This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2939935, IEEE Access

Digital Object Identifier

Maximizing infrastructure providers' revenue through network slicing in 5G

MATTEO VINCENZI[1], ELENA LOPEZ-AGUILERA[2], AND EDUARD GARCIA-VILLEGAS[3]
[1]Universitat Politecnica de Catalunya (e-mail: matteo.vincenzi@upc.edu)
[2]Universitat Politecnica de Catalunya (e-mail: elopez@entel.upc.edu)
[3]Universitat Politecnica de Catalunya (e-mail: eduardg@entel.upc.edu)

Corresponding author: M. Vincenzi (e-mail: matteo.vincenzi@upc.edu)

ABSTRACT Adapting to recent trends in mobile communications towards 5G, infrastructure owners are gradually modifying their systems for supporting the network programmability paradigm and for participating in the slice market (i.e., dynamic leasing of virtual network slices to service providers). Two-fold are the advantages offered by this upgrade: i) enabling next generation services, and ii) allowing new profit opportunities. Many efforts exist already in the field of admission control, resource allocation and pricing for virtualized networks. Most of the 5G-related research efforts focus in technological enhancements for making existing solutions compliant to the strict requirements of next generation networks. On the other hand, the profit opportunities associated to the slice market also need to be reconsidered in order to assess the feasibility of this new business model. Nonetheless, when economic aspects are studied in the literature, technical constraints are generally oversimplified. For this reason, in this work, we propose an admission control mechanism for network slicing that respects 5G timeliness while maximizing network infrastructure providers' revenue, reducing expenditures and providing a fair slice provision to competing service providers. To this aim, we design an admission policy of reduced complexity based on bid selection, we study the optimal strategy in different circumstances (i.e., pool size of available resources, service providers' strategy and traffic load), analyze the performance metrics and compare the proposal against reference approaches. Finally, we explore the case where infrastructure providers lease network slices either on-demand or on a periodic time basis and provide a performance comparison between the two approaches. Our analysis shows that the proposed approach outperforms existing solutions, especially in the case of infrastructures with large pool of resources and under intense traffic conditions.

INDEX TERMS Communication networks, 5G mobile communication, network slicing, infrastructure as a service, traffic control, admission control, queuing analysis, Markov processes, pricing, profitability.

I. INTRODUCTION

In the last decades we have assisted to the frequent emergence of new use cases for wireless networks proposed by industrial actors and governmental bodies. Consequently, network infrastructure owners have been motivated to explore new architectures and technologies for upgrading their networks and support new services, while seeking economic incentives for amortizing the associated costs. 5G, the next generation of mobile networks, is still far from its maturity in terms of deployment, however, requirements have been proposed by standardization bodies [1]–[3], and new technologies are being fine-tuned by the research community, while the resulting architectures and mechanisms are being integrated in 3rd Generation Partnership Project (3GPP) specifications [4]–[7]. In particular, network function virtualization (NFV) and software defined networking (SDN) have been proposed as the keystones for scalable and programmable networks with Quality of Service (QoS) support [6]–[8]. Besides, they are considered as the enablers of the *network slicing* paradigm, according to which, QoS-tailored portions of the network resources are dynamically isolated into customized virtual networks, namely *network slices*, that coexist within the same infrastructure. Therefore, an alternative business model has been introduced [3], [4], [9], named *slice market*, between infrastructure providers (InPs), that is, access, transport and cloud infrastructure owners, and service providers (SPs), which include mobile virtual network operators (MVNOs), over-the-top (OTT) players (e.g., streaming providers), and vertical industries (e.g., e-health, surveillance, automotive). More precisely, slices can be leased by InPs to SPs (also known as *slice tenants*) through fine-scale service level agreements (SLAs) that substitute current long-term sharing agreements.

The concept of slice market is expected to introduce a strong competition between different InPs and SPs, thus oxygenating the typically closed and monolithic ecosystem of telecommunication services and introducing the preconditions for fast innovation. Indeed, independently from the ownership of network resources, any SP could possibly enter the market of wireless services, while InPs could better manage and monetize the utilization of their resources. Therefore, from an economic point of view, the enablers of a healthy slice market for 5G are: i) the monetary incentives to InPs for building the next generation network, and ii) the fairness in the service of competing SPs. On the other hand, from a purely technical point of view, the requirements for 5G are: i) the slice isolation [8], ii) heterogeneous End-to-End (E2E) QoS guarantees for 5G use cases [1]–[6], and, iii) a prompt slice provision, suitable for short-lived services such as emergency services or surveillance [1], [2].

Excluding architectural and technological aspects that have been extensively studied in the literature, the promptness in the slice provision is mainly regulated by two factors, that is, the communication protocol adopted between SPs and InPs, and the mechanisms used at the InPs' side for admission control, resource allocation and pricing. In this context, two macro categories of slice provision approaches exist in the literature, the *on-demand* and *periodic slicing* where, respectively, slice allocation is enforced upon each slice request arrival (e.g., policy-based approaches) or periodically (e.g., auction-based approaches). In on-demand slicing, the typical communication flow for the slice provision process consists in the uncoordinated slice request submission by SPs, followed by the broadcasting of the admitted tenants by InPs. On the other hand, in periodic slicing, an intrinsic latency is systematically added by the time window used for collecting slice requests.

Within this categorization, two strategies are mainly used in the literature for resource pricing. In on-demand slicing, prices are typically set by InPs for a given bundle of resources. On the other hand, in periodic slicing, prices are determined in relation to the resource availability as well as InPs' and SPs' strategies. Besides, a bidding model is generally adopted where the minimum and maximum bid represent, respectively, the reserve price (i.e., the minimum price accepted by the InPs), and the SPs' budget (i.e., the maximum affordable price). Many contributions exist in the literature for admission control, resource allocation and billing mechanisms in virtualized wireless networks [10], however, as detailed in the next section, most of the existing approaches do not meet neither the economic conditions for a healthy slice market, nor the 5G requirements.

In this work, we propose a timely admission control mechanism for network slicing that maximizes InPs' revenues, reduces operational expenditures and guarantees slice isolation, QoS and fairness towards SPs. In this context, InPs have the joint objective of maximizing the tenants' admission rate while prioritizing the most rewarding slice requests. Therefore, from a technological point of view, InPs have the incentive to perform the slice allocation process as fast as possible once triggered by the arrival of a slice request, since every request represents a potential source of revenue. On the other hand, from a strategical point of view, the InPs have the incentive to prioritize those slice requests with higher bids and characterized by a high ratio among arrival and service rates.

In order to maximize the slice provision promptness and reduce the computational cost of the allocation process, we propose a policy-based approach, named Above Threshold (AT) policy, that maximizes InPs' revenues by admitting slice requests with associated bids greater or equal than a given threshold tariff. In this regard, we consider two kinds of policies differing in the admission strategy with respect to the resource utilization, named *State Dependent (SD)* and *State Independent (SI)* policies, respectively. In particular, the former guarantees a maximum revenue for every number of instantiated slices, that is, it depends on the available resources, while the second maximizes revenues only in the long term and, therefore, requires lower computational expenses.

In conclusion, the main contributions of this work are the proposal of an AT admission control mechanism (both SD and SI) for network slicing in 5G together with its benchmarking with reference strategies when different resource pool sizes, traffic loads, and slicing frequencies are considered. Besides, to the best of our knowledge, this is the first effort in comparing on-demand and periodic slicing with respect to fairness towards SPs, resource utilization, InP's profit, and timeliness. In particular, we compare AT policies with the Always Admit (AA) policy in the on-demand case, and with the First-Come-First-Served (FCFS) and Best Bid (BB) policies in the periodic case. Results illustrate that the proposed SI solution is capable of outperforming the evaluated reference mechanisms in terms of revenue rates to the InPs, mostly in case of intense traffic conditions, while reducing the resource utilization in exchange for a negligible loss in terms of admission rate. Besides, it requires the lowest computational expenses and guarantees the promptest admission control, especially when complex infrastructures are examined.

In the remaining of the paper, we first present the related works (Section II) and system model (Section III). Then we introduce the mathematical framework for studying the performance of on-demand slicing when both SD and SI policies are employed (Section IV). The system analysis concludes with considerations on the optimal policies and on the complexity of the evaluated solutions. Afterwards, we introduce the system setup and the results of the performance comparison between on-demand and periodic slicing, when different policies are employed (Section V). Finally, we present the conclusions of this work (Section VI).

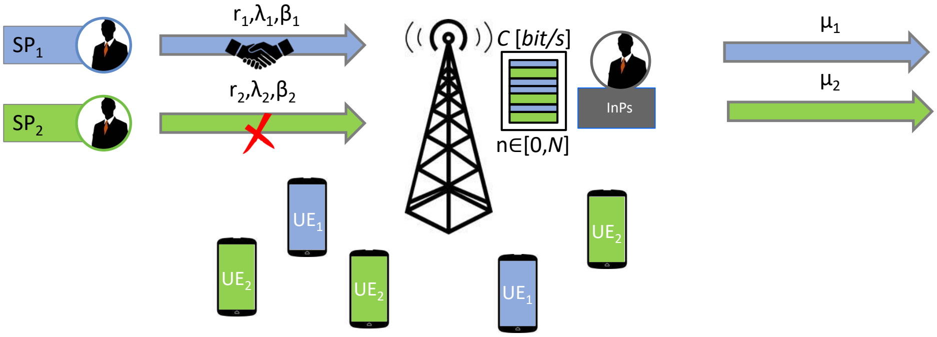


FIGURE 1. System model for slice provision when one InP leases resources to multiple SPs competing for providing service to their UEs. Colors identify the portion of resources used (e.g., channel capacity C) and the UEs served by different slice tenants. Rejected slice requests are marked with a red cross.

II. RELATED WORK

Many contributions exist in the literature for admission control, resource allocation and billing mechanisms in virtualized wireless networks [10], however, rarely both the economic conditions for a healthy slice market and 5G requirements are met. Consequently, the discussion remains open in the scientific community with respect to automated mechanisms for slice provision and pricing in 5G. In particular, [9] and [11] propose on-demand solutions to the admission control problem that maximize the InPs' profit by means of Semi-Markov Decision Processes and optimization theory, respectively. Moreover, [9] introduces the concepts of *inelastic and elastic services*, that will be used in the following, and which are associated to SLAs characterized by constant or average QoS requirements, respectively. However, both contributions lack in the review of other performance metrics relevant for 5G, for instance, fairness towards competing SPs.

On the other hand, among the proposed periodic approaches, [12], [13] employ auction theory for the study of the single/heterogeneous resource allocation problem, respectively, nevertheless, neither of the works puts a focus on network isolation, QoS support or fairness. Besides, although InPs are the entities entitled to build next generation networks, many contributions only take into account the economic return for SPs. For instance, this is the case of the spectrum leasing optimization framework presented in [14], the Fisher market slice allocation approach with strategic tenants in [15], the auction-based approach in [13] and, in general, the VCG-based auctions [16]. Finally, only limited efforts have been produced in the study of pricing schemes suitable for 5G, for instance, [17], [18] propose auction-based solutions for heterogeneous resource slicing with a per-access pricing scheme. However, in [17] the authors highlight the need for a pricing scheme based on slices' lifetime in order to account for the real resource occupation, and to reduce the risk of exaggerated slice requests and unused resources.

In conclusion, research efforts focusing in on-demand and periodic slicing tend to study complementary aspects related to the 5G slice market, therefore, we consider interesting a direct comparison between the two strategies through the same analytical framework. In this context, [19] extends the on-demand approach in [9] for the study of InPs' profits to the periodic case with heterogeneous resources. However, static InP strategies are adopted with no hint on the optimal admission strategy, nor on the fairness towards competing SPs. Reference [11] partly completes the contribution in [19] by proposing a genetic-based algorithm for online computation of the admission policy that maximizes InP's profit.

In this work, we propose a timely admission control mechanism for network slicing that takes into account the economic conditions for a healthy slice market and addresses the requirements of next-generation networks by maximizing InPs' revenues, reducing operational expenditures, and guaranteeing fairness towards SPs, slice isolation and QoS. In particular, we adopt the promptness offered by on-demand approaches for the admission of new slices, combined with pricing features typical of periodic slicing, where tariffs are set depending on the resource availability, the InPs' strategy and SPs' behavior. Indeed, we assume that SPs may have a different perception of the market and, therefore, make different bids for the same kind of slice. However, as SPs' strategies have been abundantly studied in the literature and our focus remains on InPs' perspective, we assume that SPs are irrational entities that follow a random bidding model. Moreover, we assume that tenants pay for the slices they use only if the associated SLA is met during their permanence in the network, therefore, InPs can reallocate resources only after voluntary tenants' departures.

In order to maximize the slice provision promptness and the InPs' revenue while reducing the computational cost associated to the admission decision, we propose the AT policy-based approach that admits slice requests with associated tariff-bids greater or equal than a given threshold. Such an approach is capable of maximizing tenants' admission rate while prioritizing the most rewarding slice requests and, at the same time, it minimizes the admission delay as policies can be enforced instantaneously upon each slice request arrival. In this regard, we compare the performance of both SD and SI admission strategies, which use admission thresholds that can adapt to the current resource utilization, or remain static, respectively. In this study, we model only SLAs associated to inelastic services as they are the strictest class of SLAs. Either way, an extension of this study to include elastic services can be achieved by following the modeling approach in [9]. Finally, we provide a benchmark of the proposed admission control mechanism for network slicing in 5G by comparing on-demand and periodic slicing performance (i.e., fairness towards SPs, resource utilization, InP's profit, and timeliness) with that of reference strategies (i.e., AA in the on-demand case, and FCFS and BB in the periodic case) when different resource pool sizes, traffic loads, and slicing frequencies are considered.

III. SYSTEM MODEL

In this section, we introduce the system model adopted for the analysis and, to this aim, we refer to Fig. 1. In the considered scenario, multiple user equipments (UEs) coexist within the coverage area of a given base station (BS), which belongs to a given InP. The BS represents the access point towards other network resources, such as backhaul, IP networks and cloud infrastructures. UEs can access multiple services at a time, each provided by a different SP, for instance, a given UE can surf the Internet while streaming a song in background. In Fig. 1, the different colors identify different SPs, as well as the portion of InP's resources accessed and the UEs served by different slice tenants. Within this context, different service instances of the same UE are represented as different logical UEs.

Resources are sliced independently at different BS locations and SPs are allowed to actively request network slices on a continuous time scale, while InPs monitor the resource availability and decide whether to admit them, either in real-time (i.e., on-demand slicing) or on a discrete time-scale (i.e., periodic slicing). Whenever InPs welcome a new SP, named *slice tenant*, a SLA is stipulated defining the terms for the customization and pricing of the requested slice. In other words, each SLA defines both the QoS to be guaranteed and the tariff β_s in monetary units per second (e.g., $[euros/s]$) to be paid by tenant SP_s during its permanence in the system. Finally, no distinction is provided in the system model for the labeling of different UEs or SPs, therefore, in Fig. 1, a specific tenant can be licensee of multiple network slices simultaneously, especially when SPs opt for serving different UEs by means of separate slices.

In order to specify a clear model for the SLAs, we first introduce the concept of *service (or slice) class* that we define as $c = \{\boldsymbol{r}_c, \lambda_c, \mu_c\}$. In this context, $\boldsymbol{r}_c = (r_0, \ \cdots, \ r_{\rho-1})$ represents the requirements vector, that is, the set of requirements r_i on the ρ resources accessible from the considered service area, while λ_c and μ_c are the average arrival and service rates of slice requests for the specific service class c, respectively. In particular, $T_c = 1/\mu_c$ is the average *holding (or service) time* for a specific class c, that is, the average time interval during which resources are retained by SPs providing such service. In other words, it holds $T_c = \mathbf{E}[T_{c|s}]$, where $T_{c|s}$ is the holding time of a specific tenant SP_s. Besides, we assume that InPs support a finite number ν of service classes describing the services in the slice market. Finally, if n_c is the number of slices instantiated for class c at a given time instant t, then n represents the total number of network slices instantiated in the network, that is, $n = \sum_{c=1}^{\nu} n_c$. Fixed a specific service class c, the SLA for a given tenant SP_s is defined as the tuple $\{c, \beta_s\}$, where $\beta_s T_{c|s}$ is the price paid to the InPs if the resource requirements are guaranteed during the whole holding time. As introduced in Section I, we examine only the strictest kind of SLAs, that is, those associated to inelastic services [9], characterized by constant requirements during the whole holding time. Besides, we assume that the tariff-bid β_s of a generic SP_s can vary within the interval $[\beta_m^c, \beta_M^c]$, that changes for different slice classes c as they are characterized by different associated resources and perceived value. In particular, the extremes of the bid interval represent, respectively, the minimum tariff accepted by the InP (i.e., the reserve tariff β_m^c) and the maximum tariff that SPs can afford to pay for the considered slice class (i.e., the tariff budget β_M^c). In this context, we model one resource type, that is, the channel capacity C of the access link to the BS, measured in $[bit/s]$, and we leave for future studies the extension to the case of multiple InPs with heterogeneous resources and service classes.

Hence, the definition of service class can be projected into a single resource dimension, by substituting the requirements vector with the scalar r_c, that represents the aggregate nominal rate asked by tenants for the service of UEs in the considered coverage area. Finally, in the rest of the paper, we admit only one service class, that is, all SPs ask the InPs for the same requirement on the aggregate nominal rate, thus, the notation can be simplified by removing subscript c, while SLAs of different tenants are fully described by the corresponding bids β_s.

In this case, the maximum number of slices that can be allocated simultaneously is $N = \lfloor C/r \rfloor$, and it holds $0 \leq n \leq N$. In the following, we assume that the slice request arrivals can be modeled as a Poisson stochastic process with average rate λ, and the tenants' departure as a general stochastic process with average rate μ. With regards to the pricing model, we describe different SPs' behaviors by adopting a bidding model where β_s is a random variable following a general distribution f_β over the sample space $[\beta_m, \beta_M]$.

The proposed system model is valid for both on-demand and periodic slicing, that is, when n is updated at each new admission and departure, or regularly every $T_{slicing}$ seconds. Thus, in Fig. 2, we depict an instance of the slice request, tenants' departure and bidding processes for both approaches. Besides, we highlight the possibility for the InP to reject slice requests depending on the resource availability, the received bids, and the adopted admission policy. Moreover, in the periodic case, slice requests received during a given slicing interval can be admitted at the beginning of the next interval only, when SLAs are enforced. In particular, tenants pay for slices only when they utilize resources, therefore, InPs get no revenue in the time interval between tenants' departure

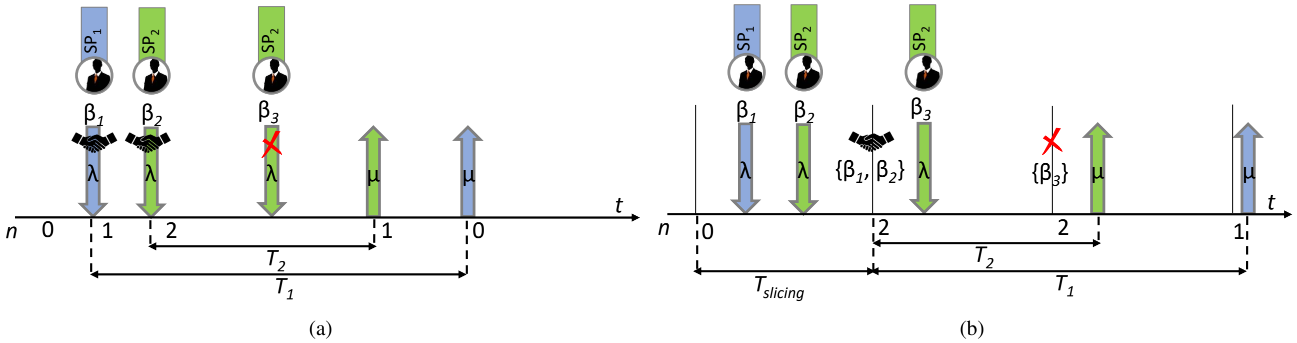


FIGURE 2. Instance of the slice request, tenants' departure and bidding processes in: a) on-demand and b) periodic slicing, when only one service class is supported and $N = 2$. Different colors identify requests and departures of different SPs, moreover, rejected requests are marked by a red cross.

and following slicing interval. We remind that, as we model the problem in function of the aggregate resource demand from the InP perspective only, multiple slice instances can correspond to the same tenant, as represented in Fig. 2.

IV. SYSTEM ANALYSIS FOR ON-DEMAND SLICING

In this section, we present the mathematical analysis for on-demand slice provision mechanisms when different policies are adopted. Regardless of the policy, the infrastructure can be represented as a cloud server farm with capacity to instantiate N equal virtual servers (i.e., the network slices) that share a common pool of jobs to be executed (i.e., the service requests of a given class). New jobs are characterized by an average arrival and service rate equal to λ and μ, respectively, and the number n of jobs executed is updated upon every new job's arrival and completion. Besides, we assume that each virtual server can handle one job at a time, in order to model the slice isolation requirement and the QoS guarantee. Therefore, thanks to the memoryless assumption on arrivals and departures, we can model the system as a $M/G/k/k$ queue[1]. Even in cases where these assumptions do not apply (e.g., non-Markovian behavior of SPs), discrete-time Markov chains could be applied. However, the needed transformations lie outside the scope of this work.

The mathematical framework offered by continuous-time Markov chain (CTMC) can be used for the mathematical analysis of the considered problem. In particular, we can refer to Fig. 3, where each state corresponds to a different tuple $(n, \mathcal{P}_n)$, whose elements describe the number of instantiated slices and the admission policy adopted at that state, respectively. Besides, the generic transition from state n to n^+ coincides either with the admission or departure of a slice tenant, and is associated with the tuple (q_{nn^+}, r_{nn^+}) representing the transition rate conditioned to the initial state and the associated reward, respectively.

The state policy $\mathcal{P}_n$ represents any possible bid-based criterion for admitting or rejecting incoming slice requests at state n:

$$\mathcal{P}_n = \begin{cases} \mathit{Admit}, & \text{if } \beta \in \mathcal{D}_n \subset [\beta_m, \beta_M] \\ \mathit{Reject}, & \text{otherwise} \end{cases} \tag{1}$$

where $\mathcal{D}_n$ is the admitted bid interval at state n. Consequently, the probability for a new slice request to be admitted at state n can be defined as $p_n(f_\beta, \mathcal{P}_n) = p_{\{\beta \in \mathcal{D}_n\}} = \int_{\mathcal{D}_n} f_\beta(\beta)\, d\beta$. State policies $\mathcal{P}_n$ can be arbitrarily chosen by the InP when resources are available in the system, that is, for states $0 \leq n \leq N-1$. On the other hand, when the system faces resource shortage (i.e., $n = N$), the only applicable policy is the rejection of any slice request, that is, $\mathcal{D}_N = \emptyset$ and, thus, $p_N = 0$. Finally, the tuple (q_{nn^+}, r_{nn^+}) associated to a transition at state n can be written as $(\lambda p_n, \beta)$ in case of admission, and as $(n\mu, 0)$ in case of departure. In conclusion, for the generic transition nn^+ it holds:

$$q_{nn^+} = \begin{cases} \lambda p_n, & \text{if } 0 \leq n \leq N-1, n^+ = n+1 \\ n\mu, & \text{if } 1 \leq n \leq N, n^+ = n-1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$r_{nn^+} = \begin{cases} \beta, & \text{if } 0 \leq n \leq N-1, n^+ = n+1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

As introduced in Section I, we assume that the InP can adopt either SD or SI policies, which differ in the capability of adapting the admission strategy to the number of slices isolated in the system. In particular, different or equal policies $\mathcal{P}_n$ are enforced at different states n, respectively. Hence, InP's strategy is represented with the *policy vector* $\boldsymbol{\mathcal{P}} = (\mathcal{P}_0, \ \cdots, \ \mathcal{P}_{N-1})$ in the SD case, while it can be fully described by the generic state policy $\mathcal{P}$ when SI approaches are adopted (i.e., $\boldsymbol{\mathcal{P}} = \mathcal{P}$).

A. STATE-DEPENDENT POLICIES

In CTMC, the stationary probability π_n associated to the generic state n of the system can be calculated through the following balance equations, when SD policies are enforced:

- $0: \quad \pi_0 \lambda p_0 = \pi_1 \mu$
- $1: \quad \pi_1(\lambda p_1 + \mu) = \pi_0 \lambda p_0 + \pi_2 2\mu$
- $n: \quad \pi_n(\lambda p_n + n\mu) = \pi_{n-1}\lambda p_{n-1} + \pi_{n+1}(n+1)\mu$

[1] It shall be noticed that, in the case of periodic slicing, the system can be modeled as a $M^X/G/k/k$ queue, since we could consider that the slice requests received within a given slicing interval arrive in batches at the beginning of the next interval.

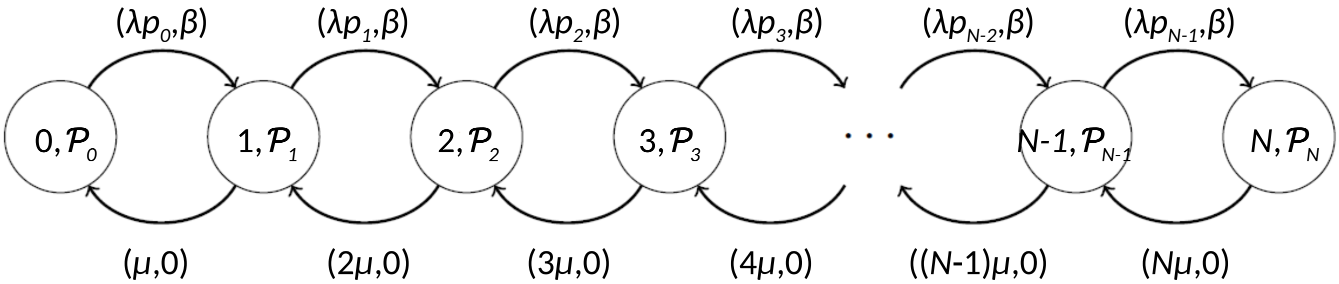


FIGURE 3. Markov chain for on-demand slicing systems, where a different number of instantiated slices n and policy $\mathcal{P}_n$ is associated to each state, while transitions are jointly represented by a transition rate q_{nn^+} and a reward r_{nn^+}.

- N : $\sum_{n=0}^{N} \pi_n = 1$

leading to:

$$\pi_0\Big(\frac{\lambda}{\mu}, f_\beta, N, \boldsymbol{\mathcal{P}}\Big) = \frac{1}{1+\sum_{i=1}^{N}(\frac{\lambda}{\mu})^i/i!\prod_{l=0}^{i-1}p_l}$$
$$\pi_{n\geq 1}\Big(\frac{\lambda}{\mu}, f_\beta, N, \boldsymbol{\mathcal{P}}\Big) = \frac{(\frac{\lambda}{\mu})^n/n!\prod_{j=0}^{n-1}p_j}{1+\sum_{i=1}^{N}(\frac{\lambda}{\mu})^i/i!\prod_{l=0}^{i-1}p_l} \tag{4}$$

Intuitively, in a low-load regime (i.e., when $\frac{\lambda}{\mu} \to 0$), the system most likely operates in states corresponding to low values of n (i.e., $\pi_0 \to 1$), independently from the bidding distribution f_β, the maximum number of slices N, and the InP's strategy $\boldsymbol{\mathcal{P}}$. The same result is obtained under high-load regime (i.e., $\frac{\lambda}{\mu} >> N$), and when a very conservative admission strategy is adopted by the InP (i.e., β_m is increased so that most of the bid distribution lies outside $\mathcal{D}_n$). Conversely, when a more permissive policy is used in high-load regime, the system behavior can be reversed (i.e., $\pi_N \approx 1$).

Following, we obtain the analytical expression for the performance metrics used to measure the efficiency of such slice provision system. The admission probability can be expressed as:

$$P_{admit}\Big(\frac{\lambda}{\mu}, f_\beta, N, \boldsymbol{\mathcal{P}}\Big) = \sum_{n=0}^{N-1}\pi_n p_n \tag{5}$$

and represents the probability for a new slice request to be admitted independently from the number of slices already instantiated in the system. According to (4) and (5), P_{admit} totally depends on the admission probability at state $n = 0$ in low-load regime (i.e., $P_{admit} \to p_0$, when $\frac{\lambda}{\mu} \to 0$). Therefore, according to (1) and to p_n's definition, the InP can improve the system's fairness (i.e., the general satisfaction of competing SPs) by widening the admission interval $\mathcal{D}_0$. In particular, the maximum admission probability in low-load regime can be reached when the state policy $\mathcal{P}_0$ admits every request (i.e., $p_0 = 1$) or, in other words, when the admission interval $\mathcal{D}_0$ includes the entire support of f_β.

The average resource utilization U in the system is defined as the ratio between the average and the maximum number of slices instantiated in the system:

$$U\Big(\frac{\lambda}{\mu}, f_\beta, N, \boldsymbol{\mathcal{P}}\Big) = \mathbf{E}[n]/N = \Big(\sum_{n=0}^{N} n\cdot\pi_n\Big)/N \tag{6}$$

Subsequently, we introduce the expected tariff $\mathbf{E}[\beta|\beta \in \mathcal{D}_n]$ paid by those slice tenants that are admitted at state n according to state policy $\mathcal{P}_n$:

$$\mathbf{E}[\beta|\beta\in\mathcal{D}_n] = \int_{-\infty}^{\infty}\beta\, p_{\{\beta|\beta\in\mathcal{D}_n\}}\, d\beta$$
$$= \frac{1}{p_n}\int_{\mathcal{D}_n}\beta\, f(\beta)\, d\beta \tag{7}$$

where $p_{\{\beta|\beta\in\mathcal{D}_n\}} = (f_\beta(\beta)\cdot 1|_{\beta\in\mathcal{D}_n})/p_n$.

The average revenue rate R_β in $[euros/s]$ for an InP applying a specific policy vector $\boldsymbol{\mathcal{P}}$ can be calculated by averaging, over all the states, the admission rate λp_n in $[admissions/s]$, times the expected price paid by admitted tenants over the average holding time, that is, $\mathbf{E}[\beta|\beta \in \mathcal{D}_n]/\mu$ in $[euros/admission]$:

$$R_\beta\Big(\frac{\lambda}{\mu}, f_\beta, N, \boldsymbol{\mathcal{P}}\Big) = \frac{\lambda}{\mu}\sum_{n=0}^{N-1}\pi_n p_n \mathbf{E}[\beta|\beta\in\mathcal{D}_n] \tag{8}$$

B. STATE-INDEPENDENT POLICIES

The analytical expressions for stationary probabilities and performance metrics of a SI system can be obtained as a particular case of the SD case. In particular, by definition of SI policy, it holds $\mathcal{P}_n = \mathcal{P}$, $\mathcal{D}_n = \mathcal{D}$ and $p_n = p$ for every state $0 \leq n \leq N-1$. Therefore, we can rewrite the stationary probabilities in (4) as:

$$\pi_n\Big(\frac{\lambda}{\mu}, f_\beta, N, \mathcal{P}\Big) = \frac{(\frac{\lambda}{\mu}p)^n/n!}{\sum_{i=0}^{N}(\frac{\lambda}{\mu}p)^i/i!}, \qquad n \geq 0 \tag{9}$$

Similarly, the system admission probability in (5) can be rewritten as:

$$P_{admit}\Big(\frac{\lambda}{\mu}, f_\beta, N, \mathcal{P}\Big) = (1-\pi_N)p \tag{10}$$

Finally, the definitions of U and $\mathbf{E}[\beta|\beta \in \mathcal{D}]$ remain unchanged, while the expression for the average revenue rate in (8) can be simplified as below and expressed as an explicit function of P_{admit}:

$$R_\beta\Big(\frac{\lambda}{\mu}, f_\beta, N, \mathcal{P}\Big) = \frac{\lambda}{\mu}P_{admit}\mathbf{E}[\beta|\beta\in\mathcal{D}] \tag{11}$$

A particular SI admission strategy is the AA policy introduced in Section I that admits every slice request regardless of the associated bid (i.e., $\mathcal{D} = [\beta_m, \beta_M]$ and $p = 1$), such that, according to (7), $\mathbf{E}[\beta|\beta \in \mathcal{D}] = \mathbf{E}[\beta]$.

C. OPTIMAL POLICY AND COMPLEXITY

In Section I, we motivated the maximization of the average revenue rate as the main InP's objective, therefore, we seek the solution $\boldsymbol{\mathcal{P}}_{opt}$ of the following maximization problem:

$$\begin{aligned} &\boldsymbol{\mathcal{P}}_{opt}\left(\frac{\lambda}{\mu}, f_\beta, N\right) = \arg\max_{\boldsymbol{\mathcal{P}}} R_\beta\left(\frac{\lambda}{\mu}, f_\beta, N, \boldsymbol{\mathcal{P}}\right) \\ &\mathcal{P}_n : \mathcal{D}_n \subset [\beta_m, \beta_M],\ n \in \mathbb{N} < N \end{aligned} \tag{12}$$

The problem highlights the dependency of the optimal policy on λ/μ, f_β and N, therefore, the InP has to compute $\boldsymbol{\mathcal{P}}_{opt}$ offline for values of λ/μ and f_β that are representative of SPs' behavior in its network in order to adopt convenient strategies accordingly[2].

In order to define the search space for the optimal policy, we remind that, according to (1), the admission interval of a generic state policy $\mathcal{P}_n$ can be any subset of the bid interval. Hence, the admission interval can be generically represented as the composition of multiple disjoint admission intervals[3]. However, in order to reduce the complexity of the problem described in (12), we propose the adoption of AT policies where an admission threshold $\dot{\beta}_n$ is set at state n, such that $\mathcal{D}_n = [\dot{\beta}_n, \beta_M]$ and $\dot{\beta}_n \geq \beta_m$. Accordingly, the system policy $\boldsymbol{\mathcal{P}}$ can be fully described by the *threshold vector* $\dot{\boldsymbol{\beta}} = (\dot{\beta}_0, \quad \cdots, \quad \dot{\beta}_{N-1})$ in the SD case and by the scalar $\dot{\beta}$ in the SI case, respectively. Thus, the search space for the optimal policy is reduced, and the problem in (12) can be transformed into an N-dimensional or mono-dimensional continuous optimization problem for SD and SI policies, respectively. On the other hand, a reduction in the achieved revenue rate is expected when compared to the optimal policy. However, as we demonstrate in the next section, the relative loss remains constrained with respect to different load regimes. Please refer to Appendix A for the performance metrics' expressions adapted for AT policies, and note that AA policies can be considered as a particular case of SI AT policies with threshold $\dot{\beta} = \beta_m$.

In order to further improve the tractability while conserving accuracy, we convert the problem into a combinatorial optimization problem by discretizing the sample space $[\beta_m, \beta_M]$ into a finite number h of intervals. Hence, the thresholds that can be used for the state policies' definition are:

$$\dot{\beta}_n = \beta_m + j\frac{(\beta_M - \beta_m)}{h}, \quad j \in \mathbb{N} < h \tag{13}$$

and the choice of a suitable value of h guarantees results' accuracy while keeping computational costs at acceptable levels, as it is demonstrated in the following section. Please find in Appendix A the combinatorial version of the problem presented in (12) adapted for AT policies, whose solution will be referred to as *optimal AT policy* in the following.

We remind that, as introduced in Section I, the objective of this work is to propose a prompt admission control mechanism for network slicing in 5G and to compare its performance with that of baseline solutions. Because proposed AT policies enable admission strategies at reduced complexity, we adopt in this study an exhaustive search of the optimal policy for demonstration purposes only, leaving for future extensions the search of a more computational efficient method. Fixed the size of the pool of resources N, the complexity of an exhaustive search for the optimal AT policy in SD and SI systems is polynomial (i.e., $\mathcal{O}(h^N)$) or linear (i.e., $\mathcal{O}(h)$), respectively, with regards to the discretization levels h. Note that, depending on the value of h, multiple solutions of the problem may exist, and, in those cases, we choose the solution that maximizes P_{admit}; that is, the solution that minimizes the Euclidean norm of the threshold vector (i.e., $||\dot{\boldsymbol{\beta}}||_2$ or $\dot{\beta}$ for SD and SI systems, respectively).

V. SYSTEM SETUP AND RESULTS EVALUATION

In this section, we present and compare the performance of different slice provision mechanisms for both on-demand and periodic slicing when different policies are employed. For the system setup, we examine different pool of resources and the extreme case where SPs follow a per-UE slicing strategy. In the case of small cells, according to [23] up to 5 simultaneously active UEs can be served, hence, we assume a maximum number of slices $N = 6$. For the traffic model, we consider low, medium and high arrival rates λ, ranging from 0.5 to 100. On the other hand, we adopt only one service class with exponentially distributed departures and unitary average service rate μ. The bid interval varies within the range $[\beta_m, \beta_M] = [0, 100]$ representing, respectively, the minimum tariff accepted by the InP and the SPs' budget. Finally, we provide results for the case where SPs make uniform bids over the admitted interval (i.e., $\beta \sim \mathcal{U}[\beta_m, \beta_M]$).

For the solution of the combinatorial problem for AT policies associated to the problem described in (12), we employ a number h of discretization levels for the bidding region that ranges from a minimum of 2 (i.e., *low* and *high bid region*) up to a maximum of $h = 10$, allowing a higher precision. Besides, we develop a tool in Matlab for the performance evaluation of the different considered mechanisms. In particular, for the case of on-demand slicing with uniformly distributed bids, AT and AA performance is evaluated according to the expressions introduced in Appendix A. On the other hand, for periodic slicing, a simulator generates instances of the request arrivals, tenants' departure and bidding processes, and enforces AT, FCFS and BB policies accordingly for different slicing intervals. Finally, we remind that the optimal AT policy is computed by means of exhaustive search, and, in the periodic case, it is obtained separately for different values of the slicing interval $T_{slicing}$.

In the remaining of this section, first we focus in on-demand slicing, computing the optimal AT policy and comparing SD and SI approaches, when AA policy is used as a benchmark. Lastly, for the periodic case, we study the

[2] The InP can estimate the SPs' traffic patterns using network tracing, and employ traffic forecasting mechanisms [20]–[22] together with machine learning tools for adapting the strategy on-the-fly.

[3] i.e., $\mathcal{D}_n = \bigcup_i \mathcal{D}_n^i$, with $\mathcal{D}_n^i = [\beta_m^i, \beta_M^i] \subset [\beta_m, \beta_M]$ and $\mathcal{D}_n^i \cap \mathcal{D}_n^j = \emptyset$, $\forall i \neq j$.

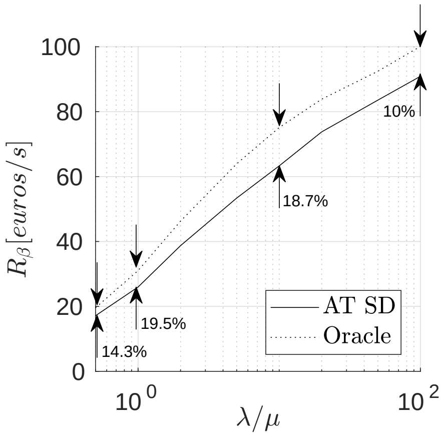


FIGURE 4. Assessment of the revenue loss for AT policy with respect to an ideal Oracle, when SD systems are considered, $N = 1$, and $h = 10$.

optimal AT policy for different slicing intervals, and we compare the performance with that of FCFS and BB policies.

A. ON-DEMAND SLICING EVALUATION

In Section IV-C, we anticipated that a reduced-complexity solution to the problem introduced in (12) exists in the form of AT policy with discretized thresholds, but this approach may suffer some penalty on the revenue. Consequently, we now study the limits of its performance by comparing the average revenue rate of the optimal AT policy with that of an ideal tool we named *Oracle*. In particular, in this context, we consider the most flexible type of AT policy, that is, the SD approach with maximum definition over the bid interval (i.e., $h = 10$). Oracle, on the other hand, is capable of recognizing the most rewarding bids. Oracle is applied a posteriori (i.e., once the simulation is finished) and, therefore, it can apply admission decisions based on its full knowledge of all the events in the simulation (i.e., slice requests, tenants' departures and bids). Hence, Oracle is only used for benchmarking purposes as it cannot be implemented in practice.

In Fig. 4, we present the average revenue rate for both optimal AT policy and Oracle with respect to the load regime (i.e., λ/μ) in logarithmic scale. To this aim, we study the most resource-limited case (i.e., $N = 1$), which leaves AT policies with the least flexibility in terms of resource availability, for counterbalancing Oracle's knowledge of future events. We remind that InPs aim at the joint maximization of admission rate and prioritization of highest bids and that, according to Section IV-A, resources are exhausted (i.e., $\pi_N \approx 1$) in high-load regime (i.e., when $\frac{\lambda}{\mu} >> N$). Consequently, when a larger pool of slice requests is received by InPs, the latter are motivated to adopt a more selective admission criterion by raising the bid threshold, which leads to a revenue enhancement at the expense of the admission probability (i.e., according to (5) it holds $P_{admit} \approx 0$). It can be observed from the figure that both Oracle and AT policies can achieve a logarithmic increase with respect to $\frac{\lambda}{\mu}$. On the other hand, a loss in revenues is expected with respect to Oracle, as raising the admission threshold translates in revenue maximization in the long term, while Oracle is capable of selecting best bids over each realization of the slice request process. The graph shows that the loss in revenues remains bounded for any load regimes, and, in particular, a 14.3% loss is experienced when few revenue opportunities are available (i.e., $\frac{\lambda}{\mu} \to 0$), it increases to 19.5% when arrivals are N times the departures (i.e., $\frac{\lambda}{\mu} \approx N$), while it reduces for high-load regimes (i.e., $\frac{\lambda}{\mu} >> N$). For instance, AT policies undergo a loss in revenue of 10% when $\frac{\lambda}{\mu} = 100$. Therefore, AT policies offer a sub-optimal but viable solution to the generic optimization problem represented in (12).

Before comparing the optimal strategies in SD and SI systems, we study the influence of discretization over the complexity of the optimization problem and the accuracy of results. In particular, in order to study the feasibility of adopting an exhaustive search for benchmarking analysis, we provide the computation times associated to an exhaustive search of the optimal AT policy in our system setup for an infrastructure capable of hosting up to six slice tenants (i.e., $N = 6$). To this aim, we employ an Intel(R) Core(TM) i9-7900X CPU @ 3.30GHz with 64GB of RAM, and results reveal that when $h = 4$, 1.8 ms are necessary for a SI system against the 4.5 ms for a SD system. Besides, when $h = 10$, 6.6 s are necessary for a SI system against the 64.5 minutes for SD systems. Therefore, within the considered system setup, computation times remain limited for both systems, although SI systems are preferable when big infrastructures are being studied, and when many combinations of $\frac{\lambda}{\mu}$ and f_β have to be considered for modeling SPs' behavior.

Comparing the performance accuracy for SD and SI systems, we represent in Fig. 5 the average revenue rate offered by AT policies when different discretization levels h are used. The figure proves that both systems react the same way to discretization, except for some specific values of h showing very small differences in revenue due to the lower degrees of freedom of SI systems. For instance, for $N = 6$ and $h = 8$, a 1.2% difference in revenue rate can be observed between the two systems. Besides, a floor exists for R_β when a minimum number of discretization levels h is used, or, in other words, that a solution to the problem described in (12) can be sought in the discrete domain with no significant performance loss when a suitable accuracy is adopted. In particular, the constraint on h is approximately independent of the size of the resource pool (i.e., N), however, it is more evident in high-load regimes, as a better granularity allows a more rewarding bid selection over a bigger pool of service requests. For instance, according to Fig. 5, InPs may decide to apply a minimum number of discretization levels equal to $h = 2$ and $h = 4$ when $\lambda/\mu = 0.5$ and $\lambda/\mu = 100$, respectively, in order to jointly minimize complexity and the loss in revenue opportunities. However, in the following, we adopt $h = 10$ for a better graphical detail.

In order to study the behavior of SD and SI systems adopt-

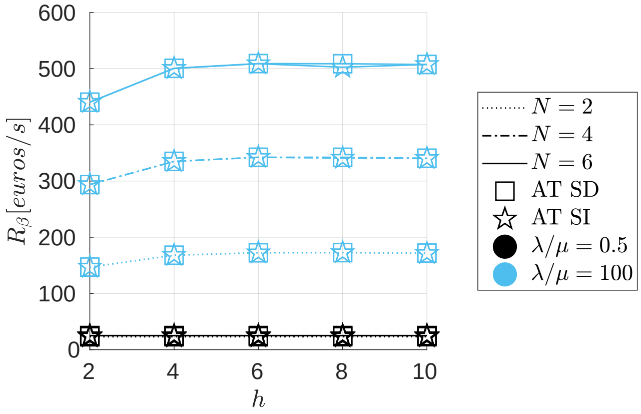


FIGURE 5. Average revenue rate for SD and SI AT policies with respect to the discretization granularity h, for different values of N, and λ/μ.

ing AT policies under different load regimes and systems sizes, we represent in Fig. 6a and 6b the optimal policies for both solutions, when $h = 10$ discretization levels are used for all values of N and λ/μ. When comparing the two graphs, it can be observed that, independently of the load regime λ/μ and of the size of the resource pool N, similar AT policies are optimal for SD and SI systems. In particular, in low-load regime (i.e., $\lambda/\mu = 0.5$), the low arrival rate of service requests and the small holding time of slice tenants encourage the InPs to adopt in both systems low admission thresholds, thus, maximizing revenues by increasing the admission probability. On the other hand, in high-load regime (i.e., $\lambda/\mu = 100$), the system is saturated (i.e., $\mathbf{E}[n] \approx N$) and suffers from resource scarcity due to the high arrival rate of slice requests and the big holding time of slice tenants. Hence, InPs are motivated to increase the admission threshold in order to block the less rewarding slice requests.

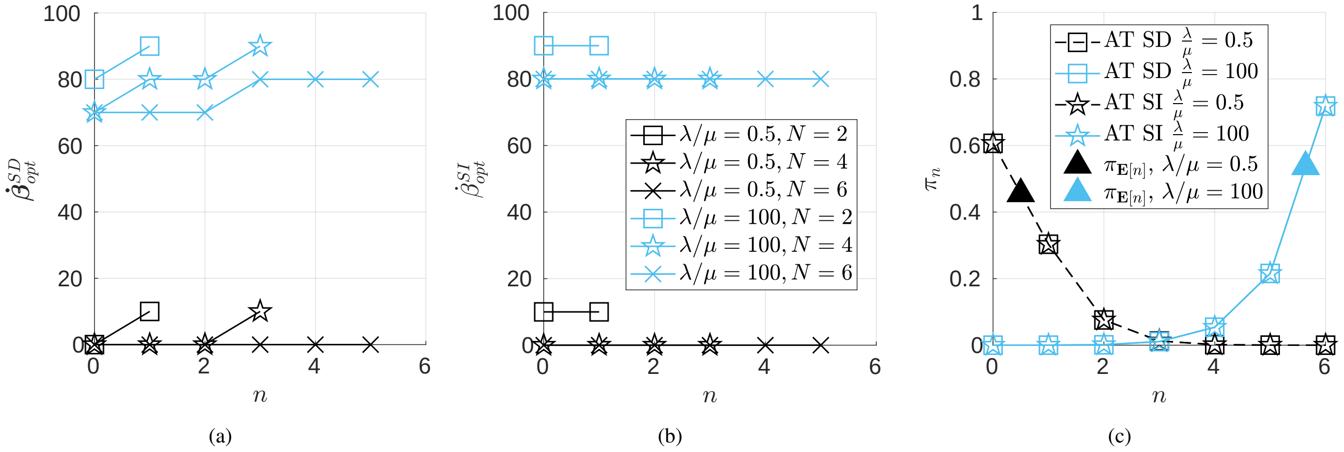
In both load regimes, the higher flexibility of SD systems enables step-like policies, where lower admission thresholds are adopted when the system is far from saturation, while higher ones are employed when the system is about to exhaust its resources. Moreover, with increasing size of the resource pool N, SD systems tend to be less selective by relaxing the policy when far from saturation, in order to achieve a better balance between admission probability and revenue rate. Despite different strategies can be generally considered optimal for SD and SI systems, it can be noted that the difference in the admission thresholds adopted at each state n is, at most, equal to the discretization step (i.e., $|\dot{\beta}_n^{SD} - \dot{\beta}_n^{SI}| \leq (\beta_M - \beta_m)/h, n \in \mathbb{N} < N$). Therefore, independently from the load regime and the pool of resources, the optimal policy for the two approaches leads to the same system behavior, on average, that is, to the same stationary probabilities π_n, as illustrated by Fig. 6c for the case $N = 6$. This aspect, in turn, translates into a close performance matching, as demonstrated below.

After having computed the optimal admission thresholds for on-demand AT policies, we now compare the performance of SD and SI approaches with that of an AA policy when different load regimes and pools of resources are considered. In particular, in Fig. 7, we study the admission probability P_{admit}, the average revenue rate R_β and the average resource utilization U when $h = 10$ bid levels are used. Firstly, it can be observed that, by enforcing the constraint on the discretization accuracy (i.e., $h \geq 4$), a close performance match can be obtained between SD and SI approaches not only for the average revenue rate but also for the other performance metrics. This result holds independently from the load regime λ/μ and the size of the resource pool N.

In low-load regime (i.e., $\lambda/\mu = 0.5$) it can be observed that the performance metrics of different admission strategies (i.e., AT or AA) are very close and tend to coincide when big resource pools are considered. Indeed, due to the limited revenue opportunities, AT strategies imitate the behavior of the AA approach by admitting as many requests as possible (see Fig. 6a and 6b), resulting in high admission probabilities (Fig. 7a). However, in resource-limited systems (i.e., $N = 2$), the higher flexibility of SD approaches is capable of guaranteeing a slightly higher admission probability when compared to SI strategies. At the same time, due to the low rate of service requests, the average number of instantiated slices (i.e., $\mathbf{E}[n]$) remains approximately constant, independently from the size of the pool of resources (i.e., N). Therefore, according to (6), the average resource utilization decreases with respect to N (Fig. 7b), while the average revenue rate does not vary (Fig. 7c).

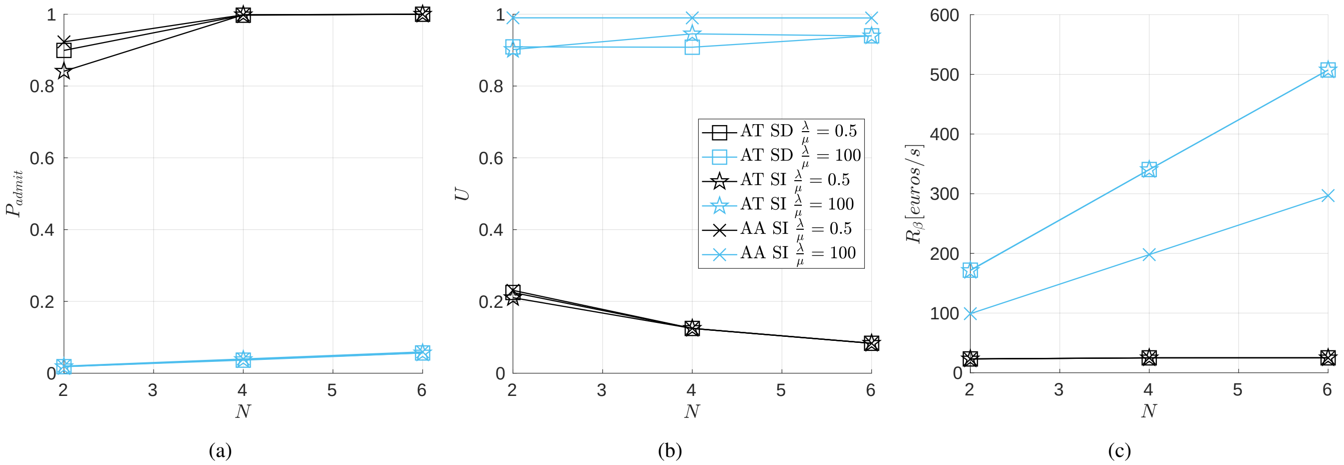
In high-load regime (i.e., $\lambda/\mu = 100$), the average admission probability decreases with respect to the low-load regime for AT policies in both SD and SI systems (Fig. 7a). However, as results coincide with those for the AA policy, this is not the consequence of the adoption of higher admission thresholds in AT policies, but rather of the limited resources with respect to the demand. Consequently, both the admission probability and the average operational expenditures (i.e., $\mathbf{E}[n]$) increase linearly with the size of the resource pool N, as more resources can be accessed by competing SPs. Therefore, according to its definition in (6), the average resource utilization U remains approximately constant with respect to N (Fig. 7b). However, the more restrictive admission strategy of AT policies is demonstrated by a slightly lower utilization when compared to AA policy, especially for SD systems due to their greater flexibility. Likewise, because of the higher revenue opportunities, the revenue rate is higher than the one achievable in low-load regime and increases linearly with respect to the resource pool size N, as represented in Fig. 7c. Besides, due to the higher admission thresholds, AT policies are capable of admitting the most rewarding slice requests and consistently offer much higher revenue rates when compared to the AA strategy (i.e., 68.6% improvement).

In conclusion, AT policies provide a great advantage in terms of revenue rate and resource utilization while conserving the admission probability of less restrictive strategies,



(a) (b) (c)

FIGURE 6. Optimal AT policy $\dot{\boldsymbol{\beta}}_{opt}$ in a) SD and b) SI systems with different N and λ/μ, and c) stationary probabilities π_n in SD and SI systems with $N = 6$ and different λ/μ. Besides, it is represented the interpolation of π_n corresponding to the average state $\mathbf{E}[n]$ (i.e., $\pi_{\mathbf{E}[n]}$). $h = 10$ in all the graphs.



(a) (b) (c)

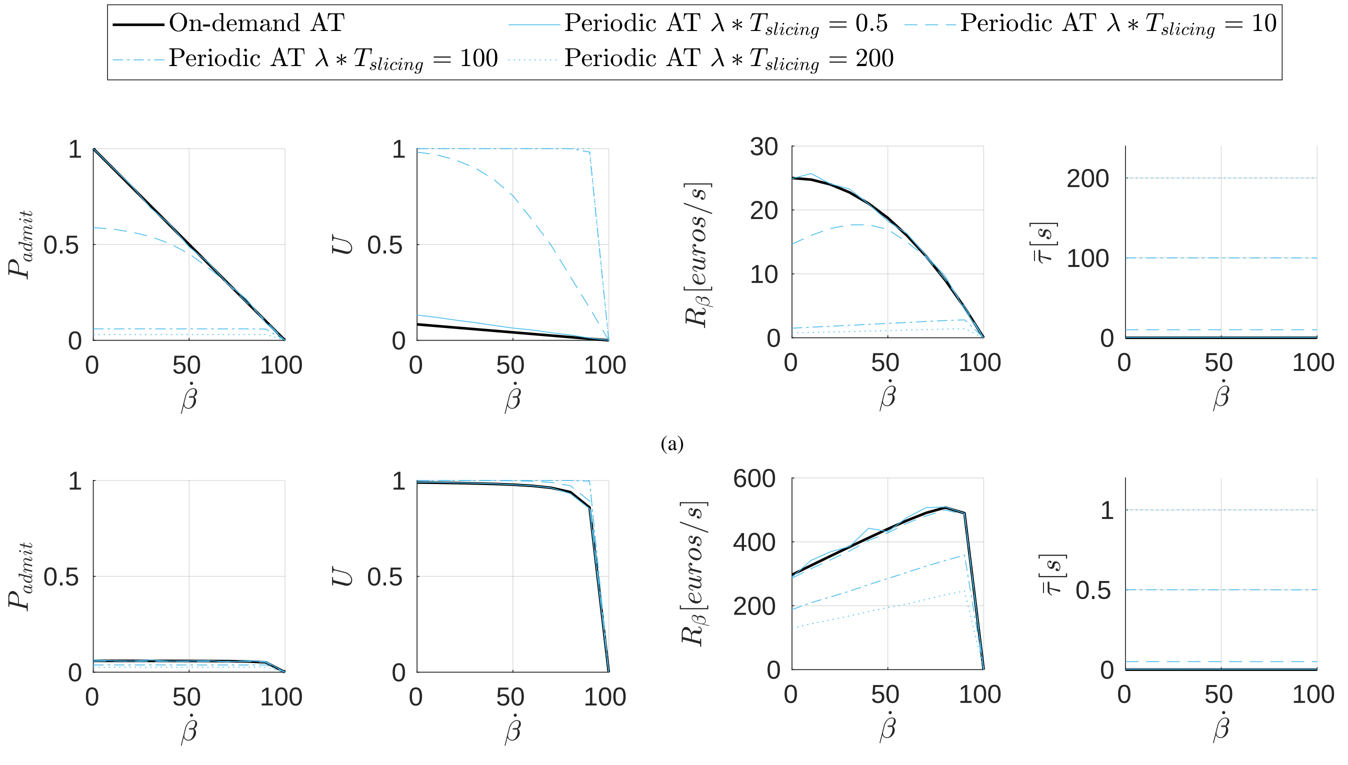
FIGURE 7. Performance of on-demand systems: a) admission probability, b) average resource utilization and c) average revenue rate. SD and SI AT policies with $h = 10$ and AA policies are compared.

such as the AA policy. Besides, when sufficient accuracy is adopted for the bid interval discretization (i.e., $h \geq 4$), SI AT policies are reduced complexity solutions of the problem represented in (12) when compared to SD policies, at the expense of a slightly lower admission probability for resource-limited systems.

B. ON-DEMAND AND PERIODIC SLICING COMPARISON

In the remaining of this section, we first compare the performance of on-demand and periodic slicing mechanisms when AT policy is adopted. Afterwards, the comparison is extended to reference admission control strategies (i.e., the AA policy in on-demand case and the FCFS and BB policies in the periodic case). The analysis introduced in Section IV can be extended to the periodic case by using discrete-time Markov chains (DTMCs), where transitions among states take place at regular time intervals. Therefore, P_{admit}, U, R_β and the optimal AT policy $\dot{\boldsymbol{\beta}}_{opt}$ become dependent on the slicing interval $T_{slicing}$. In this context, extending the model introduced in Section III, n represents the number of slices instantiated and reserved during a given slicing interval, considering also those tenants that fulfilled their SLA within the considered interval (i.e., tenants leaving the system and interrupting their contribution to InPs' revenues). Therefore, the definition of U in (6) takes on a connotation of average resource reservation for periodic slicing, however, for the sake of comparability, we maintain same name and symbol as for on-demand slicing. As shown in previous paragraphs, both SD and SI AT strategies can be utilized for this comparison when sufficient discretization accuracy is guaranteed, thus, in the following, we consider only SI policies due to the lower complexity needed for computing the optimal policy.

In Section I, we highlighted that, once policies are defined, the promptness of a specific slice admission method strictly depends on the delay added by the communication flow between SPs and InPs and the complexity for computing the admission decision. In order to provide a complete comparison between on-demand and periodic systems, we introduce in this context a new performance metric measuring the delay



(a)

(b)

FIGURE 8. Performance of on-demand and periodic slicing with respect to the admission threshold $\dot{\beta}$ when SI AT policies are adopted and: a) $\lambda/\mu = 0.5$, b) $\lambda/\mu = 100$. $N = 6$ and $h = 10$ are considered in all the graphs, and, in the periodic case, performance metrics are estimated over $N_{slicing} = 10000$ slicing intervals.

added by the admission control mechanism. In particular, we define the average waiting time $\bar{\tau}$ as the average time delay from service request arrivals, up to their admission or blockage. For on-demand slicing, it holds $\bar{\tau} = 0$ because, according to Section I, slice requests are evaluated right upon arrival. On the other hand, in periodic slicing, $\bar{\tau}$ is the average time interval between slice request arrivals and the beginning of next slicing interval. Therefore, exploiting the properties of Poisson processes, the instants t_a corresponding to slice requests arrivals within the *k-th* slicing interval are uniformly distributed (i.e., $t_a \sim \mathcal{U}[kT_{slicing}, (k+1)T_{slicing}]$, with $k \in \mathbb{N}_0$). Hence, $\bar{\tau} = \mathbf{E}[T_{slicing} - t_a] = T_{slicing}/2$ independently from the adopted policy. With respect to the computation of the admission decision, both AA and FCFS strategies introduce null delay, as they only enforce the admission decision whenever resources are available. Assuming that the optimal admission thresholds are pre-computed for different values of λ/μ, f_β, and N, the same holds for AT policies. Finally, the BB admission mechanism implies the implementation of sorting algorithms with higher computational expenses than previous strategies, however, as better processors are made available every year, we assume that the dominant component of the total delay is $\bar{\tau}$ for all the analyzed strategies.

In order to compare how AT policies behave in on-demand and periodic strategies, we analyze how the performance metrics vary with respect to the admission threshold $\dot{\beta}$ defined in (13) and slicing interval $T_{slicing}$. In particular, in Fig. 8, we provide the representation of the admission probability P_{admit}, the average resource utilization U, revenue rate R_β, and waiting time $\bar{\tau}$ for the whole range of admission thresholds and slicing intervals defined in the system setup. On the other hand, without loss of generality, only a fixed system dimension is considered (i.e., $N = 6$). Finally, Fig. 8a and 8b illustrate the cases with low and high-load regimes (i.e., $\lambda/\mu = 0.5$ and $\lambda/\mu = 100$), respectively.

With respect to the system's fairness P_{admit} and the utilization of resources U, it can be observed from Fig. 8 that both are monotonically decreasing functions of $\dot{\beta}$, for every load regime and admission strategy (i.e., either on-demand or periodic). Therefore, a global maximum exists for both performance metrics over the admitted bid interval and it coincides with the most permissive threshold (i.e., $\dot{\beta} = 0$), while they tend to decrease when less permissive strategies are enforced. Besides, periodic slicing provides same performance as on-demand slicing when a small number of arrivals takes place per slicing period (i.e., $\lambda T_{slicing} = 0.5$). On the other hand, when slices are offered less frequently than the service rate (i.e., $T_{slicing} \geq 1/\mu$), the number of SPs competing within the same slicing interval increases, and a higher optimal AT threshold is adopted. Accordingly, the admission probability decreases, and the resource reservation deviates from the resource utilization of the on-demand case. Note that for very high values of $\lambda T_{slicing}$ the level of saturation is comparable to that of on-demand slicing mechanisms in case of high-load regimes (i.e., $P_{admit} \to 0$ and $U \to 1$).

On the other hand, R_β manifests different behavior and shows a global maximum depending on the load regime and slicing strategy. When the number of competing SPs is low (i.e., $\lambda/\mu = 0.5$ in the case of on-demand slicing, joint to $T_{slicing} < 1/\mu$ for the periodic slicing case), R_β is a monotonically decreasing function of $\dot{\beta}$. As limited revenue opportunities exist, the unconditional admission (i.e., $\dot{\beta} = 0$[4]) outperforms any other admission criterion. However, when the load regime increases in on-demand slicing, or when lower slicing frequencies are adopted in periodic slicing (i.e., $T_{slicing} \geq 1/\mu$), the competition among SPs increases and R_β becomes a concave function of $\dot{\beta}$. We remind that InPs have the joint objective of maximizing the admission rate and the resulting revenue, hence, when slice requests exceed the resource availability, on the one hand, revenue opportunities increase, on the other hand, the resources tend to be exhausted. Therefore, an optimal admission threshold exists as a tradeoff between the maximization of the admission rate and the prioritization of the most rewarding requests. To confirm what we just said, independently from the load regime, the horizontal coordinate that maximizes R_β corresponds to a value of P_{admit} not too far from its maximum. Besides, the optimal AT threshold also reduces U with respect to its maximum, thus limiting the operational expenditures while guaranteeing maximum revenue. Finally, it is confirmed that the average waiting time $\bar{\tau}$ is null for on-demand slicing, while it increases with respect to the slicing interval for periodic slicing (i.e., $\bar{\tau} = T_{slicing}/2$).

After having studied how the performance metrics vary with respect to the adopted threshold and to the enforced slicing interval, we analyze now the properties of the optimal AT policy for on-demand and periodic cases. In particular, in Fig. 9, we represent $\dot{\beta}_{opt}$ as a function of $\lambda T_{slicing}$, while considering different resource pool sizes (i.e., $N = 2$, $N = 4$, and $N = 6$), as well as low and high-load regimes (i.e., $\lambda/\mu = 0.5$ and $\lambda/\mu = 100$). First, we can observe that, for small values of $\lambda T_{slicing}$, the optimal AT policy for periodic slicing is well approximated by the one for on-demand slicing for all pools of resources and load regimes. Indeed, the high slicing frequency makes periodic slicing systems receive fewer slice requests per slicing interval, thus approximating the behavior of on-demand slicing. Besides, we can observe how, for increasing number of arrivals per slicing interval (i.e., $\lambda T_{slicing}$), the optimal AT policy for periodic slicing becomes more selective than in the on-demand case, tending to the maximum admitted threshold for every λ/μ and N.

In order to benchmark the optimal AT policy in both the on-demand and periodic cases, we compare its performance with that of reference slicing mechanisms. In particular, in the on-demand case, we consider the AA policy that admits all slice requests, independently from the associated bids, whenever resources are available. Note that, in the case of inelastic slices only, AA coincides with the admission strategy proposed in [9]. On the other hand, in the periodic case, we study the adaptation of AA to discrete time case, which operates as a FCFS policy within a given slicing interval. Finally, for periodic slicing we also provide comparison with the BB policy that, within a given slicing interval, admits requests with highest bids up to resource exhaustion. Hence, in Fig. 10, we represent the admission probability P_{admit}, the average resource utilization U, the average revenue rate R_β, and the average waiting time $\bar{\tau}$ as a function of $\lambda T_{slicing}$. The comparison is performed over the whole range of slicing intervals according to the system setup, while, without loss of generality, only a fixed system dimension is adopted (i.e., $N = 6$). Besides, low, medium and high-load regimes (i.e., $\lambda/\mu = 0.5$, $\lambda/\mu = 10$ and $\lambda/\mu = 100$) are illustrated in Fig. 10a, 10b and 10c, respectively.

First, it can be observed how, in on-demand slicing, AT always outperforms AA in terms of offered revenues and resource utilization at the cost of a small loss in admission probability. Besides, AT and AA policies for on-demand slicing act as best-case scenario for their natural extensions to periodic slicing, that is, periodic AT and FCFS policies, respectively. In particular, FCFS well approximates the AA performance for low values of $\lambda T_{slicing}$, while it provides worse performances for less frequent slicing (i.e., $T_{slicing} \geq 1/\mu$).

Observing into more detail the performances of different periodic slicing schemes, periodic AT proves to be more selective and resource efficient than the other two policies, in the sense that it is characterized by a slightly lower admission probability and by the reservation of less resources for the revenue maximization. Besides, FCFS represents the lower bound in terms of revenue rate with respect to periodic AT and BB policies. Indeed, for low values of $\lambda T_{slicing}$, BB behaves like a FCFS policy, while periodic AT improves revenues by rejecting low bids and keeping resources for future requests with higher bids. On the other hand, when sufficient service requests are received within a given slicing interval, BB outperforms the unconditional admission of FCFS and tends to the revenue rate offered by the periodic AT policy. Finally, for slicing intervals greater than one tenth of the service time (i.e., $T_{slicing} \geq 0.1/\mu$), periodic AT and BB offer comparable revenue rates. The effectiveness of the most rewarding policies (i.e., periodic AT and BB) is emphasized when high values of λ/μ are explored, that is, when more revenue opportunities exist. On the other hand, independently from the adopted policy, the admission probability decreases and the resource utilization increases inevitably due to the limited resources with respect to the demand. With respect to the average waiting time $\bar{\tau}$, it is null for on-demand strategies and for very frequent slicing (i.e., $T_{slicing} \approx 0$), while it increases linearly with $T_{slicing}$ for periodic slicing (i.e., $\bar{\tau} = T_{slicing}/2$), regardless of the analyzed mechanism.

In conclusion, a slicing system that employs the optimal

[4] We highlight that, even though a null threshold is enforced, positive revenue rates are possible, on average, as SPs' behavior is modeled according to a uniform random bid distribution.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2939935, IEEE Access

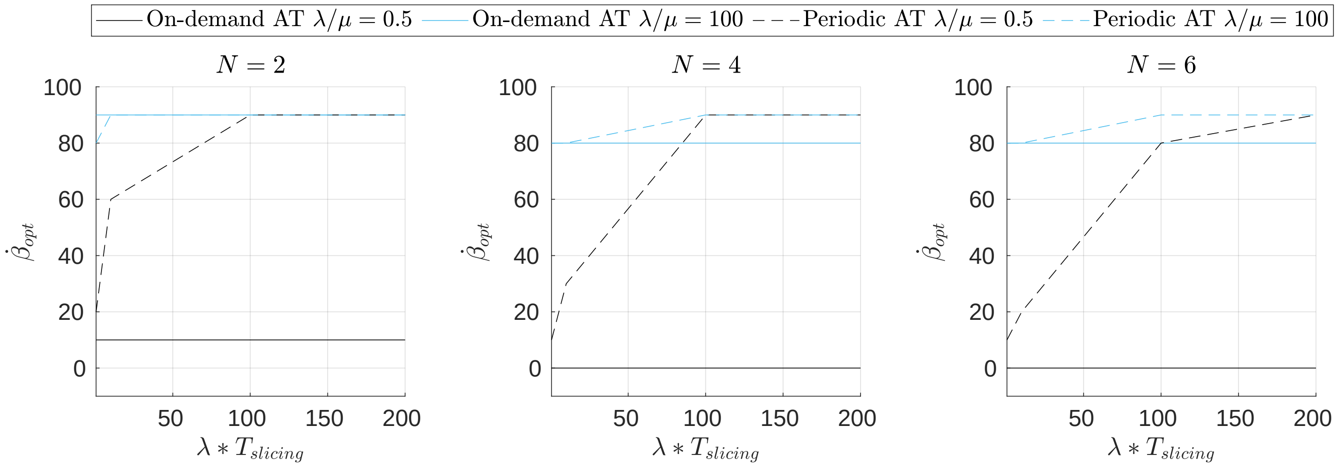


FIGURE 9. Optimal threshold $\dot{\beta}_{opt}$ for on-demand and periodic slicing when SI AT policy is used and different values of N are considered. $h = 10$ is considered in all graphs and, in the periodic case, the optimal threshold is computed over $N_{slicing} = 10000$ slicing intervals.

AT admission policy (with respect to load regime, bid distribution and pool of resources) outperforms all the considered reference mechanisms, either on-demand or periodical. Indeed, it offers the highest revenue rate and smallest resource utilization, with a negligible loss in terms of admission rate. Besides, on-demand slicing solutions enable null waiting time for slice requests.

VI. CONCLUSIONS

In this work, we proposed a slice provision mechanism for enabling the slice market envisioned for 5G. The proposed approach consists in a policy that selects the most rewarding bids offered by SPs (i.e., AT policy), and a reduced complexity solution is provided for adapting the optimal policy to different resource pool sizes, traffic loads and SPs behavior. We demonstrated that it enhances the slice provision promptness, with QoS guarantees and fairness towards SPs, while guaranteeing two-fold economic incentives to InPs: revenue maximization and reduction of operational expenditures. Besides, we presented a comparison of the proposal's performance with reference strategies, both when enforced upon every service request (i.e., on-demand slicing) or at regular time-intervals (i.e., periodic slicing). In particular, we adopt always-admit policy (i.e., AA) in on-demand slicing, and first-come-first-served (i.e., FCFS) and best bid (i.e., BB) policies in periodic slicig.

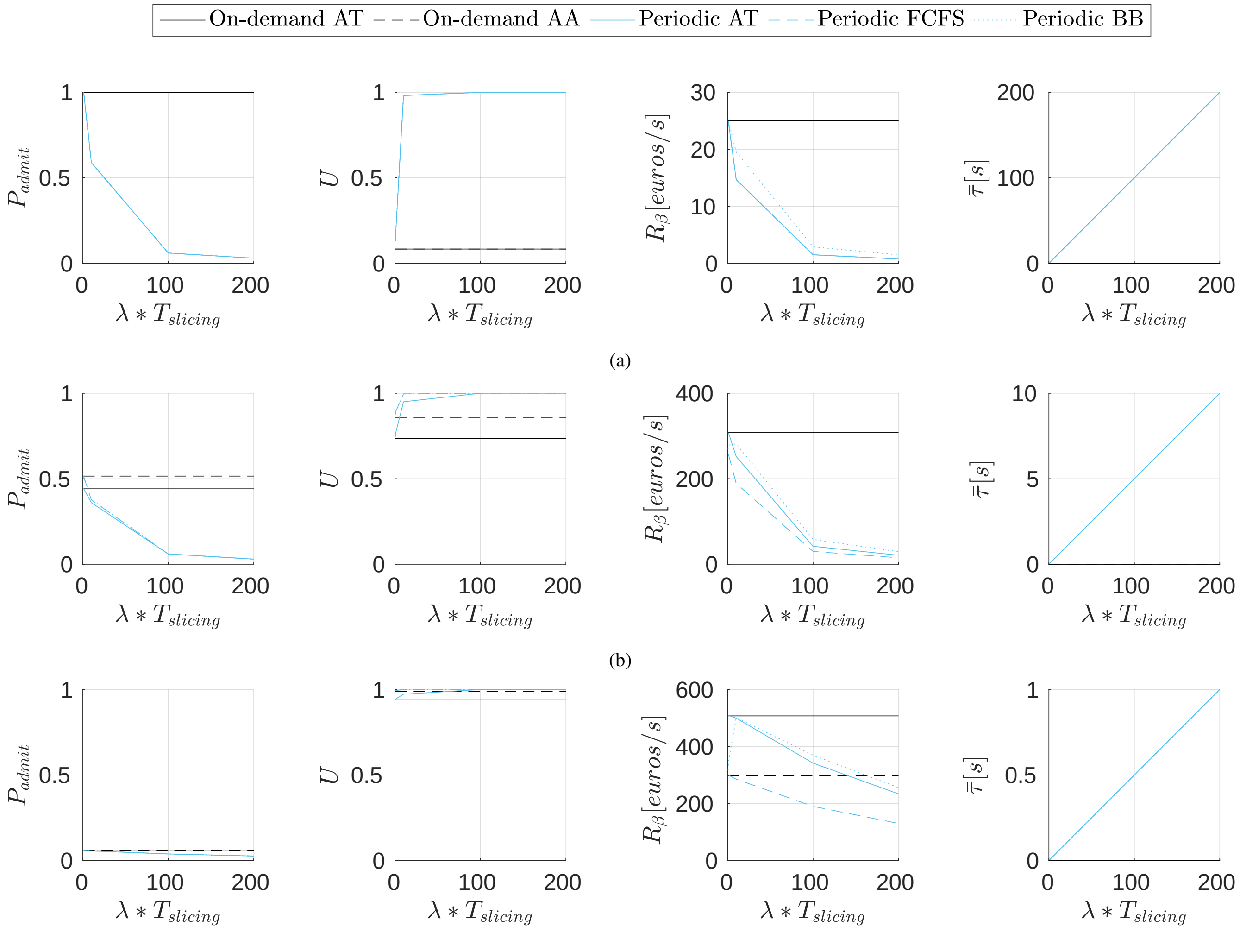
Provided that the optimal bid threshold is chosen for the current network conditions, the proposed AT policy in on-demand slicing outperforms the other considered mechanisms, including a best bid selection strategy for periodic slicing. Our optimal AT policy offers the highest revenue rates while reducing operational expenditures and offering real-time slicing, in exchange for a negligible loss in terms of fairness towards SPs. On the other hand, if only periodic slicing is possible, AT policy still offers the same advantages, however, slice requests experience a waiting time different from zero, which is independent from the adopted strategy and it decreases with the slicing frequency. Finally, AT approaches enable reduced complexity solutions when compared to other strategies, such as the BB policy. The effectiveness in terms of revenues is highlighted especially in systems characterized by limited resources and high-load regimes. In our future studies, we plan to include the case with elastic services, when different service classes are examined. Besides, we consider modeling SPs as fully rational entities that adapt their bidding strategies to their perception of the market. Finally, computational efficient methods for the search of optimal admission policies will be examined and proposed for the integration in real systems.

.

APPENDIX A ON-DEMAND SLICING WITH AT POLICY

We adapt here to the case of AT policies the expressions provided in Sections IV-A and IV-B for the performance metrics of on-demand slicing, and of the combinatorial version of the optimization problem in (12) for such policies. In particular, as the InP admits slice requests at state n only when the tariff-bid is higher than threshold $\dot{\beta}_n$, the admission probability at state n is $p_n(f_\beta, \dot{\beta}_n) = 1 - CDF(\dot{\beta}_n)$. It is straightforward that p_n is a monotonically decreasing function of $\dot{\beta}_n$ as $\frac{dp_n}{d\dot{\beta}_n} = -f_\beta(\beta) \leq 0$. Besides, for the most conservative and permissive admission strategies it holds, respectively, $p_n(f_\beta, \beta_m) = 1$ and $p_n(f_\beta, \beta_M) = 0$.

The admission probability P_{admit}, the average resource utilization U and the average revenue rate R_β remain unchanged. On the other hand, the expected tariff-bid for tenants admitted at state n equals $\mathbf{E}[\beta|\beta \geq \dot{\beta}_n] = \frac{1}{p_n}\int_{\dot{\beta}_n}^{\beta_M} \beta\, f_\beta(\beta)\, d\beta$, that is a non-negative function of $\dot{\beta}_n$ (i.e., according to Leibniz's integral rule $\frac{d\mathbf{E}[\beta|\beta \geq \dot{\beta}_n]}{d\dot{\beta}_n} = \frac{f_\beta(\beta)}{p_n}\Big(\mathbf{E}[\beta|\beta \geq \dot{\beta}_n] - \dot{\beta}_n\Big) \geq \dot{\beta}_n \frac{f_\beta(\beta)}{p_n}\Big(\frac{1}{p_n}\int_{\dot{\beta}_n}^{\beta_M} f_\beta(\beta)\, d\beta - 1\Big) = 0$). For the most conservative and permissive admission strategies it holds $\mathbf{E}[\beta|\beta \geq \beta_m] = \mathbf{E}[\beta]$ and $\mathbf{E}[\beta|\beta \geq$



(a)

(b)

(c)

FIGURE 10. Performance metrics for network slicing with respect to $\lambda T_{slicing}$ when SI AT and AA policies are adopted for on-demand approaches and AT, FCFS and BB policies for periodic approaches. For the periodic case, the optimal threshold is calculated for each value of $\lambda T_{slicing}$, over $N_{slicing} = 50000$ slicing intervals, besides results are provided for: a) $\lambda/\mu = 0.5$, b) $\lambda/\mu = 10$ and c) $\lambda/\mu = 100$. $N = 6$ and $h = 10$ are considered in all graphs.

$\beta_M] = \beta_M \geq \mathbf{E}[\beta|\beta \geq \beta_m]$, respectively.

In the particular case of uniformly distributed bids, it holds for AT policies $p_n = \frac{\beta_M - \dot{\beta}_n}{\beta_M - \beta_m}$, $\mathbf{E}[\beta|\beta \geq \dot{\beta}_n] = \frac{\beta_M + \dot{\beta}_n}{2}$ and for AA policy $p = 1$, $\mathbf{E}[\beta|\beta \geq \dot{\beta}] = \mathbf{E}[\beta] = \frac{\beta_M + \beta_m}{2}$. Finally, the average revenue rate for the three policies can be written as:

$$R_\beta^{SD} = \frac{1}{2}\frac{\lambda}{\mu}\frac{1}{\beta_M - \beta_m}\sum_{n=0}^{N-1}\pi_n(\beta_M^2 - \dot{\beta}_n^2)$$

$$R_\beta^{SI} = \frac{1}{2}\frac{\lambda}{\mu}(1 - \pi_N)\frac{\beta_M^2 - \dot{\beta}_n^2}{\beta_M - \beta_m}$$

$$R_\beta^{AA} = \frac{1}{2}\frac{\lambda}{\mu}(1 - \pi_N)(\beta_M + \beta_m)$$

In conclusion, we introduce the combinatorial version of the problem described in (12) for AT policies:

$$\dot{\boldsymbol{\beta}}_{opt}\left(\frac{\lambda}{\mu}, f_\beta, N\right) = \arg\max_{\dot{\boldsymbol{\beta}}} R_\beta\left(\frac{\lambda}{\mu}, f_\beta, N, \dot{\boldsymbol{\beta}}\right)$$

$$\dot{\beta}_n = \beta_m + j\frac{(\beta_M - \beta_m)}{h}, n, j \in \mathbb{N}\, \text{and}\, n < N, j < h$$

ACKNOWLEDGMENT

This work was supported in part by the EU Horizon 2020 Research and Innovation Programme (5GAuRA) under Grant 675806, and in part by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya.

REFERENCES

[1] 3GPP, TR 22.891, "Feasibility Study on New Services and Markets Technology Enablers; Stage 1," Rel. 14, Dec. 2018.

[2] 3GPP, TS 22.261, "Service requirements for the 5G system; Stage 1," Rel. 16, Dec. 2018.

[3] 3GPP, TS 28.530, "Aspects;Management and orchestration; Concepts, use cases and requirements," Rel. 15, Dec. 2018.

[4] 3GPP, TR 28.801, "Study on management and orchestration of network slicing for next generation network," Rel. 15, Jan. 2018.

[5] 3GPP, TR 21.915, "Release 15 Description; Summary of Rel-15 Work Items," Rel. 15, Feb. 2019.

[6] 3GPP, TS 23.501, "System Architecture for the 5G System; Stage 2," Rel. 15, Dec. 2018.

[7] 3GPP, TS 28.527, "Life Cycle Management (LCM) for mobile networks that include virtualized network functions; Stage 2," Rel. 15, Jun. 2018.

[8] Ying Zhang, "SDN and NFV in 5G," in *Network Function Virtualization: Concepts and Applicability in 5G Networks*, Hoboken, NJ, USA: John Wiley &amp; Sons, Inc., Jan. 2018.

[9] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing," presented at the 2017 IEEE Conf. on Comput. Commun. (IEEE INFOCOM 2017), Atlanta, GA, 2017, pp. 1-9.

[10] U. Habiba and E. Hossain,"Auction Mechanisms for Virtualization in 5G Cellular Networks: Basics, Trends, and Open Challenges," in *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2264-2293, third quarter 2018.

[11] B. Han, J. Lianghai and H. D. Schotten, "Slice as an Evolutionary Service: Genetic Optimization for Inter-Slice Resource Management in 5G Networks," in *Access, IEEE*, vol. 6, pp. 33137-33147, 2018.

[12] G. Sun et al., "Coalitional Double Auction for Spatial Spectrum Allocation in Cognitive Radio Networks," in *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, pp. 3196-3206, Jun. 2014.

[13] P. Lin, X. Feng, Q. Zhang and M. Hamdi, "Groupon in the Air: A three-stage auction framework for Spectrum Group-buying," in *Proc. 2017 IEEE Conf. on Comput. Commun. (IEEE INFOCOM 2017)*, Turin, Apr. 2013, pp. 2013-2021.

[14] Y. Zhang, S. Bi, and Y. J. A. Zhang, "A two-stage spectrum leasing optimization framework for virtual mobile network operators," presented at the 2016 IEEE Int. Conf. on Commun. Syst. (ICCS), Dec 2016, pp. 1-6.

[15] P. Caballero, A. Banchs, G. de Veciana and X. Costa-Pérez, "Network slicing games: Enabling customization in multi-tenant networks," presented at the 2017 IEEE Conf. on Comput. Commun. (IEEE INFOCOM 2017), Atlanta, GA, 2017, pp. 1-9.

[16] M. Morcos, T. Chahed, Lin Chen, J. Elias and F. Martignon, "A two-level auction for C-RAN resource allocation," presented at the 2017 IEEE Int. Conf. on Commun. Workshops (ICC Workshops), IEEE, Paris, May. 2017, pp. 516-521.

[17] A. Jarray and A. Karmouch, "Decomposition Approaches for Virtual Network Embedding With One-Shot Node and Link Mapping," in *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 1012-1025, Jun. 2015.

[18] F. Esposito, D. Di Paola and I. Matta, "On Distributed Virtual Network Embedding With Guarantees," in *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 569-582, Feb. 2016.

[19] B. Han, D. Feng and H. D. Schotten, "A Markov Model of Slice Admission Control," in *IEEE Netw. Letters*, vol. 1, no. 1, pp. 2-5, March 2019.

[20] J. Perez-Romero, J. Sanchez-Gonzalez, O. Sallent, R. Agusti, "On Learning and Exploiting Time Domain Traffic Patterns in Cellular Radio Access Networks," in *Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science*, vol. 9729, Springer, Jun. 2016.

[21] U. Paul, L. Ortiz, S. R. Das, G. Fusco, M. Madhav Buddhikot, "Learning Probabilistic Models of Cellular Network Traffic with Applications to Resource Management," 2014 IEEE Int. Symp. on Dyn. Spectr. Access Netw. (DYSPAN), McLean, VA, 2014, pp. 82-91.

[22] Y. Zang, F. Ni, Z. Feng, S. Cui, Z. Ding, "Wavelet Transform Processing for Cellular Traffic Prediction in Machine Learning Networks," 2015 IEEE China Summit and Int. Conf. on Signal and Inf. Process. (ChinaSIP), Chengdu, 2015, pp. 458-462.

[23] METIS-II, "Performance evaluation framework," Deliverable D2.1, no. January, p. 10, 2016. [Online]. Available at: https://metis-ii.5g-ppp.eu/wp-content/uploads/deliverables/METIS-II_D2.1_v1.0.pdf

MATTEO VINCENZI is a Ph.D. student at the Department of Network Engineering of Universitat Politècnica de Catalunya (UPC). He received his M.Sc. Degree in telecommunications engineering from the University of Bologna in 2014. He worked as a researcher at Mavigex S.r.l., and he participated as Early Stage Researcher (ESR) in the European funded project Application-aware User-centric pRogrammable Architectures for 5G multi-tenant networks (5GAURA), an Innovative Training Network (ITN) of the Marie Skłodowska-Curie Actions (MSCA). His main research interests are in the area of programmable wireless communication systems, network slicing, network sharing.

ELENA LOPEZ AGUILERA is an Associate Professor and a member of the Wireless Networks Group (WNG) in the Networks Engineering Department of Universitat Politècnica de Catalunya, BarcelonaTech (UPC). She received her M.Sc. degree in telecommunications engineering from the UPC in 2001, and her Ph.D. in 2008. Her main research interests include the study of IEEE 802.11 WLANs, Internet of Things enabling technologies in heterogeneous scenarios, and 5G networks. Her experience also comprises QoS, security, radio resource management, location mechanisms, and wake-up radio systems.

EDUARD GARCIA-VILLEGAS received his M.Sc. and Ph.D. degrees from the Technical University of Catalonia, BarcelonaTech (UPC) in 2003 and 2010, respectively. He is an associate professor at the same university and a member of the Wireless Networks Group (WNG). He participates in the activities of the IEEE P802.11 WG as a voting member. He also participates in the research developed within the i2CAT Foundation. His research interests include IEEE 802.11 WLANs, radio resource management in wireless networks, IoT enabling technologies and 5G architecture.