with channel and context metrics," in *Proc. ACM MMSys*, 2020.

[19] F. Mehmeti and T. F. L. Porta, "Analyzing a 5G dataset and modeling metrics of interest," in *Proc. IEEE MSN*, 2021.

[20] N. Nikaein and R. Knopp, "OpenAirInterface (OAI) RAN Code: LAYER2/MAC/mac.h." https://gitlab.eurecom.fr/oai/openairinterface5g/blob/develop/openair2/LAYER2/MAC/mac.h. Accessed: 2023-04-17.

[21] S. M. Matinkhah and W. Shafik, "Smart grid empowered by 5G technology," in *Proc. SGC*, 2019.

[22] H. C. Leligou, T. Zahariadis, L. Sarakis, E. Tsampasis, A. Voulkidis, and T. E. Velivassaki, "Smart grid: a demanding use case for 5G technologies," in *Proc. IEEE PerCom*, 2018.

[23] S. Ayvaşık, E. Babaians, A. Papa, Y. Deshpande, A. Jano, W. Kellerer, and E. Steinbach, "Demo: Remote robot control with haptic feedback over the Munich 5G research hub testbed," in *Proc. IEEE WoWMoM*, 2023.

[24] ETSI, "5G; service requirements for the 5G system: 3GPP TS 22.261 version 17.11.0 release 17." www.etsi.org, 2022. Technical Specification.

[25] ETSI, "5G; service requirements for cyber-phyiscal control applications in vertical domains: 3GPP TS 22.104 version 17.7.0 release 17." www.etsi.org, 2022. Technical Specification.

[26] B. Finley and A. Vesselkov, "Cellular IoT traffic characterization and evolution," in *Proc. IEEE WF-IoT*, 2019.

[27] B. Goektepe, S. Fähse, L. Thiele, T. Schierl, and C. Hellge, "Subcode-based early HARQ for 5G," in *Proc. IEEE ICC WKSHPS*, 2018.

[28] ETSI, "5G; NR; multiplexing and channel coding: 3GPP TS 38.212 version 17.5.0 release 17." www.etsi.org, 2023. Technical Specification.

[29] G. Pocovi, T. Kolding, and K. I. Pedersen, "On the cost of achieving downlink ultra-reliable low-latency communications in 5G networks," *IEEE Access*, vol. 10, 2022.

[30] ETSI, "5G; NR; user equipment (UE) radio transmission and reception; part 1: Range 1 standalone: 3GPP TS 38.101-1 version 17.8.0 release 17." www.etsi.org, 2023. Technical Specification.

[31] R. Prasad and A. Sunny, "Scheduling slice requests in 5G networks," *IEEE ACM Trans. Netw.*, vol. 31, no. 6, 2023.

[32] P. Nikolaidis, A. Zoulkarni, and J. Baras, "Data-driven bandwidth adaptation for radio access network slices." https://arxiv.org/abs/2311.17347, 2023. Accessed: 2024-01-12.

[33] F. Lotfi, F. Afghah, and J. Ashdown, "Attention-based open ran slice management using deep reinforcement learning," in *Proc. IEEE GLOBECOM*, 2023.

[34] R. Challa, V. V. Zalyubovskiy, S. M. Raza, H. Choo, and A. De, "Network slice admission model: Tradeoff between monetization and rejections," *IEEE Systems Journal*, vol. 14, no. 1, 2020.

[35] Q. Qin, N. Choi, M. R. Rahman, M. Thottan, and L. Tassiulas, "Network slicing in heterogeneous software-defined RANs," in *Proc. IEEE INFOCOM*, 2020.

[36] J. Zheng, A. Banchs, and G. de Veciana, "Constrained network slicing games: Achieving service guarantees and network efficiency," *IEEE ACM Trans. Netw.*, vol. 31, no. 6, 2023.

[37] A. Gharehgoli, A. Nouruzi, N. Mokari, P. Azmi, M. R. Javan, and E. A. Jorswieck, "AI-based resource allocation in end-to-end network slicing under demand and CSI uncertainties," *IEEE Trans. Netw. Service Manag.*, vol. 20, no. 3, 2023.

[38] M. Zangooei, M. Golkarifard, M. Rouili, N. Saha, and R. Boutaba, "Flexible RAN slicing in Open RAN with constrained multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 2, 2024.

[39] Q. Li, Y. Wang, G. Sun, L. Luo, and H. Yu, "Joint demand forecasting and network slice pricing for profit maximization in network slicing," *IEEE Trans. on Netw. Sci. Eng.*, vol. 11, no. 2, 2024.

[40] B. Han, V. Sciancalepore, D. Feng, X. Costa-Perez, and H. D. Schotten, "A utility-driven multi-queue admission control solution for network slicing," in *Proc. IEEE INFOCOM*, 2019.

[41] M. Dai, G. Sun, H. Yu, and D. Niyato, "Maximize the long-term average revenue of network slice provider via admission control among heterogeneous slices," *IEEE ACM Trans. Netw.*, vol. 32, no. 1, 2024.

[42] F. Debbabi, R. Jmal, L. Chaari Fourati, R. Taktak, and R. L. Aguiar, "Adaptive admission control for 6G network slicing resource allocation (A2C-NSRA)," in *Proc. AINA*, Springer Nature Switzerland, 2024.

[43] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proc. ACM MobiCom*, 2017.