Overselling Spectrum Slices in Radio Access Networks (RANs): An Optimal Auction Design Perspective

Cariappa K S
Department of Computer Science
University of Cincinnati, OH
Email: kotranca@mail.uc.edu

Swastik Brahma
Department of Computer Science
University of Cincinnati, OH
Email: brahmask@ucmail.uc.edu

Anthony Macera
Information Directorate
Air Force Research Laboratory, Rome, NY
Email: anthony.macera.1@us.af.mil

MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM) | 979-8-3503-7423-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/MILCOM61039.2024.10773777

***Abstract*—This paper presents the design of a novel optimal auction mechanism that enables a Mobile Network Operator (MNO) to intelligently oversell slices of its spectrum to a carefully chosen set of Wireless Tenants (WTs). The paper uses Queueing theory to model and analyze the Quality of Service (QoS) experienced by WTs under overselling, and integrates the theory with auction theory to ensure that the developed mechanism duly satisfies WTs' QoS requirements. It is shown that, under inherent uncertainties associated with communication processes, our overselling methodology enhances important performance criteria, such as the MNO's utility and spectrum utilization, beyond what is permitted by traditional approaches that do not allow overselling. Insights into the number of frequencies that an MNO should slice its spectrum into have been provided. Numerous simulation results are presented which show the performance advantages of our proposed mechanism.**

***Index Terms*—Slicing, Overselling, Auctions, Queueing Theory.**

I. Introduction

The advancement of Network Function Virtualization (NFV) and Software Defined Networking (SDN) techniques have enabled the concept of *Radio Access Network (RAN) slicing* [1], [2]. RAN slicing enables a Mobile Network Operator (MNO) to intelligently divide its physical radio infrastructure into multiple logical networks or *slices*, and then *reserve* the slices for different Wireless Tenants (WTs) in a goal-driven business-oriented manner [3]. Such reservation of infrastructural slices helps to support myriad applications over a shared radio infrastructure while being able to meet their desired Quality of Service (QoS) requirements. RAN slicing is a key technology for meeting the demands of next generation wireless applications in diverse sectors [2], [4], including tactical battlefield networks, Internet of Things (IoT), connected vehicles, and healthcare. RAN slicing, however, is still in its infancy, with various unresolved challenges in the area.

While reservation of slices can enable Service Level Agreements (SLA) between the MNO and WTs to be established and met, a novel question that arises is: *what if some WTs eventually do not need to transmit data as planned using their reserved slices?* Such events, which are inherently hard to predict, can lead to under-exploitation of the radio infrastructure and greatly sacrifice resource utilization– a problem that has remained grossly underexplored. For example, [1], [2] explores algorithmic aspects of slice creation and management, [5] studies power allocation schemes for communication performance optimization in a slicing context, [6] employs machine learning techniques for slicing, and [3] employs game theory [7] to study competitive slicing strategies. However, prior work has overlooked the inherent uncertainties in WT's traffic, that can make them to not always have data to transmit using their reserved slices. While [8], [9] have taken a preliminary perspective into the topic, the aforementioned problem remains largely unaddressed and its mitigation approaches grossly undertheorized. We aim to fill this void in this paper.

Specifically, in this paper, to mitigate the aforementioned problem, which stems from the inherent uncertainties associated with traffic characteristics of WTs, we propose to allow the MNO to intelligently *oversell (i.e., overbook)*[1] slices of its usable electromagnetic spectrum. In particular, considering an orthogonal frequency-division multiplexing (OFDM) system, where the MNO slices its spectrum into multiple orthogonal frequencies [5], we build on the field of mechanism design [7], [10] to present the design of a novel auction mechanism that allows the MNO to optimally *oversell* the slices (frequencies), i.e., use them to serve more WTs than the number of available slices. The novel contributions of the paper are as follows:

- We present the design of a novel optimal auction mechanism that allows an MNO to optimally 'oversell' slices of its spectrum to a carefully selected set of WTs depending on their bidding behaviors and traffic characteristics. Among others, our mechanism can enforce desirable properties such as truthfulness of bidding strategies.
- Modeling the traffic characteristics of WTs as a random process, we employ Queueing theory [11] to characterize the QoS experienced by WTs when overbooking is allowed. Further, we integrate the theory with mechanism design to ensure that our developed overselling-enabled auction methodology can satisfy WTs' QoS requirements.
- We provide insights into the number of frequencies that the MNO's spectrum should be sliced into given a set of WTs that are contending to be served by the MNO.
- We present computationally efficient techniques to implement our designed auction mechanism.
- We present numerous simulations which demonstrate the performance advantages of our proposed mechanism.

This work was supported by the U.S. National Science Foundation (NSF) under Award Number CCF-2302197 and by the University of Cincinnati (UC).

[1]We use the terms 'overselling' and 'overbooking' synonymously.

II. Formulation of the Auction Design Problem

Consider an MNO that has sliced its range of usable spectrum into M frequencies, and is conducting an auction to determine the set of WTs that it should serve using its M frequencies over a time period T. Consider collisions (packet drops) to occur if multiple WTs transmit simultaneously on the same frequency, making the MNO's frequency slices to become *overbooked* if the MNO decides to serve more than M WTs. In such a scenario, consider that there are N WTs, numbered 1 to N, participating in the auction with WT i, $i \in \{1, \cdots, N\}$, having a *true* valuation of v_i per unit of data that it successfully transmits over the time period T. To model the MNO's uncertainty regarding the true valuation of WT i, $i \in \{1, \cdots, N\}$, consider it to be a random variable with $f_i : [a_i, b_i] \rightarrow \mathbb{R}_+$ being its probability density function (PDF) and $F_i : [a_i, b_i] \rightarrow [0, 1]$ being its cumulative distribution function (CDF). Here, a_i and b_i are the lowest and highest possible valuations of WT i, respectively. Consider that the packet generation characteristics of every WT over the time period T follows a Poisson process with rate λ, and that the transmission delay of its packets using any one of the M frequencies is exponentially distributed with a mean $1/\mu$.

To participate in the auction, suppose that every WT i, $i \in \{1, \cdots, N\}$, sends its valuation (potentially in a falsified form) as its bid to the MNO, with $\mathbf{v} = (v_1, \cdots, v_N)$ denoting the vector of bids that the MNO receives. In such a scenario, the MNO's auction mechanism can be described by two functions:

- *WT selection function*, $\mathbf{p}(\mathbf{v}) = \big(p_1(\mathbf{v}), \cdots, p_N(\mathbf{v})\big)$, where $p_i(\mathbf{v}) \in \{0, 1\}$ is a Boolean variable, with $p_i(\mathbf{v}) = 1$ denoting that the MNO decides to serve WT i over the time period T (and $p_i(\mathbf{v}) = 0$ indicating otherwise); and
- *Payment function*, $\mathbf{x}(\mathbf{v}) = \big(x_1(\mathbf{v}), \cdots, x_N(\mathbf{v})\big)$, where $x_i(\mathbf{v})$ is the payment that WT i makes to the MNO.

A. Expected Utilities of the MNO and WTs

The MNO's utility from the above auction mechanism is

$$U^{MNO}(\mathbf{p}, \mathbf{x}) = \int_V \sum_{i=1}^{N} x_i(\mathbf{v}) f(\mathbf{v}) d\mathbf{v} \tag{1}$$

where $V = [a_1, b_1] \times \cdots \times [a_N, b_N]$ denotes the set of all possible combinations of WTs' valuations, $f(\mathbf{v}) = \prod_{i=1}^{N} f_i(v_i)$ is the joint density function on V for the vector of valuations $\mathbf{v} = (v_1, \cdots, v_N)$, and $d\mathbf{v} = dv_1 \cdots dv_N$. Further, the expected utility that WT i, $i \in \{1, \cdots, N\}$, gets from the auction from bidding its *true* valuation $v_i \in [a_i, b_i]$ is

$$U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) = \int_{V_{-i}} \big(p_i(\mathbf{v}) v_i h\big(n(\mathbf{v})\big) \frac{TR}{\mu} - x_i(\mathbf{v})\big) f_{-i}(\mathbf{v}_{-i}) d\mathbf{v}_{-i} \tag{2}$$

Note that in (2):

- $n(\mathbf{v}) = \sum_{i=1}^{N} p_i(\mathbf{v})$ denotes the number of WTs that the MNO decides to serve for a given $\mathbf{v}$,
- given $n(\mathbf{v})$, $h\big(n(\mathbf{v})\big)$ denotes the expected successful packet transmission rate of a WT served by the MNO,
- R denotes the capacity of each frequency (in terms of the number of data units that a frequency can transmit per unit time), which makes $h\big(n(\mathbf{v})\big)\frac{TR}{\mu}$ the expected amount of data that a WT served by the MNO transmits over the time period T,
- $V_{-i} = [a_1, b_1] \times \cdots \times [a_{i-1}, b_{i-1}] \times [a_{i+1}, b_{i+1}] \times \cdots \times [a_N, b_N]$ denotes the set of all possible combinations of WTs' valuations other than WT i, $f_{-i}(\mathbf{v}_{-i}) = \prod_{j \in \{1, \cdots, N\}, j \neq i} f_j(v_j)$ denotes the joint density function on V_{-i} for the vector of valuations $\mathbf{v}_{-i} = (v_1, \cdots, v_{i-1}, v_{i+1}, \cdots, v_N)$, and $d\mathbf{v}_{-i} = dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_N$.

Moreover, given that WT i's *true* valuation is $v_i \in [a_i, b_i]$, the expected utility that WT i gets from bidding a *falsified* valuation $w_i \in [a_i, b_i]$, hoping to make an undue profit, is

$$\tilde{U}_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) = \int_{V_{-i}} \big(p_i(w_i, \mathbf{v}_{-i}) v_i h\big(n(w_i, \mathbf{v}_{-i})\big) \frac{TR}{\mu} - x_i(w_i, \mathbf{v}_{-i})\big) f_{-i}(\mathbf{v}_{-i}) d\mathbf{v}_{-i} \tag{3}$$

where $(w_i, \mathbf{v}_{-i}) = (v_1, \cdots, v_{i-1}, w_i, v_{i+1}, \cdots, v_N)$.

In such a scenario, in this paper, we aim to *design the functions* $\mathbf{p}(\mathbf{v})$ *and* $\mathbf{x}(\mathbf{v})$ *such that they allow the MNO to* ***optimize*** *its revenue by* ***overselling*** *the M frequencies while satisfying certain constraints.*

B. Auction Design as an Optimization Problem

Design of the optimal auction mechanism for the MNO can be formulated as the following optimization problem:

$$\max_{\mathbf{p}, \mathbf{x}} \; U^{MNO}(\mathbf{p}, \mathbf{x})$$

subject to:

$$U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) \geq 0, \; \forall i \in \{1, \cdots, N\} \tag{4a}$$

$$U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) \geq \tilde{U}_i^{WT}(\mathbf{p}, \mathbf{x}, w_i), \; \forall i, \forall v_i, w_i \in [a_i, b_i] \tag{4b}$$

$$h(n(\mathbf{v})) \geq \psi, \forall \mathbf{v} \in V \tag{4c}$$

$$p_i(\mathbf{v}) \in \{0, 1\}, \; \forall i \; \in \{1, \cdots, N\}, \forall \mathbf{v} \in V \tag{4d}$$

The above four constraints are explained below:

- *Individual-Rationality (IR) constraint* (4a) justifies participation of the WTs in the auction by ensuring that their expected utilities are non-negative.
- *Incentive-Compatibility (IC) constraint* (4b) disincentivizes WTs from lying about their valuations during bidding by ensuring that honest reporting of valuations form a Nash Equilibrium (NE).
- *QoS constraint* (4c) ensures that the expected packet transmission rate ($h(.)$) of every WT that the MNO serves satisfies a prescribed threshold ψ.
- *Selection Parameter constraint* (4d) ensures that every WT's selection parameter is Boolean. *Note, there is no constraint that restricts how many WTs the MNO decides to serve to allow overbooking of the M frequency slices.*

III. Analysis of the Auction Design Problem

For a given bid v_i, the expected amount of data that WT i, $i \in \{1, \cdots, N\}$, transmits over the time period T is

$$Q_i(\mathbf{p}, v_i) = \int_{V_{-i}} p_i(v_i, \mathbf{v}_{-i}) h\big(n(v_i, \mathbf{v}_{-i})\big) \frac{TR}{\mu} f_{-i}(\mathbf{v}_{-i}) d\mathbf{v}_{-i} \tag{5}$$

Using (5), we first present a simplified characterization of (4b).

Lemma 1. *The IC constraint in* (4b) *holds if the following two conditions hold $\forall i \in \{1, \cdots, N\}$:*

$$if \; v_i \geq w_i, \;\; then \;\; Q_i(\mathbf{p}, v_i) \geq Q_i(\mathbf{p}, w_i) \tag{6a}$$

$$\int_{a_i}^{v_i} Q_i(\mathbf{p}, w_i) dw_i = U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) - U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) \tag{6b}$$

Proof. Consider $v_i, w_i \in [a_i, b_i]$ with $v_i \geq w_i$, $i \in \{1, \cdots, N\}$. Now, if v_i is WT i's true valuation per unit data that it successfully transmits while it bids the falsified valuation w_i, the expected utility that the WT gets can be expressed using (2), (3) and (5) as

$$\tilde{U}_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) = U_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) + (v_i - w_i)\, Q_i(\mathbf{p}, w_i) \quad (7)$$

To ensure that WT i does not have an incentive to bid such a falsified valuation w_i, imposing the IC constraint (4b), we get

$$U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) \geq U_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) + (v_i - w_i)\, Q_i(\mathbf{p}, w_i) \quad (8)$$

Similarly, considering WT i's true valuation to be w_i while it bids the falsified valuation v_i, the IC constraint implies that

$$U_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) \geq U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) + (w_i - v_i)\, Q_i(\mathbf{p}, v_i) \quad (9)$$

From (8) and (9), we get

$$(v_i - w_i)\, Q_i(\mathbf{p}, w_i) \leq U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) - U_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) \leq (v_i - w_i)\, Q_i(\mathbf{p}, v_i) \quad (10)$$

Note that (6a) is clearly implied by (10). Further, letting $v_i = w_i + \delta$, the above inequality can be rewritten as:

$$\delta\, Q_i(\mathbf{p}, w_i) \leq U_i^{WT}(\mathbf{p}, \mathbf{x}, w_i + \delta) - U_i^{WT}(\mathbf{p}, \mathbf{x}, w_i) \leq \delta\, Q_i(\mathbf{p}, w_i + \delta) \quad (11)$$

Clearly, (11) implies that $Q_i(\mathbf{p}, w_i)$ is Riemann-integrable, from which (6b) follows, concluding the proof. □

Using Lemma 1, we can simplify the optimization problem in (4) to the form given in the following theorem.

THEOREM 1. *For $(\mathbf{p}, \mathbf{x})$ to represent the MNO's optimal auction mechanism, $\mathbf{p}$ should be such that it maximizes*

$$\int_V \sum_{i=1}^{N} \left(\left(v_i - \frac{1 - F_i(v_i)}{f_i(v_i)} \right) p_i(\mathbf{v}) h\big(n(\mathbf{v})\big)\, \frac{TR}{\mu} \right) f(\mathbf{v})\, d\mathbf{v} \quad (12)$$

subject to constraints (4c) *and* (4d)*, and the payment made by WT i, $i \in \{1, \cdots, N\}$, should follow*

$$x_i(\mathbf{v}) = p_i(\mathbf{v}) v_i h\big(n(\mathbf{v})\big) \frac{TR}{\mu} - \int_{a_i}^{v_i} p_i(w_i, \mathbf{v}_{-i}) h\big(n(\mathbf{v})\big) \frac{TR}{\mu} dw_i \quad (13)$$

Proof. The MNO's expected utility (1) can be re-written as:

$$U^{MNO}(\mathbf{p}, \mathbf{x}) = \int_V \sum_{i=1}^{N} \left(x_i(\mathbf{v}) - v_i p_i(\mathbf{v}) h\big(n(\mathbf{v})\big) \frac{TR}{\mu} \right) f(\mathbf{v}) d\mathbf{v} + \int_V \sum_{i=1}^{N} v_i\, p_i(\mathbf{v})\, h\big(n(\mathbf{v})\big)\, \frac{TR}{\mu} f(\mathbf{v})\, d\mathbf{v} = -\sum_{i=1}^{N} \int_{a_i}^{b_i} U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) f_i(v_i)\, dv_i + \sum_{i=1}^{N} \int_V v_i\, p_i(\mathbf{v})\, h\big(n(\mathbf{v})\big)\, \frac{TR}{\mu} f(\mathbf{v})\, d\mathbf{v} \big(\text{using (2)}\big) \quad (14)$$

Now, using (6b), we have

$$\begin{aligned} \int_{a_i}^{b_i} U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) f_i(v_i)\, dv_i &= \int_{a_i}^{b_i} U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) f_i(v_i)\, dv_i \\ &+ \int_{a_i}^{b_i} \int_{a_i}^{v_i} Q_i(\mathbf{p}, w_i)\, f_i(v_i)\, dw_i\, dv_i = U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) \\ &+ \int_{a_i}^{b_i} \int_{w_i}^{b_i} f_i(v_i)\, Q_i(\mathbf{p}, w_i)\, dv_i\, dw_i = U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) \\ &+ \int_V \big(1 - F_i(v_i)\big) p_i(\mathbf{v}) h\big(n(\mathbf{v})\big) \frac{TR}{\mu} f_{-i}(\mathbf{v}_{-i}) d\mathbf{v} \end{aligned} \quad (15)$$

Substituting (15) into (14), we get:

$$U^{MNO}(\mathbf{p}, \mathbf{x}) = \int_V \sum_{i=1}^{N} \left(v_i - \frac{1 - F_i(v_i)}{f_i(v_i)} \right) p_i(\mathbf{v})\, h\big(n(\mathbf{v})\big)\, \frac{TR}{\mu} f(\mathbf{v})\, d\mathbf{v} - \sum_{i=1}^{N} U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) \quad (16)$$

In (16), $\mathbf{x}$ only appears in the last term of the MNO's utility. Now, from (6b), note that for WT i, $i \in \{1, \cdots, N\}$, if $U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) \geq 0$, we get $U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) \geq 0$, $\forall v_i \in [a_i, b_i]$, which leads to the satisfaction of the IR constraint (4a). Thus, setting $U_i^{WT}(\mathbf{p}, \mathbf{x}, a_i) = 0$ in (16), $\forall i \in \{1, \cdots, N\}$, is optimal as it would both maximize the MNO's utility as well as satisfy the IR constraint, which implies, using (6b), that

$$U_i^{WT}(\mathbf{p}, \mathbf{x}, v_i) - \int_{a_i}^{v_i} Q_i(\mathbf{p}, w_i) dw_i = 0 \quad (17)$$

Substituting (2) and (5) into (17), we get (13), with the MNO's utility thereby becoming (12). This proves the theorem. □

A. Determination of Auction Outcomes

We now present how to find Theorem 1's prescribed set of WTs that the MNO should serve and their payments.

1) Optimal Selection of WTs

For WT i, $i \in \{1, \cdots, N\}$, let us define

$$\theta_i(v_i) = v_i - \frac{1 - F_i(v_i)}{f_i(v_i)} \quad (18)$$

We refer to $\theta_i(v_i)$ as the *virtual valuation* of WT i. Note that the MNO's expected utility in the optimal auction mechanism (12) is maximized if $\mathbf{p}(\mathbf{v})$ is such that it maximizes

$$\frac{TR}{\mu} h\big(n(\mathbf{v})\big) \sum_{i=1}^{N} \theta_i(v_i)\, p_i(\mathbf{v}) \quad (19)$$

for all $\mathbf{v} \in V$ (subject to constraints (4c) and (4d)). To explore maximization of (19), suppose that the MNO, after receiving the vector of bids $\mathbf{v}$ from N WTs, labels the WTs in non-increasing order of their virtual valuations $\big($such that $\theta_1(v_1) \geq \theta_2(v_2) \geq \cdots \geq \theta_N(v_N)\big)$. For notational simplicity, suppose that we drop the argument, viz. $\mathbf{v}$, of $n(\cdot)$ and $p_i(\cdot)$. In such a scenario, for a given n, note that the *maximum* value (say denoted as $\theta^{(n)}$) of the term $\sum_{i=1}^{N} \theta_i(v_i)\, p_i$ in (19) is attained when $p_i = 1$ ($\forall i \in \{1, \cdots, n\}$) and $p_i = 0$ ($\forall i \in \{n+1, \cdots, N\}$), leading to $\theta^{(n)} = \theta_1(v_1) + \cdots + \theta_n(v_n)$.

Thus, to find optimal $\mathbf{p}$ that maximizes (19) subject to (4c) and (4d), the MNO would have to find n^* such that

$$n^* = \arg\max_n \frac{TR}{\mu} h\big(n\big) \theta^{(n)} \quad (20a)$$

$$\text{s.t.} \quad 0 \leq n \leq N \text{ and } h(n) \geq \psi \quad (20b)$$

with optimal $\mathbf{p}^*$ subsequently being $p_i^* = 1$ ($\forall i \in \{1, \cdots, n^*\}$) and $p_i^* = 0$ ($\forall i \in \{n^* + 1, \cdots, N\}$). Note, the above process of finding optimal $\mathbf{p}$ is computationally efficient since it only involves sorting WTs' virtual valuations, and inspecting the $N+1$ values that n can assume (to solve (20)).

2) Determination of WTs' payments

The payment that a WT should make can be found using (13). Note, a WT that the MNO does not serve, does not make any payment, since, for advertised bid $v_i \in [a_i, b_i]$ of WT i, if $p_i(v_i, \mathbf{v}_{-i}) = 0$, we have: *a*) the first term of (13) is 0; and *b*) the second term of (13) is also 0 (over bids in $[a_i, v_i]$) since its integrand, which is clearly non-negative, is a monotonically increasing function of the bid of WT i as follows from (6a). Next, to find the payment of WT i that the MNO serves,

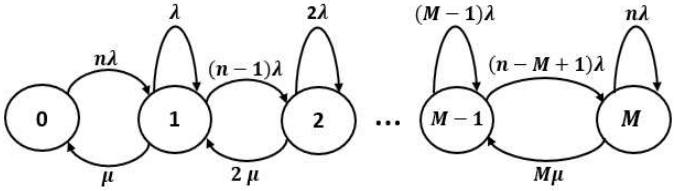


Fig. 1. Markov chain with M frequency slices under n WTs being served.

note that the integrand in (13), viz. $p_i(w_i, \mathbf{v}_{-i})h\big(n(\mathbf{v})\big)\frac{TR}{\mu}$, is a monotonically increasing *step* function of WT i's bid w_i. This is because, as follows from (6a), the integrand in (13) is a monotonically increasing function of bid w_i of WT i, which when combined with the fact that $p_i(\mathbf{v})$ is Boolean (making (13)'s integrand behave as a step function), explains the aforementioned nature of the integrand in (13). Exploiting this nature, one can employ techniques such as the Bisection method to iteratively identify points of discontinuities of (13)'s integrand over $[a_i, v_i]$, progressively compute the area under the integrand function between consecutive points of discontinuities (to compute the integration over consecutive discontinuity points), while adding the found areas to compute the second term of (13). The computed second term can then be subtracted from the first term of (13) (which is clearly straightforward to compute) to find the needed payment.

IV. Queueing Theoretic Analysis of Steady States

As noticable from Theorem 1, determination of WTs that the MNO should serve and their payments depend on the expected QoS that would be experienced by them (i.e., $h(.)$). Thus, using Queueing theory [11], considering the MNO to serve n users, we now characterize $h(\cdot)$ and its properties.

To do so, consider an on-demand frequency allocation scheme where a WT being served by the MNO is assigned a frequency (e.g., by the Base Station (BS) in the MNO's infrastructure) to transmit when the WT has a data packet to send (which is relinquished by the WT when the transmission ends with the WT being reassigned a frequency when it has another data packet to send). Such on-demand frequency allocation, e.g., can correspond to the *grant-based* scheduling methodology of 5G, where data packet transmission requests of devices are dynamically mapped to frequencies [12]. For expositional simplicity, consider that the WTs do not have capabilities to buffer any packets (implying that packets generated by a WT when all frequencies are busy, or when it is in the process of transmitting one, are dropped). As mentioned earlier, consider that the data packet generation characteristics of every WT over the time period T follows a Poisson process with rate λ and that the transmission delay of its packets using any one of the available frequencies is exponentially distributed with a mean $1/\mu$.

A. Markov Chain and Steady State Analysis

The Markov chain depicting the state transitions of the MNO's system, considering that the MNO is serving n WTs and that its usable spectrum has been sliced into M frequencies, is shown in Fig. 1. Specifically, the MNO's system being in state m, $m \in \{0, 1, \cdots, M\}$, reflects that m frequencies are being simultaneously used at a point in time (by m different WTs) within the time period T. The self-loops in the figure reflect *packet drop rates* from the different states. In particular, note that, when the system is in state m, $m \in \{0, 1, \cdots, M-1\}$, the m transmitting WTs will aggregately generate packets at the rate $m\lambda$, all of which will be dropped due to their lack of buffering capabilities (as denoted by state m's self-loop). In such a scenario, the remaining $n-m$ WTs, $m \in \{0, 1, \cdots, M-1\}$, will aggregately generate packets at the rate $(n-m)\lambda$, resulting in the MNO's system to transition to state $m+1$ from state m at the rate $(n-m)\lambda$, as shown in the figure. However, when the MNO's system is in state M, denoting that M WTs are transmitting simultaneously, the n WTs being served by the MNO will aggregately generate packets at the rate $n\lambda$, all of which will be dropped (as denoted by the self-loop from state M). Further, since the transmission delay of a WT's packet is exponentially distributed with a mean $1/\mu$, the system will transition to state $m-1$ from state m at the rate $m\mu$, $m \in \{1, 2, \cdots, M\}$.

1) Steady-State Probabilities

Let π_m be the probability of the MNO's system being in state m at a random point in time within the time period T, $m \in \{0, 1, \cdots, M\}$, and let $\rho = \frac{\lambda}{\mu}$. The next theorem characterizes the steady-state values of the probabilities.

Theorem 2. *The steady-state probabilities of the Markov chain depicted in Fig. 1 are* $\pi_0 = \frac{1}{\sum_{i=0}^{M}\binom{n}{i}\rho^i}$ *and* $\pi_m = \binom{n}{m}\rho^m\pi_0,\ \forall m = \{1, 2, \cdots, M\}$

Proof. For the Markov chain in Fig. 1 to be stable, the balance equations imply that:

$$\pi_0 n\lambda = \pi_1\mu \Rightarrow \pi_1 = \binom{n}{1}\rho\pi_0 \tag{21a}$$

$$\pi_1(n-1)\lambda = \pi_2\ (2\mu) \Rightarrow \pi_2 = \binom{n}{2}\rho^2\pi_0 \tag{21b}$$

$$\cdots$$

$$\pi_{M-1}(n-M+1)\lambda = \pi_M(M\mu) \Rightarrow \pi_M = \binom{n}{M}\rho^M\pi_0 \tag{21c}$$

Further, we have:

$$\pi_0 + \pi_1 + \cdots + \pi_{M-1} + \pi_M = 1 \tag{22}$$

Substituting (21) into (22), and simplifying, yields $\pi_0 = \frac{1}{\sum_{i=0}^{M}\binom{n}{i}\rho^i}$, with $\pi_m = \binom{n}{m}\rho^m\pi_0$ following from (21), $\forall m = \{1, 2, \cdots, M\}$. This proves the theorem. □

B. Characterization and Analysis of $h(\cdot)$

First, we characterize $h(\cdot)$ using the Markov chain in Fig. 1.

Remark 1. *Given that the MNO is serving n WTs while having M frequency slices, the expected packet transmission rate of each WT, viz. $h(n)$, that is being served is:*

$$h(n) = \frac{1}{n}\big[n\lambda - \big\{\sum_{m=1}^{M-1}\pi_m m\lambda + \pi_M n\lambda\big\}\big] \tag{23}$$

Remark 1 follows from the fact that, as can be noted from Fig. 1, the total expected packet drop rate is $\sum_{m=1}^{M-1}\pi_m m\lambda + \pi_M n\lambda$. This implies, noting that the total aggregate packet generation rate of the n WTs that the MNO is serving is $n\lambda$, the total expected packet transmission rate of the n WTs is $n\lambda - \big\{\sum_{m=1}^{M-1}\pi_m m\lambda + \pi_M n\lambda\big\}$, which divided by n yields the expected packet transmission rate of each WT.

Next, we study $h(\cdot)$'s behavior as the number of WTs that the MNO serves increases for a given number of frequency slices. To do so, let us use $h(n;M)$ to denote the expected packet transmission rate of a WT when the MNO serves n WTs while having sliced its spectrum into M frequencies.

LEMMA 2. *Given M frequencies, $h(n=M;M) > h(n=M+1;M)$.*

Proof. Considering M frequency slices, when the MNO serves M WTs, i.e., $n=M$, using (23), the expected packet transmission rate of each WT is $h(n=M;M)=\frac{1}{M}(M\lambda - \sum_{m=1}^{M}\pi_m m\lambda)$, where π_m is the steady-state probability of the MNO's system being in state m, $m\in\{0,\cdots,M\}$. Using Th. 2, substituting $\pi_m=\binom{M}{m}\rho^m\pi_0$, where it can be shown that $\pi_0=\frac{1}{(1+\rho)^M}$, into $h(n=M;M)$, $m=\{1,\cdots,M\}$, and subsequently simplifying using the binomial theorem, we get

$$h(n=M;M)=\frac{\lambda}{1+\rho} \tag{24}$$

Again, given M frequencies, when the MNO serves $M+1$ WTs, i.e, $n=M+1$, using (23), the expected packet transmission rate of each WT is $h(n=M+1;M)=\frac{1}{M+1}\Big[(M+1)\lambda-\big\{\sum_{m=1}^{M-1}\pi'_m m\lambda+\pi'_M(M+1)\lambda\big\}\Big]$, where π'_m is the steady-state probability of the MNO's system being in state m, $m\in\{0,\cdots,M\}$. Using Th. 2, substituting $\pi'_m=\binom{M+1}{m}\rho^m\pi'_0$, where it can be shown that $\pi'_0=\frac{1}{(1+\rho)^{M+1}-\rho^{M+1}}$, into $h(n=M+1;M)$, $m\in\{1,\cdots,M\}$, and subsequently simplifying using the binomial th., we get

$$h(n=M+1;M)=\lambda-\pi'_0\lambda[\rho(1+\rho)^M+\rho^{M+1}-\rho^M] \tag{25}$$

Subtracting (24) from (25), while using the above characterization of π'_0, we get $h(n=M+1;M)-h(n=M;M)$

$$=\frac{\lambda\rho}{1+\rho}\left[1-\frac{(1+\rho)^{M+1}-\rho^{M+1}+\rho^{M-1}}{(1+\rho)^{M+1}-\rho^{M+1}}\right] \tag{26}$$

which is clearly negative since $(1+\rho)^{M+1}-\rho^{M+1}+\rho^{M-1} > (1+\rho)^{M+1}-\rho^{M+1}$. This completes the proof. □

Note that, given M frequency slices, unlike Lemma 2, analytically studying the behavior of $h(\cdot)$ as the number of WTs that the MNO serves increases beyond $M+1$ becomes mathematically intractable. This is because such analysis encounters a partial binomial sequence (finding a closed-form sum of which remains an unsolved problem). However, we have conducted extensive simulations which have shown that, for a given M, $h(\cdot)$ shows a non-increasing trend as the number of WTs that the MNO serves increases.

Specifically, Fig. 2 shows how $h(n;M)$ varies with M (no. of frequencies) and n (no. of served WTs) for two scenarios, viz. $\lambda=100$ pkts/sec with $\mu=400$ pkts/sec, and $\lambda=100$ pkts/sec with $\mu=200$ pkts/sec. As can be seen, for both scenarios, $h(\cdot)$ is a constant whenever $n\le M$. This is because, when $n\le M$, only n frequencies are needed to serve the n WTs, making $h(\cdot)$ to be characterizable by (24), which is independent of n and M (making it a constant for a given λ and μ). To corroborate this, e.g., for $n\le M$ under $\lambda=100$ pkts/sec and $\mu=400$ pkts/sec, $h(\cdot)$ obtained from (24) is $\frac{100}{1+0.25}=80$, which tallies with the figure. Moreover, as can be seen, for both values of λ and μ in the figure, for any given M, $h(\cdot)$ decreases when n becomes larger than M, corroborating

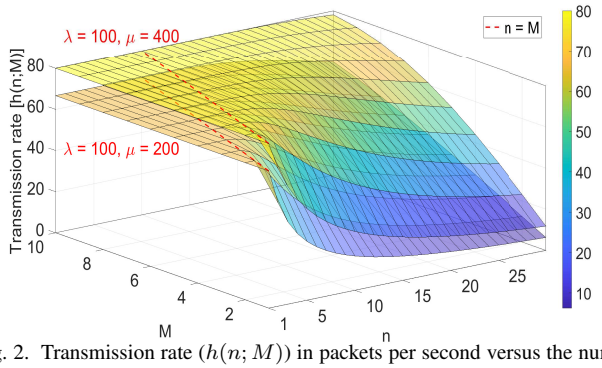


Fig. 2. Transmission rate ($h(n;M)$) in packets per second versus the number of frequencies (M) and the number of served (n) of the total WTs (N).

Lemma 2 as well as showing the general decreasing behavior of $h(\cdot)$ as the degree of overbooking increases.

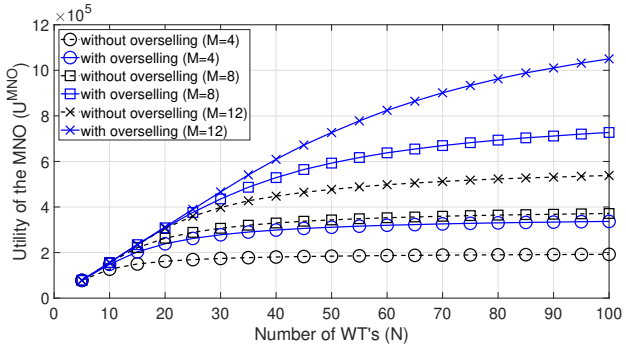
REMARK 2. *Note that, for a given M, Lemma 2 and Fig. 2 imply that if $h(n;M)<\psi$ (i.e., the QoS constraint* (4c) *is dissatisfied), then $h(n+k;M)<\psi$, $\forall k\in\{0,\cdots,N-n\}$. This can be used to enhance the computational efficiency of solving* (20) *by exploring whether the constraint $h(n;M)\ge\psi$ is satisfied in increasing order of n, $0\le n\le N$, and terminating the exploration to return $n^*=\underset{n\in\{0,\cdots,N'\}}{\arg\max}\ \frac{TR}{\mu}h(n)\theta^{(n)}$, where $N'(\le N)$ is the highest value of n for which $h(n;M)\ge\psi$.*

V. SIMULATION RESULTS

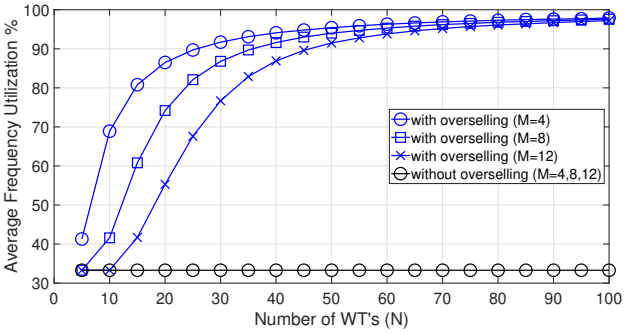
In this section, we provide simulation results to show the performance advantages of our developed auction mechanism. Fig. 3 studies how the performance of our mechanism scales with the number of WTs (N) that participate in the bidding process, with $T=60$ secs, $R=50$ Megabits per second (Mbps), $\lambda=200$ pkts/sec, $\mu=400$ pkts/sec, the valuation of each WT for successfully transmitting an Mb of data being uniformly distributed over $[10,50]$, and $\psi=120$ pkts/sec in the QoS constraint in (4c). Note that, 'with overselling' in the figures refers to the operating point prescribed by Th. 1, while 'without overselling' refers to sol. of (4) with the constraint $\sum_{i=1}^{N}p_i(\mathbf{v})\le M$ added (which can be found using a similar procedure as described in Sec. III-A1 with the constraint $0\le n\le N$ replaced with $0\le n\le M$).

As can be seen from Fig. 3(a), for any M, the MNO obtains an enhanced utility 'with overselling' over what is obtained 'without overselling' as N varies. Again, from Fig. 3(b), note that the *average frequency utilization* (average percentage of time over the time period T each frequency is used) 'with overselling' is higher than that of the 'without overselling' case. *These observations clearly show the performance advantages of our proposed overbooking methodology.* In fact, it can be noted from Fig. 3(b) that the average frequency utilization remains constant with increase of N in the 'without overselling' case, emphasizing the inability of such an approach to duly exploit increase in demand to enhance spectrum utilization.

Further, as can be seen from Fig. 3(a), for any given M, the MNO's utility 'with overselling' increases at a faster rate with N than the rate at which it increases 'without overselling'. This is because, with increasing N, while increase in the MNO's utility 'without overselling' can only be attributed to



(a) Utility of the MNO vs. the number of WTs (N)



(b) Frequency utilization vs. the number of WTs (N)

Fig. 3. Performance of our mechanism vs. the number of WTs (N)

the enhanced ability to find WTs that are willing to make higher payments for transmitting their data, in the 'with overselling' case, not only does the MNO's utility benefit from the above advantage as N increases, but also from the ability to oversell its frequency slices to more WTs (while satisfying the QoS constraint). Implications of this trend, which allows a higher degree of overselling as N increases, results in the average frequency utilization to increase with N for the 'with overselling' case, as seen in Fig. 3(b).

In Fig. 4, given that the MNO has a certain range of usable spectrum, we study how the MNO's utility scales with the number of frequency slices (M) that it uses, both for the 'with overselling' and 'without overselling' cases. Note that, the transmission capacity (R) of each frequency slice, which impacts the transmission delay ($1/\mu$) of a packet, depends on M. Specifically, considering that the MNO has B MHz of usable spectrum which is sliced into M frequencies, we consider $R = (B/M)\log(1 + SNR)$ (implying that the capacity of each frequency slice decreases with M). Using such a modeling, Fig. 4 plots the MNO's utility with varying M, considering $N = 15$, $B = 20$ MHz, $T = 60$ secs, $SNR = 31$ dB, $\lambda = 300$ pkts/sec, the valuation of each WT for successfully transmitting each Mb of data being uniformly distributed over $[10, 50]$, $\psi = 150$ pkts/sec, and with $1/\mu$ set to 10 Kb/R (considering a packet size of 10 Kb and R calculated for each M as mentioned above). As expected, the optimal value of M for the with and without overselling cases differ. Moreover, as can be seen, *the MNO obtains an enhanced utility when slicing its usable spectrum into fewer frequencies (and overselling them), taking advantage of the higher capacity of each frequency slice, than when using more frequency slices*

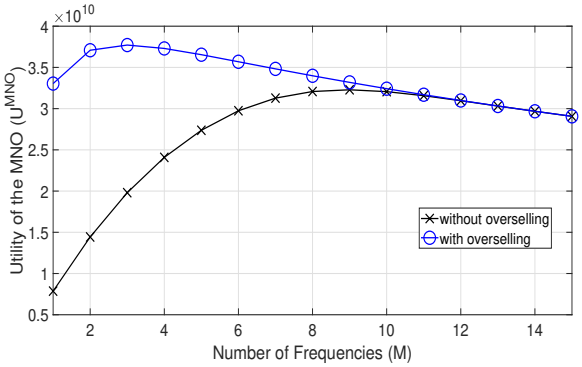


Fig. 4. MNO's utility vs. M for a given range of spectrum

(and allowing only as many WTs as the number of slices). This again shows the merit of overselling.

VI. Conclusion

This paper presented the design of a novel optimal auction mechanism that allows an MNO to intelligently oversell slices of its spectrum to a set of WTs based on their advertised bids and communication characteristics. Queueing theoretic analyses have been presented that characterize the QoS experienced by WTs under overselling and such results have been integrated with the presented auction model to ensure that WTs' QoS requirements are met. It has been shown that, under inherent uncertainties associated with communication processes, our designed overselling methodology enhances the MNO's revenue and the associated spectrum utilization beyond what is permitted by schemes where overselling in not permitted. Insights have been provided into the number of frequencies that the MNO should slice its spectrum into. Several simulation results have been presented that show the performance advantages of the proposed methodology.

References

[1] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: Ran slicing for a flexible and cost-effective multi-service mobile network architecture," in *ACM MobiCom '17*, p. 127–140.

[2] Y. Chen, R. Yao, H. Hassanieh, and R. Mittal, "Channel-Aware 5g RAN slicing with customizable schedulers," in *NSDI'23*, pp. 1767–1782.

[3] M. Srinivasan and C. S. R. Murthy, "Efficient spectrum slicing in 5g networks: An overlapping coalition formation approach," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1299–1316, 2020.

[4] A. Castañares, D. K. Tosh, and C. A. Kamhoua, "Slice aware framework for intelligent and reconfigurable battlefield networks," in *MILCOM'21*.

[5] F. Saggese, M. Moretti, and P. Popovski, "Power minimization of downlink spectrum slicing for embb and urllc users," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 11 051–11 065, 2022.

[6] M. Setayesh, S. Bahrami, and V. W. Wong, "Resource slicing for embb and urllc services in radio access network using hierarchical deep learning," *IEEE Trans. on Wireless Commun., 21 (11), 8950-8966*, 2022.

[7] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[8] L. Zanzi, V. Sciancalepore, A. Garcia-Saavedra, and X. Costa-Perez, "Ovnes: Demonstrating 5g network slicing overbooking on real deployments," in *IEEE INFOCOM'18*, pp. 1–2.

[9] J. X. Salvat, L. Zanzi, A. Garcia-Saavedra, V. Sciancalepore, and X. Costa-Perez, "Overbooking network slices through yield-driven end-to-end orchestration," in *ACM CoNEXT '18*, p. 353–365.

[10] R. B. Myerson, "Optimal auction design," *Mathematics Operations Research*, vol. 6, no. 1, pp. 58–73, 1981.

[11] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of Queueing Theory*. Wiley, 2018.

[12] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5g: physical and mac-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.