A Profit-Maximizing Strategy of Network Resource Management for 5G Tenant Slices

Bin Han, Member, IEEE, Di Feng, Lianghai Ji, Student Member, IEEE, and Hans D. Schotten, Member, IEEE

Abstract—Supported by the emerging technologies of Network Function Virtualization (NFV) and network slicing, 5G networks allow tenants to rent resources from mobile network operators (MNOs) in order to provide services without possessing an own network infrastructure. The MNOs are therefore facing the problem of deciding if to accept or decline the resource renting requests they receive. This paper builds a stochastic model that describes the MNO’s revenue and opportunity cost of accepting a contract, and therewith proposes a strategy that is analytically derived to maximize the expected profit at every decision.

Index Terms—Network slicing, multi-tenant network, profit model, network resource management, 5G network optimization

I. INTRODUCTION

NETWORK slicing was proposed by the Next Generation Mobile Networks (NGMN) Alliance [1], since then it has become one of the hottest topics in the filed of future 5th Generation (5G) mobile communication networks. Generally, the concept of network slicing can be understood as creating and maintaining multiple independent logical networks (slices) on a common physical infrastructure, each slice operates a separate business service. Enabled and supported by the emerging technologies of software defined networks (SDN) and network function virtualization (NFV), network slicing exhibits great potentials, not only in supporting specialized applications with extreme performance requirements, but also in benefiting the mobile network operators (MNOs) with increased revenue [2].

As pointed out by Rosti et al. [2], a sliced mobile network manages its infrastructure and virtual resources in independent scalable slices, each slice runs a homogeneous service with simple business model. Thus, an MNO can dynamically and flexibly create, terminate and scale its slices to optimize the resource utilization for a better revenue or profit.

In a previous paper [3], we have proposed a profit optimization model for sliced mobile networks that run in the traditional business mode: the MNOs with network resources implement the slices and provide all network services directly to their end-users. In this case, an MNO has full a priori knowledge about the service demand and the cost/revenue models of its every slice. It is able to scale the slices according to their profiting efficiencies, in order to achieve the maximal overall profit under the resource constraint. This is a classic multi-objective optimization problem (MOOP), in which the main challenge is to solve the optimum, or at least to find a satisfactory solution, with an affordable computing effort.

Unfortunately, this model does not apply to the slices run by tenants such as mobile virtual network operators (MVNOs), which are considered to play an important role in 5G networks [4]. Tenants are third-parties that provide services without owning any network infrastructure, e.g. utility/automotive companies and over-the-top service providers such as YouTube. To implement services, they have to be granted by MNOs with network resources, including radio / infrastructure resources and virtualized resource blocks. In legacy networks, every tenant makes its contractual agreement with the MNO(s), to pay a fixed and coarsely estimated annual/monthly fee for these resource sharing concepts. In the context of network slicing, in contrast, the resources are first bundled into slices before granted to tenants upon demand. Depending on the type, size and lifetime of granted slice, the fee is specified. This approach improves the sharing efficiency and the resource utilization rate. However, as such slices are maintained by tenants, the MNO has neither insight into their efficiencies of making revenue at end-users, nor authentication to rescale or terminate them during their lifetime. Instead, the MNO formulates the fee rate for different resource bundles, and chooses if to accept or decline the slice requests from tenants, like discussed in [5]. In this case, the MNO cannot jointly optimize all slices in a fully dynamic approach, but only attempt to make the best decision for every received request, which is a problem of decision theory and operations research.

In this paper, we will focus on the case of tenant slices, and propose an economic model that evaluates the profit of an MNO to accept a certain slice request from tenant. Based on this model we propose a decision strategy to maximize the expected overall network profit at every decision step. The rest part of the paper is organized as follows: In Sec. II we simplify the problem described above to an approximate business model. Then we build the profit model in Sec. III starting with an ideal simple case and then approach to the complex reality step-by-step, on every step we deduce a profit-maximizing decision strategy from the proposed profit model. At the end we close the paper with our conclusion and outlooks in Sec. V.

II. SIMPLIFIED BUSINESS PROBLEM

A. Fundamental Assumptions

Our study begins with some basic assumptions and approximations on the business case. First of all, in most countries
and regions, the mobile network infrastructure is controlled by only a few or even one single MNO, i.e. the network resource market is never perfectly competitive but highly oligarchy or monopoly. Hence, in this work we consider the case with only one MNO, ignoring the competition between different MNOs.

Second, the MNO holds a resource pool, which contains resources of certain types. Self-evidently, resource of every type is limited in amount. To rent these resources to tenants, a list of available contracts is provided by the MNO, every type is limited in amount. To rent these resources to tenants, a list of available contracts is provided by the MNO, every contract defined by a resource bundle, a contract period and a periodical payment. Every resource bundle is specified for a reference slice of certain type and size. We assume that the list of available contracts are predefined and remains consistent.

When a tenant requires resources to implement a slice, it selects one from the available contracts, requests to possess the corresponding resource bundle for the contract period. The MNO then decides if to accept or decline the request. Upon acceptance, the contract is confirmed and the tenant periodically pays the quota defined in the contract. If denied, the requested resource bundle will not be dedicated, and can be flexibly exploited for the MNO’s own slices to make revenue.

We consider that a confirmed contract cannot be terminated or modified within its period. We neglect the priority of contract renewals over new contract establishments, i.e. a tenant obtains no advantage for its future requests from the current contract. We also consider only nonelastic slices and neglect resource multiplex over slices, i.e. no resource can be allocated to multiple contracts simultaneously. Therefore, when accepting the current request, the MNO also loses some opportunity of accepting potential better deals in future.

The requests arrive stochastically. Usually, the arriving rate remains on a certain level and the intervals between different arrivals are independent from each other. Hence, it is reasonable to consider the number of arriving requests in a certain period as Poisson distributed. To simplify the model we consider an enough short unit period so that the request arrivals can be approximated as a Bernoulli process. We also assume that the MNO possess full a priori knowledge about the statistics of arriving requests (resource bundle and contract period), which we consider as consistent.

### B. Model Setup

To normatively describe the simplified business model above, we define the following sets, variables and mappings:

- $\Psi = [0, 1]^N$: a general $N$-dimensional non-negative Euclidean space to measure normalized network resource bundles in reference to the maximal resource pool, where $N$ is the number of resource types.
- $\psi_t \in \Psi$: the normalized measure of idle resource pool available at discrete time $t \in \mathbb{N}$.
- $\Omega_t \subset \Psi$: the finite, discrete set of resource bundles defined by all contract options provided by the MNO with its idle resources at time $t$. Each element in $\Omega_t$ is a possible resource bundle requested at time $t$. The case of no request arrival is considered as a null-request, i.e. $\omega_{null} = 0^N \in \Omega_t, \forall t \in \mathbb{N}$. We also consider that the entire resource pool is idle at $t = 0$ so that $\Omega_0$ is a predefined set of resource bundles defined by the list of all available contracts, and generally we have a generation function:

$$\Omega_t = G(\psi_t) = \{\omega|\omega \leq \psi_t\} \cap \Omega_0, \forall t \in \mathbb{N}\tag{1}$$

- $T \subset \mathbb{N}^+$: the finite, discrete set of contract periods in all contract options, inf$(T) = 1$ denotes a unit time period.
- $P \subset \mathbb{R}$: the finite, discrete set of payments defined by all contract options.
- $\Omega_0 \overset{F}{\rightarrow} [0, 1]$: a probability measure on $\Omega_0$ representing the probabilities of upcoming request for different contracts in each unit time period.
- $\Omega_0 \times \mathcal{P}(\Omega_0) \overset{f}{\rightarrow} [0, 1]$: a probability measure representing the probabilities of upcoming request for different contracts that can be supported by a given idle resource pool.

We consider requests for oversize contracts beyond the support of current idle resource pool as equivalent of null contracts, hence generally:

$$f(\omega, \Omega) = \begin{cases} 0 & \omega > \sup(\Omega) \\ g(\omega_{null}) + \sum_{\mu > \sup(\Omega)} g(\mu) & \omega = \omega_{null} \\ g(\omega) & \text{otherwise} \end{cases} \tag{2}$$

- $\Psi \times \Omega_0 \overset{F}{\rightarrow} \Psi$: the decision strategy that determines if to accept the arrived requests. It maps from the current idle resource pool and a received request to the idle resource pool at the next period, i.e. $F(\psi_t, \omega_t) = \psi_{t+1}$.
- $\Omega_0 \times T \overset{P}{\rightarrow} P$: a pricing function that maps the resource bundle and contract period $(\omega, T)$ to the corresponding periodical payment, the payment of a null contract is zero i.e. $p(\omega_{null}, T) = 0, \forall T \in T$.
- $\Omega_0 \overset{q}{\rightarrow} \mathbb{R}$: a function that maps a resource bundle to the corresponding revenue it can generate within one unit time period through its deployment in MNO’s own slices. In this work we consider $q$ as a linear function of $\omega$.

### III. Profit Model and Decision Strategy

Now we analyze the MNO’s profit of accepting a resource request from tenant. In the following part of the paper, we use the term $(\psi_t, \omega_t, T_t)$ to represent the request arriving at $t$.

#### A. Two-Step Decision, Non-Expiring Contract

We start with a simple two-step model, where both the MNO and the tenant only act two unit periods, i.e. $t \in \{0, 1\}$. We also assume that no contract will expire, which means $T_t = 2 - t$. Now for the two periods we have two idle resource pools $\psi_0, \psi_1$. As there is no contract expiry, no resource is released at $t = 1$ so that $\psi_1 = \begin{cases} \psi_0 - \omega_0 & (\psi_0, \omega_0, 2) \text{ accepted;} \\ \psi_0 & \text{otherwise.} \end{cases} \tag{3}$

Obviously, as the MNO only operates two periods, it should accept any contract request at $t = 1$ as long as its resource

\footnote{Note that here with $\omega \leq \psi_t$ we denote $\omega$ is not greater than $\psi_t$ in any of the $N$ dimensions.}

\footnote{$\mathcal{P}(\Omega_0)$ denotes the power set of $\Omega_0$.}
pool supports. So the focus is on the decision at \( t = 0 \). Given any contract request \((\psi_0, \omega_0, 2)\) arrived at \( t = 0 \), the expected payoff of accepting this request is

\[
\Gamma_1(\psi_0, \omega_0) = (1 + \beta)p(\omega_0, 2) - C_1(\psi_0, \omega_0),
\]

where \( \beta \in (0,1) \) is the discount factor to describe the time value of money \((\text{TVM})\) \([6]\), which is determined by the capital market. \( C_1 \) is the Opportunity Cost \((\text{OC})\) of this contract:

\[
C_1(\psi_0, \omega_0) = (1 + \beta)q(\omega_0)
+ \beta \sum_{\omega \in \Omega_0} [f(\omega, \Omega_0) - f(\omega, G(\psi_0 - \omega_0))]p(\omega, 1).
\]

Obviously, to maximize the profit, the MNO is supposed to follow the optimal strategy of binary decision:

\[
\psi_1 = F_{1, \text{opt}}(\psi_0, \omega_0) = \begin{cases} 
\psi_0 - \omega_0 & \Gamma_1(\psi_0, \omega_0) \geq 0; \\
\psi_0 & \text{otherwise},
\end{cases}
\]

where case 1 denotes acceptance and case 2 for declination.

**B. Multi-Step Decision, Non-Expiring Contract**

Then we progress towards the multi-step model, where the MNO and the tenant act \( t_{\text{max}} \) unit periods. Once again, as of this step we still consider non-expiring contracts, so that the request arriving at \( t \) has a contract period of \( T_t = t_{\text{max}} - t \) and hence a periodical payment of \( p(\omega_t, t_{\text{max}} - t) \). Taking it into account that time present value of any future payment \( x \) in \( \Delta t \) periods from current under a discount factor \( \beta \) is \( x \beta^{\Delta t} \), we can compute the total present value of all payments for the requested contract \((\psi_t, \omega_t, t_{\text{max}} - t)\) as

\[
p_{t_{\text{max}}}(\omega_t) = \sum_{t = 1}^{t_{\text{max}} - 1} \beta^{t-1} p(\omega_t, t_{\text{max}} - t).
\]

Similarly, the present value of exploiting the resource bundle \( \omega_t \) on the MNO’s own slice for \( t_{\text{max}} - t \) can be computed as

\[
q_{t_{\text{max}}}(\omega_t) = \sum_{t = 1}^{t_{\text{max}} - 1} \beta^{t-1} q(\omega_t).
\]

Thus, the OC of any request arriving at \( t \) is

\[
C_{t_{\text{max}}}(\psi_t, \omega_t) = q_{t_{\text{max}}}(\omega_t) + \sum_{t = 1}^{t_{\text{max}} - 1} \beta^{t-1} \sum_{\omega \in G(\psi_t)} [f(\omega, \Omega_t) - f(\omega, G(\psi_t - \omega_t))]p(\omega, t_{\text{max}} - t).
\]

Therefore, let

\[
\Gamma_{t_{\text{max}}}(\psi_t, \omega_t) = p_{t_{\text{max}}}(\omega_t) - C_{t_{\text{max}}}(\psi_t, \omega_t),
\]

the profit-maximizing decision by the MNO should be

\[
\psi_{t+1} = F_{t_{\text{max}}, \text{opt}}(\psi_t, \omega_t) = \begin{cases} 
\psi_t - \omega_t & \Gamma_{t_{\text{max}}}(\psi_t, \omega_t) \geq 0; \\
\psi_t & \text{otherwise},
\end{cases}
\]

Note that as the OC is not promised to be convex about the decision sequence \([\psi_0, \ldots, \psi_{t_{\text{max}} - 1}]\), \(11\) is not guaranteed to achieve the global optimum of \( C_{t_{\text{max}}}(\psi_t, \omega_t) \), but a local maximum. Besides, according to \(9\), the OC \( C_{t_{\text{max}}} \) depends on the expected idle resource pool in future \( \psi_{t+1} > \psi_t \), which is determined by the target decision strategy \( F_{t_{\text{max}}, \text{opt}} \), whose solution \(11\) relies on \( C_{t_{\text{max}}} \). This closed loop encourages to apply an iterative approach, as briefly described in Fig. 1.

As \( C_{t_{\text{max}}} \) has non-negative terms and bounded partial sums, it converges with increasing \( t_{\text{max}} \). Therefore the sequence \( C_{t_{\text{max}}}^i \) where

\[
\text{for } i = 0 \text{ to } t_{\text{max}} \text{ do }
\begin{align*}
& \text{Generate a request sequence according to } g(\omega) ; \\
& \text{Determine the decision sequence according to } F_{t_{\text{max}}}^i; \\
& \text{Compute the OC } C_{t_{\text{max}}}^i(\psi_i, \omega) \text{ under the new decision sequence for all } \omega \in \Omega_0 \text{ according to } 9; \\
& \text{Update the strategy } F_{t_{\text{max}}}^{i+1} \text{ with the OC according to } 11; \\
& \text{if } i \geq 1, \sum_{\omega \in \Omega_0} |C_{t_{\text{max}}}^i(\psi_i, \omega) - C_{t_{\text{max}}}^{i-1}(\psi_i, \omega)| < \gamma \text{ then return } F_{t_{\text{max}}}^{i, \text{opt}}; \\
& \text{else Update the strategy } F_{t_{\text{max}}}^{i+1} \leftarrow F_{t_{\text{max}}}^{i, \text{opt}}; \\
& \text{end return } F_{t_{\text{max}}}^{t_{\text{max}}, \text{opt}}.
\end{align*}
\]

Fig. 1: An iterative algorithm to approximate the target strategy. In each iteration the strategy is tested and correspondingly updated with a simulated request sequence, and the convergence is evaluated by the estimated total OC of the simulated request sequence under the updated strategy.

In real world, the MNO and tenants shall be considered as long-term or even eternally operating, i.e. \( t_{\text{max}} \to +\infty \). According to \(11\), the sequence \((\psi_0, \psi_1, \ldots)\) monotonically decreases, and \(G(\psi_t)\) converges to an empty set:

\[
\lim_{t \to +\infty} G(\psi_t) = \emptyset.
\]

Therefore we know that \( C_{t_{\text{max}}}^i(\psi_t, \omega_t) \) is bounded as \( t_{\text{max}} \to +\infty \). Meanwhile, as \( p \) is bounded and \( \beta \in (0,1) \), \( p_{t_{\text{max}}} \) also converges to a bounded value as \( t_{\text{max}} \to +\infty \), so the proposed approach also applies to the infinite-step case.

Generally, in the case of non-expiring contract, at every step of decision, the impacts of all historical decisions about previous requests are completely reflected in the current idle resource pool without any extra influence in the future. Hence, the non-expiring contract model is a Markov model, and the proposed decision strategy is also Markovian as well.

**C. Multi-Step Decision, Expiring Contract**

Subsequently we bring the issue of contract expiry into our discussion. Consider flexible contract periods \( T_t < t_{\text{max}} - t \).
permitted, the resource bundle assigned to it will be released to the MNO’s resource pool after the expiry. Thus, the idle resource pool size $\psi_t$ is not monotonically decreasing with $t$, but jointly determined by all previous requests and decisions during $[0, t-1]$. Hence, the model becomes non-Markovian and the Markovian decision strategy \((11)\) does not apply as it lacks information about contract periods and previous decisions.

As a solution to this, given an arbitrary request \((\psi_t, \omega_t, T_t)\) arriving at $t$, we denote the indicator function

\[
I_t(t, T_t) = \begin{cases} 
1 & \tau \leq t \leq t + T_t, (\psi_t, \omega_t, T_t) \text{ accepted; } \\
0 & \text{otherwise}
\end{cases}
\]

(13)

to represent its validity at time $t$. Thus, we can represent the resource bundle reserved for it at any time $t$ as

\[
\omega_t^{T_t} = \omega_t I_t(t, T_t)
\]

(14)

Then we define the scalar

\[
\bar{\omega}_t = \sum_{\tau=0}^{t-1} \omega_t^{T_t}
\]

(15)

to track and aggregate the resource bundles currently reserved by all previously accepted contracts. Thus, instead of \((\psi_t, \omega_t, T_t)\), now we use \((\bar{\omega}_t, \omega_t, T_t)\) to represent a contract request. The equations \((7)\) and \((8)\) become

\[
p_{\text{max}}^{T_t}(\omega_t) = \sum_{\tau=t}^{t+T_t-1} \beta^{t-\tau} p(\omega_t, T_t),
\]

(16)

\[
q_{\text{max}}^{T_t}(\omega_t) = \sum_{\tau=t}^{t+T_t-1} \beta^{t-\tau} q(\omega_t),
\]

(17)

respectively, and thus the OC of accepting the request is

\[
C_{\text{max}}^{T_t}(\bar{\omega}_t, \omega_t) = q_{\text{max}}^{T_t}(\omega_t) + \sum_{\tau=1}^{t-1} \beta^{t-\tau+1} \times \sum_{\omega \in G(\psi_t)} [f(\omega, \Omega_t) - f(\omega, G(\psi_0 + \bar{\omega}_t - \omega_t^{T_t}))]p(\omega, T_t).
\]

(18)

Defining the payoff function

\[
\Gamma_{\text{max}}^{T_t}(\bar{\omega}_t, \omega_t) = p_{\text{max}}^{T_t}(\omega_t) - C_{\text{max}}^{T_t}(\bar{\omega}_t, \omega_t),
\]

the non-Markovian profit-maximizing decision strategy is

\[
F_{\text{max}}^{T_t, \text{opt}}(\psi_t, \omega_t) = \begin{cases} 
(\psi_0 + \bar{\omega}_t + 1 - \omega_t, \omega_t^{T_t}) & H_{\text{max}}^{T_t}(\psi_t, \omega_t) \geq 0; \\
(\psi_0 + \bar{\omega}_t + 1, \omega_t^{T_t}) & \text{otherwise}
\end{cases}
\]

(20)

In finite-step cases where $t_{\text{max}}$ is finite, the number of possible sequences \(((\omega_t, T_t))_{0 \leq t \leq t_{\text{max}}} \text{ is limited so that } C_{\text{max}}^{T_t} \text{ is bounded and the iterative approach in Fig. 1 still applies. However, when } t_{\text{max}} \to +\infty, \text{ as } (\psi_t)_{t \in \mathbb{N}} \text{ is not monotonic about } t, \text{ (12) fails to hold and no more convergence is guaranteed. In this case, an artificial finite } t_{\text{max}} \text{ is needed, like we did in the footnote 4.}

IV. DISCUSSIONS

So far we have closed our study on the MNO’s decision strategy under the assumptions in Sec. II-A. Nevertheless, concerning the strength of our assumptions, some discussions about their feasibilities may be necessary.

The first concern can be raised by the monopoly model with only one MNO, because the tenants are may fail to obtain resources to maintain services in this case. Practically, a pre-ordering mechanism can be applied to allow early requests before the due of slice creation. Thus, if a request is declined, the tenant is still able to reattempt with another contract option with better chance of acceptance. Moreover, it is true that in practice there is usually not only one but several MNOs, i.e. the market is actually oligarchy. While providing the tenant more alternative options when its request is rejected by one MNO, this fact does not necessarily conflict with our results under the monopoly assumption, because oligarchy markets differ from monopoly ones only in the supply, demand and pricing mechanism, but not in the decision making logic [8], which we focus on in this paper. Nevertheless, the competition between MNOs in oligarchy markets worths further study.

Another doubt may arise about the exclusion of unexpected contract termination and modification. Certainly they are ignored here for model simplification, and can be eventually involved in future work by introducing another random variable with corresponding statistics. Similarly, in this work we have ignored the preference in renewing old expiring contracts over making new ones. For a better approximation to reality we can consider all contracts as non-expiring, and apply the random termination event instead to describe the tenant cancellation.

At last, the assumption that MNO possesses full a priori knowledge about the request statistics may be argued. In practice, although the a priori model is hard to obtain, it can be estimated in a Bayesian approach from the a posteriori historical records that every MNO keeps, as long as it remains consistent. Even in case of non-stationarity, short-term consistence can still be approximated with periodical updates.

V. CONCLUSION

In this work, in an operations research perspective we have investigated the 5G network resource management problem of creating tenant slices upon request, aiming at a strategy that maximizes the expected MNO revenue at every binary decision. Different cases have been studied and an iterative algorithm proposed. The convergence of proposed method has been mathematically derived. For future works, numerical experiments are expected to evaluate the converging speed and the revenue gain of the proposed algorithm in both short and long terms, elastic slices shall also be considered to enable resource multiplexing between slices.

REFERENCES

[1] “5G White Paper,” Next Generation Mobile Networks, White Paper, 2015.
[2] P. Rost, C. Mannweiler et al., “Network slicing to enable scalability and flexibility in 5G mobile networks,” IEEE Communications Magazine, vol. 55, no. 5, pp. 72–79, 2017.
[3] B. Han, S. Tayade, and H. D. Schotten, “Modeling Profit of Sliced 5G Networks for Advanced Network Resource Management and Slice Implementation,” in The 22nd IEEE Symposium on Computers and Communications (ISCC), July 2017.

[4] P. Rost, A. Banchs et al., “Mobile Network Architecture Evolution toward 5G,” IEEE Communications Magazine, vol. 54, no. 5, pp. 84–91, 2016.

[5] D. Bega, M. Gramaglia et al., “Optimising 5g infrastructure markets: The business of network slicing,” in The 36th IEEE International Conference on Computer Communications (INFOCOM), 2017.

[6] B. E. Needles and S. V. Crosson, Managerial Accounting. Nelson Education, 2013.

[7] R. G. Bartle and D. R. Sherbert, Introduction to Real Analysis. Wiley New York, 2000, vol. 2.

[8] A. Mas-Colell, M. D. Whinston, J. R. Green et al., Microeconomic Theory. Oxford University Press New York, 1995, vol. 1.