IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 8, 2020, accepted November 24, 2020, date of publication November 26, 2020,
date of current version December 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040949

Resource Allocation for Network Slicing in Mobile Networks

ALBERT BANCHS[1,2], (Senior Member, IEEE), GUSTAVO DE VECIANA[3], (Fellow, IEEE), VINCENZO SCIANCALEPORE[4], (Senior Member, IEEE), AND XAVIER COSTA-PEREZ[4,5,6], (Senior Member, IEEE)

[1]Department of Telematics Engineering, University Carlos III of Madrid, 28911 Leganes, Madrid
[2]IMDEA Networks Institute, 28918 Madrid, Spain
[3]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, USA
[4]NEC Laboratories Europe, 69115 Heidelberg, Germany
[5]i2cat Foundation, 08034 Barcelona, Spain
[6]ICREA, 08010 Barcelona, Spain

Corresponding author: Albert Banchs (banchs@it.uc3m.es)

The work of Albert Banchs was supported in part by the H2020 5G-TOURS European project under Grant 856950, and in part by the Spanish State Research Agency (TRUE5G project) under Grant PID2019-108713RB-C52/AEI/10.13039/501100011033. The work of Gustavo de Veciana was supported by NSF Grant CNS-1910112.

ABSTRACT This paper provides a survey of resource allocation for network slicing. We focus on two classes of existing solutions: (*i*) reservation-based approaches, which allocate resources on a reservation basis, and (*ii*) share-based approaches, which allocate resources based on static overall shares associated to individual slices. We identify the requirements that a slice-based resource allocation mechanism should satisfy, and evaluate the performance of both approaches against these requirements. Our analysis reveals that reservation-based approaches provide a better level of isolation as well as stricter guarantees, by enabling tenants to explicitly reserve resources, but one must pay a price in terms of efficiency unless reservations can be updated very dynamically; in particular, efficiency falls below 50% when reservations are performed over long timescales. We provide further comparisons in terms of customizability, complexity, privacy and cost predictability, and discuss which approach might be more suitable depending on the network slices' characteristics. We also describe the additional mechanisms required to implement the desired resource allocations while meeting the latency and reliability requirements of the different slice types, and outline some issues for future work.

INDEX TERMS Mobile networks, network slicing, beyond 5G, resource allocation.

I. INTRODUCTION

NETWORK SLICING FOR 5G

Beyond supporting tight requirements in terms of latency, reliability and throughput, 5G incorporates profound changes in architectural design. One of the key novel concepts is *network slicing*, which enables the infrastructure to be 'divided' into several *logical slices*. Each slice can invoke (virtual) network functions running on the common infrastructure, and tailor them to meet its specific requirements [1], [2]. In this way, slices can be customized to support specific mobile services [3], providing far more flexibility than RAN sharing approaches in 4G networks [4]. The network slicing framework has the potential to address the complexity of managing diverse multi-service requirements, but it is critical that this is achieved cost-effectively through efficient sharing of network resources.

The associate editor coordinating the review of this manuscript and approving it for publication was Zihuai Lin.

NETWORK SLICING MODEL

Network slicing makes room for new players in the mobile network ecosystem, formalizing the separation between *infrastructure providers* (which provide communication and computational resources) and *network slice tenants* (which acquire slices to provide services to their customers). This model is analogous to that introduced in cloud computing, where Infrastructure as a Service (IaaS) providers make available computational resources such as CPU, disk or memory to the tenants. Network slicing is geared at enabling an ecosystem akin to the cloud compute business model. However,

providing network resources to support mobile services is an intrinsically different problem to that in cloud computing, since (*i*) radio resources can be particularly scarce, making over-provisioning extremely costly, and (*ii*) we cannot assign *any* radio or edge compute resource to a user indistinctly, since users may need to be served by nearby nodes.

SLICE-BASED RESOURCE ALLOCATION FOR MOBILE SERVICES

Given the dynamic nature of mobile user loads, the 5G system calls for novel approaches to enable slice-based dynamic management and allocation of resources across the network. Each tenant will typically enter into a Service Level Agreement (SLAs) with the infrastructure provider which ideally (*i*) allows network slice tenants to manage the performance of their customers, while (*ii*) enabling the infrastructure to achieve economies of scale by multiplexing the traffic of multiple network slices.[1] Then slice-based allocations resulting from the SLAs will then be translated to specific customer-level allocations through typically more complex mechanisms involving scheduling at user and slice level.

This paper presents a survey of resource allocation for network slicing, analyzing and comparing the existing approaches for resource distribution across slices. The focus of this paper is on resource allocation approaches that decide the amount of resources to be allocated to each slice. This involves the allocation resources such as radio or edge computing with location constraints, where a user needs to be allocated resources from a neighboring node or base station, and these cannot be exchanged with resources from other nodes. Beyond the approaches studied here, complementary mechanisms are required to schedule the resources of each node while meeting the specific service requirements such as URLLC (Ultra Reliable Low Latency Communications) or mMTC (massive Machine Type Communications). Such low-level resource allocation schemes are not the main focus of this paper, and are discussed in Section VII.

While there are some other surveys in the literature about network slicing (see, e.g., [5]), their focus is rather on the architectural principles and enabling technologies; in contrast, the focus of this paper is on the resource allocation models, comprising the criteria to allocate resources among slices as well as the implications on several fronts: architectural, pricing, performance, etc. The papers in [6], [7] have a similar focus to ours, however their contribution is mostly limited to presenting the possible resource allocation models for network slicing, while our emphasis is on the analysis the advantages and performance of the different approaches, going into substantially more depth. On another front, [8] reviews the different problems that need to be addressed for network slicing; this paper focuses on one of these problems (namely, resource allocation), providing a much deeper insight on the possible solutions that may be adopted for this problem.

The key contributions of this paper are as follows:

- We identify the key requirements that a resource allocation mechanism should satisfy and the functionality that it should provide. Our analysis of the requirements for network slicing is novel and, to the best of our knowledge, deeper than previous analyses in the literature.
- We present two classes of resource allocation mechanisms proposed in the literature. While there are papers in the literature focusing on the operation of individual mechanisms, our description here goes beyond the operation details and addresses the underlying fundamental concepts such as the involved timescales or the sharing gains.
- We evaluate each of the two classes of mechanisms against the requirements identified earlier, showing the advantages and disadvantages of each approach and presenting both quantitative results and qualitative arguments. We are not aware of any such analysis in the literature.
- Based on our analysis and results, we provide a comparison of both approaches and discuss in which scenarios it may be more suitable to rely on reservation-based approaches and which ones are better suited for share-based approaches.
- We discuss the mechanisms at the different levels that would be required to implement the resource allocations for network slicing and we identify some issues for future work.

II. SLICE-BASED RESOURCE ALLOCATION: KEY REQUIREMENTS

We shall begin by introducing the key requirements for the design of slice-based resource allocation mechanisms.

CUSTOMIZABILITY

A key goal is to enable tenants to customize the resource allocation and functionality of their slices to meet the needs of their customers. For slices serving mobile users, we will typically have temporal load variations across different network nodes and thus it is important to tailor resource allocations to follow such variations. To this end, well-defined interfaces should be provided enabling tenants to dynamically adapt their slices' allocations to meet the spatiotemporal varying customer demands.

COMPLEXITY

The complexity and implementation overheads should be kept low. These overheads may arise due to excessive signaling associated with the dynamic reconfiguration of slices, their set up and tear down, as well as the computational costs to make such decisions. Note, however, that the complexity of slice-based resource allocation solutions should be traded off against the level of customizability.

[1]Note that the SLA for network slices needs to be abstract and at a high level to allow for an easy interface with the tenants.

EFFICIENCY

To be cost effective, the infrastructure provider will want to achieve a high utilization of the network's communication and compute resources. This translates to reduced capital and operational expenses and typically comes from flexible sharing, i.e., statistical multiplexing across the traffic of multiple slices.

ISOLATION

Most tenants will want a degree of protection and isolation that ensures that their SLAs will not be compromised by the behavior of other tenants. This is indeed one of the main features of the network slicing principle: each slice should be perceived as a 'virtual' network that is effectively 'isolated' from other slices on the network. Isolation has implications in terms of ***resource guarantees***, as it makes the resources provided to a slice independent of the other slices; this is essential for services such as URLLC which require very strict guarantees. Naturally, there is a tradeoff between isolation and efficiency, as the latter improves when relaxing isolation requirements.

PRIVACY

Since tenants sharing infrastructure resources may be competing with one another, it is important to minimize the leakage of sensitive information from one tenant to another. For example, a tenant should only be able to make coarse, if any, inferences of other tenants' customer demands and performance. Not unlike cloud computing services, privacy is tied to isolation and thus typically comes at an increased cost and/or loss of efficiency.

COST PREDICTABILITY

Tenants tend to prefer resource allocation models that lead to predictable costs. In cloud computing, this is typically done by providing a range of products over various timescales, where commitments over longer timescales typically result in lower costs to tenants. Similarly, in the context of mobile services, one would expect longer-term SLAs to provide more predictable cost models.

III. SLICE-BASED RESOURCE ALLOCATION APPROACHES

In the literature there are, broadly speaking, two classes of resource allocation approaches: share-based and reservation-based. The first class relies on tenants agreeing to share the overall network resources based on pre-agreed fixed shares. In the second class, tenants issue specific reservation requests for resources, which may be accepted or rejected by the infrastructure provider depending on resource availability.

Multiple schemes have been proposed for each of the above classes. To analyze their advantages and possible issues, we shall focus on two representative schemes capturing the salient features of each class:

- *Share-based approach* [7], [9]–[12][2]: each slice purchases an overall *network share*. This share can be understood as the 'budget' allocated to the slice, which the slide can distribute among the network's nodes (e.g., base stations or data centers). Then, the resources at each node are shared among the slices in proportion to their budget allocations at the node. Thus, the total resource allocation of a slice will ultimately depend on its share, allowing the slice to choose how to subdivide its share across nodes.
- *Reservation-based approach* [13]–[19]: each slice requests a certain amount of resources individually at each network node, and the infrastructure may accept or reject the request. In the former case, the infrastructure guarantees that the slice will be provided with the reserved resources as long as it needs them.

A critical characteristic underlying the above approaches is the timescales at which resource allocations are made or adjusted. Although the timescales may depend on the implementation of each specific mechanism, the following general remarks apply in the analysis of this paper:

- *Months/days timescale*: In the share-based approaches, the shares purchased by slices are typically considered to be rather static (e.g., they may depend on the monetary contribution of a network operator sharing the infrastructure with other operators). Thus, it is reasonable to assume that such slice shares are updated over long timescales, say on a monthly or daily basis.
- *Days/hours timescale*: In the reservation-based approach, reservations address the needs of a slice which typically issues the corresponding requests over a timescale that may span from hours to days. Note that performing reservations on shorter timescales would involve potentially heavy signaling in addition to complex admission control and resource allocation algorithms. In what follows, as well as in most approaches in the literature [17], [18], we assume that reservations are made on an hourly or daily basis.
- *Minutes/seconds timescale*: In the share-based approach, slices may vary their budget allocation across nodes on quite short timescales, within minutes or even less. Indeed, such operation only requires conveying the budget allocation from each tenant to the individual nodes, and the allocation is then performed locally at each node. This is a lower-complexity operation than that involved in making reservations. In line with similar approaches in the literature, such as SON [20], this can be performed on a minute or sub-minute basis.

Fig. 1 shows three example of how resources may be allocated between slices in a network with two slices and four users per slice, for different distributions of the users across nodes and under the following two approaches: (*i*) a share-based approach, where both slices have the same

[2]The work in [7] also discusses other resource allocation approaches in addition to the share-based approach.

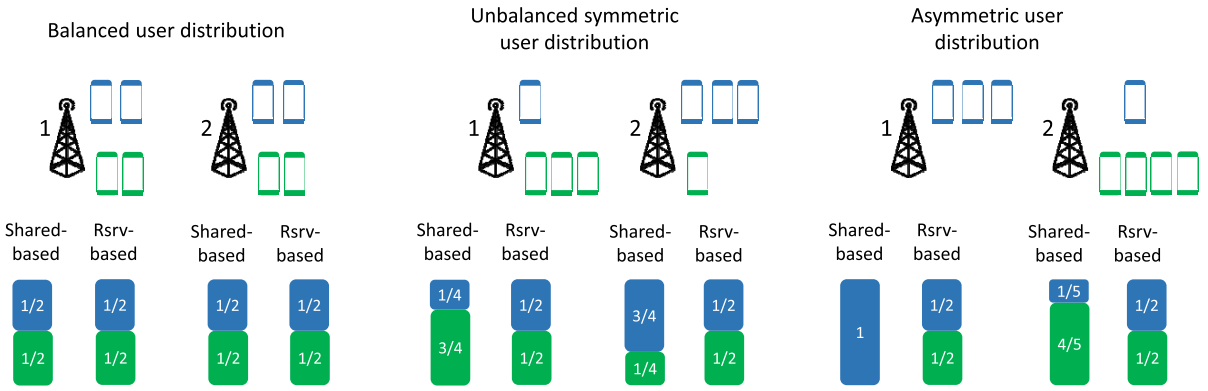


FIGURE 1. Examples of share-based and reservation-based allocations. There are two slices (green and blue). On the top, we show the number of users of each slice associated with each node. On the bottom, we show the fraction of each node's resources allocated to each slice for the each approach (shared-based and reservation-based).

network share and distribute their budget proportionally to the number of users at each network node, and (*ii*) a reservation-based approach, where both slices reserve half of the resources at each node. Note that, under the above considerations on timescales, it is reasonable to assume that (*i*) with the share-based approach, shares are allocated over long timescales (months/days) but the division of the share of a slice across nodes is performed at rather fast timescales (minutes/seconds) and can thus follow user loads, and (*ii*) with the reservation-based approach, reservations are performed at intermediate timescales (days/hours) and thus may not follow changes in user load in a timely fashion.

Ideally, we would like resource allocations to (*i*) provide a similar amount of resources to all users, given that both slices have the same share and the same number of users, and (*ii*) protect a slice from the (potentially greedy) behavior of the other slices. We observe from Fig. 1 that in the case of balanced user distributions, the two approaches provide the same allocation: they both share resources equally among all users, thus meeting the goals stated above. In the symmetric unbalanced case, under the share-based approach slices receive more resources at the nodes where they have more users, leading to a more even distribution of resources across users than the reservation-based approach (and thus providing a better overall allocation). Finally, in the asymmetric case, the performance of the blue slice is harmed by the green slice in the second node under the share-based approach, while the reservation-based approach provides more protection to the blue slice; indeed, under the share-based approach the user on the blue slice in the second node receives a small amount of resource due to the green slice being very unbalanced, while in the reservation-based approach, the blue slice is isolated from such behavior on the green slice.

The above example shows that, while share-based approaches may better adjust to current user load distribution, they also provide a smaller level of protection for the tenants. More broadly, Table 1 illustrates the main features of the share-based and reservation-based approaches in terms of the underlying resource allocation concept, their reaction to congestion, the timescales involved and the guarantees provided. As far as the underlying cost model is concerned, it is natural to assume that (*i*) under the share-based approach, tenants will be charged based on their share, while (*ii*) with reservation-based approaches, tenants will be charged based on the reservations they perform, following either a fixed pricing strategy (i.e., independent of the demand) or, alternatively, a dynamic one (i.e., demand-dependent). This has some implications on issues such as the cost predictability or the potential outage or unavailability of resources; note that in this paper we are only concerned on such fundamental issues, and not the specific business model or pricing strategy of the infrastructure provider (which is out of the scope of this paper).

The two resource allocation approaches studied in this paper are being considered in ongoing standardization efforts.

TABLE 1. Resource allocation approaches for network slicing.

	Share-based	Reservation-based
Allocation concept	Based on a fixed share assigned to each tenant	Based on the reservation requests issued by tenants
Reaction to congestion	All tenants see their resources reduced proportionally to their share	The requests of some of the tenants are not admitted into the network
Timescale	The overall shares are allocated over a long timescale, but the division of the shares among nodes may be adjusted on a minute timescale	Reservations are typically made on an hour or day timescale
Guarantees	Tenants are guaranteed their share of the overall resources	A request may not be admitted, but once admitted resources are guaranteed on a per-node basis
Cost model	Tenants will naturally be charged a cost that depends on their share	Tenants will typically be charged depending on their reservations of network resources

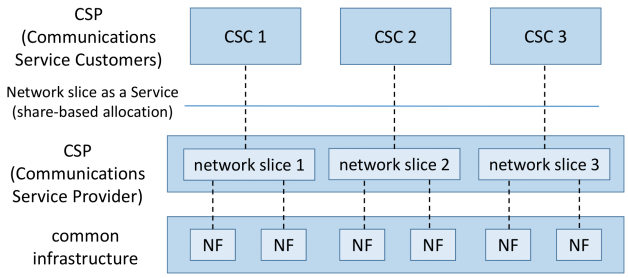


FIGURE 2. 3GPP management of share-based models.

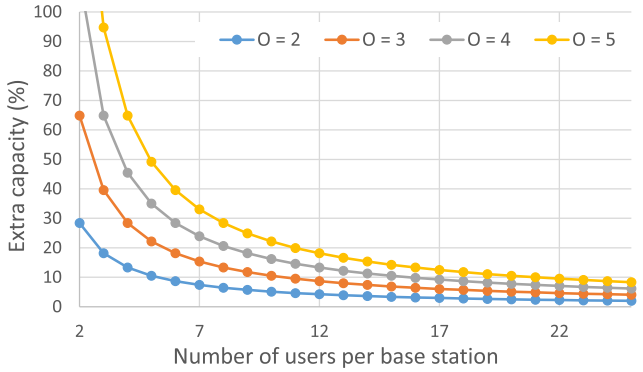


FIGURE 3. Capacity gains achieved by dynamic resource sharing: extra capacity (in %) that would be required by static slicing in order to provide the same performance as a dynamic share-based resource allocation, as a function of the number of users per base station and the number of slices (O). Source: [11].

In the 5G specifications [21], 3GPP has detailed the lifecycle management of a network slice through four different phases: (*i*) network slice preparation, (*ii*) installation, (*iii*) operation, and (*iv*) decommissioning. Once the network slice template is chosen, the infrastructure verifies whether the slice request can be accommodated and reserves the corresponding resources. In particular, network slice capacity planning and on-boarding procedures are performed in the first phase, slice resources are allocated and configured in the second phase, the supervision and performance monitoring is performed in the third phase, and resources are freed in the fourth phase. Such phases might be identified as part of a reservation-based approach, as they involve requests for resources reservations, their allocation, management and termination, respectively.

Beyond reservation-oriented operations, 3GPP also introduces a management model wherein different players may participate in the network slicing negotiation possibly following a share-based approach, as shown in Fig. 2. In particular, the communication service provider (CSP) may decide to offer a predefined network slice as an available service (namely network-slice-as-a-service) to multiple Communication Service Customers (CSCs), which may compete for the management of the slice resources in a share-based fashion. In turn, the CSCs may act as a CSP that offers its own services on top of the network slice instance.

In the next two sections, we analyze share-based and reservation-based approaches presented in this section in view of the requirements introduced in Section II.

IV. ANALYSIS OF SHARE-BASED APPROACHES

In the following, we provide an analysis of the share-based approaches against the requirements that should ideally be met by a resource allocation approach.

One of the key advantages of share-based approaches is their potential for improved ***efficiency***. When the slices' loads are time varying, it is desirable to have allocations which are dynamic and can adjust to such variations. A number of share-based schemes have been proposed in the literature [9]–[11] which allow for flexible and dynamic resource allocations. The benefits resulting from such dynamic resource slicing have received substantial attention (see, e.g., [11]). The gains achievable by sharing resources dynamically when demands are stochastic are illustrated in Fig. 3, which depicts the additional capacity required by a static partitioning of the resources (referred to as 'static slicing') to achieve the same performance as an optimal dynamic share-based scheme. We observe that such gains are substantial and grow when decreasing the cell load and increasing the the number of tenants, reaching 100% extra capacity in some cases.

The ***complexity*** of operating a network infrastructure under a share-based approach is relatively low. The network only needs to receive the budget distribution of each tenant and allocate the nodes' resources proportionally to the budgets. By contrast, the complexity on the tenant's side can be high, as the tenant needs to (*i*) decide the share needed to satisfy the service requirements, (*ii*) choose the budget distribution across nodes at each point in time, and (*iii*) possibly limit the number of customers to guarantee the service quality for active customers.

The ***overhead*** of a shared-based scheme is relatively low as well. It involves signaling the total budget of each node from the network to the tenants, and the budget distribution from the tenants to the network. While there may be several iterations in which tenants modify their budget distribution, these iterations can take place at a centralized controller, transferring the resulting allocation to the nodes afterward.

With share-based approaches, the network only ***guarantees*** tenants an overall share of the entire network. While a priori this does not provide guarantees at a node level, an important result reported in [10] shows that with share-based approaches, a tenant is guaranteed a better performance than with a static allocation of resources at each individual network node (i.e., 'static slicing'). This implies that we are actually providing tenants with some sort of guarantees in terms of node-level allocations. Furthermore, by limiting the number of users in the slice, tenants may leverage their overall share to realize statistical service guarantees for their users, as shown in [9]. Note that the such guarantees are only provided in terms of the overall resources for a user; to deliver guarantees on latency and reliability, a scheduling algorithm needs to be implemented by the tenant or the network in order to schedule such resources according to the desired guarantees.

The ***customizability*** of the share-based approach is enabled by letting tenants communicate their preferences to the infrastructure; this is done by dynamically subdividing the tenants' share or budget amongst the nodes. Such an approach has been widely studied in the context of economics and game theory, which refer to this as a Fisher market [22]; in such markets, buyers (in our case slices) have fixed budgets (in our case network shares) and bid for resources within their budget. The application of such a framework to share-based network slicing is developed in [10].

In terms of ***protection and isolation***, a priori share-based approaches provide a poor level of isolation, as the allocation of a tenant depends on the budget distribution of the other tenants, and hence may be affected by their behavior. However, the result mentioned above on the superior performance over a static allocation of resources implies some level of protection, as a tenant is guaranteed a better performance than a static allocation with perfect isolation.

When considering a distributed system such as the share-based one, ***stability*** is an important feature. Indeed, the manner in which a tenant distributes its budget amongst nodes may depend on the other tenants' distributions, and thus allocations could potentially bounce back and forth without converging. In [11], stability was studied in the context of elastic users which have concave utility functions, showing that for tenants supporting this type of users, an equilibrium exists and is reached when tenants selfishly maximize their own performance. In [9], a similar analysis was conducted for inelastic users with a minimum rate requirement. In contrast with the elastic users setting, in this case an equilibrium may not exist; moreover, even if an equilibrium exists, it may not be reached when tenants unilaterally optimize their performance.

In so far as ***privacy*** is concerned, the share-based approach leaks information to the tenants about the total budget allocated at each node, as this information is needed to enable tenants to estimate the implications of their budget allocation decisions. In a network with many tenants, the information about the nodes' total budgets may reveal very little sensitive information about individual slices. However, if there are only a few slices in the network, this information may allow a tenant to infer the spatial demands of other tenants, thus revealing potentially sensitive information.

Finally, one of the main strengths of the share-based approach is the ***predictability of the cost***. Indeed, the cost of this approach is typically tied to the share purchased by the tenant, which corresponds to a long-term contract and thus provides a highly predictable cost.

V. ANALYSIS OF RESERVATION-BASED APPROACHES

Next, we analyze the reservation-based approaches against the requirements introduced in Section II.

The ***efficiency*** of the reservation-based approach highly depends on the timescales of the reservations. Fig. 4 exhibits the efficiency of network slicing resource allocations measured with a real-world dataset [23] as a function of the reservation durations. Results are provided for different network levels ℓ, ranging from edge resources ($\ell = 1$) to central resources ($\ell = L$) and intermediate levels ($1 < \ell < L$). We observe that the loss of efficiency can grow as high as a factor of 10 for edge resources (efficiencies around 0.1 for $\ell = 1$) and a factor of 2 for cloud resources (efficiencies around 0.5 for $\ell = L$). The reason for this is that, when performing a reservation for a long period of time, we cannot adjust to the traffic dynamics and need to make the reservation for the peak demand during the period. As the reservation-based approach involves a fairly high complexity and tenants are not likely to be able to determine their needs on a very fine time granularity, reservations typically involve fairly long periods [17]. As a result, one may expect rather low efficiency when dealing with reservation-based approaches. This contrasts with the share-based approaches analyzed previously, which are expected to provide much higher efficiency at all network levels by re-allocating network resources more dynamically.

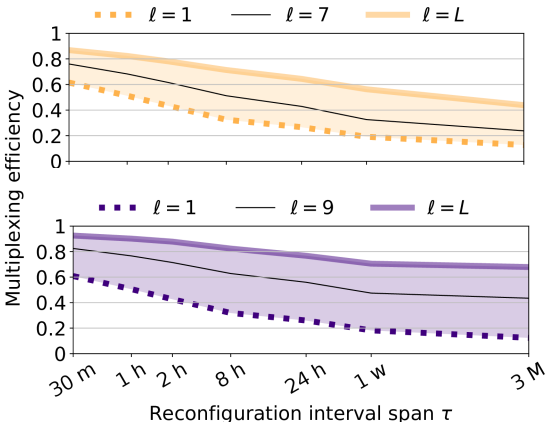


FIGURE 4. Efficiency of the reservation-based approach versus the resource reconfiguration periodicity τ. Dashed and solid colored lines denote the $\ell = 1$ (edge resources) and $\ell = L$ (central resources), while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city. Source: [23].

Reservation-based schemes typically involve a fairly high ***complexity*** on the network side. In order to provide the desired guarantees, complex admission control algorithm need to be implemented [17], coupled with traffic forecasting [18] along with some mechanisms to implement the resource reservations [16]. While machine learning approaches have been effectively used for these purposes [16]–[18], these solutions pose some issues in terms of learning time, computational resources, collection of data, etc. In addition to their complexity, reservation-based schemes also suffer from a fairly high signaling ***overhead***, involving both signaling between the network and the tenants to perform the reservations as well as signaling inside the network to convey in a timely fashion the information needed at the various points in the network.

The main strengths of the reservation-based approach are the ***guarantees*** provided to those requests that are admitted, along with the associated ***protection and isolation*** in the

usage of the reserved resources. Indeed, with this approach a tenant can reserve a fixed amount of resources at each node, which are guaranteed to the tenant independently of the demands of other tenants. In this way, full isolation is provided.

Reservation-based schemes provide a good level of ***customizability***: each tenant can reserve the desired allocation at each node and distribute the reserved resources among its users as it likes, thus enabling the provisioning of a customized service to each user. The level of customizability, however, is constrained by the timescales involved in reservations: as a tenant cannot efficiently perform a new reservation every time a user moves from one node to the other, resource allocations cannot be adapted to the current user distribution of a tenant, which harms customizability.

In terms of ***stability***, the reservation-based approach is stable by nature. Indeed, after a tenant issues reservation requests to satisfy its needs, regardless whether those are admitted or not, the tenant is not expected to take further actions. Thus, the system will not experience a chain of actions that puts its stability at risk.

As for ***privacy***, the only information leaked by the system corresponds to accepting or rejecting a reservation request. Based on this, a tenant may infer the demands of other tenants, specially when the tenants' aggregate demands at a node are high and force to reject reservation requests. However, to gather any meaningful data, a tenant would likely need to issue many (real or fake) requests, which would presumably be costly. Thus, one may consider that in practice reservation-based approaches offer a good level of privacy-preservation.

Finally, the ***cost predictability*** of reservation-based schemes will depend on the adopted business model. With fixed pricing, costs will be highly predictable, but requests may congest the network leading to rejecting incoming requests and thus making resource availability rather unpredictable. By contrast, by adopting a variable pricing approach, one may prevent congestion by increasing the prices; however, this leads to an unpredictable behavior in terms of cost.

VI. COMPARISON OF RESOURCE ALLOCATION APPROACHES

We next provide a comparison of the two approaches, share-based and reservation-based slicing, in terms of the requirements discussed in Section II. Table 2 presents a detailed discussion of each requirement, and the main conclusions are as follows:

- At a high level, the main difference between the two approaches is that reservation-based schemes provide ''hard'' service guarantees to admitted slices, but this comes at a price in complexity, efficiency and overheads.
- With reservation-based schemes, fairly complex mechanisms are run by the network; instead, share-based schemes can rely on rather simple algorithms on the network side, bringing part of the performance management complexity to the tenants.
- While reservation-based schemes provide protection by design, share-based schemes also provide some level of protection by ensuring that performance is at least as good as under static slicing.
- In terms of privacy and predictability of costs, both schemes are comparable.

From the above analysis, it follows that the key advantages of share-based approaches are high efficiency, low complexity and overhead, and cost predictablity, while the advantages of reservation-based approaches are harder guarantees and protection, more stability and better privacy. Thus, there is no clear winner between share-based versus reservation-based network slicing: the preferred option will depend on the choice of the economic and performance model, driven by business considerations, as well as other practical and engineering considerations:

- Let us consider 'large' tenants serving a substantial number of customers over a broad region. We posit that such large tenants, supporting many diverse and dynamic mobile users, will find it more attractive to buy a share of the overall network, while a tenant having a relatively small number of users or very localized traffic would likely find it more effective to reserve specific resources as needed.
- A tenant with unpredictable or time varying demands across a broad region might find a share-based scheme more cost-effective versus requiring constant changes in reservations or addressing its changing demands through over-provisioned reservations. A tenant with predictable demands could, by contrast, enter into proactive cost-effective agreements based on reservations upfront for its precise needs on different timescales.
- A tenant with very strict requirements, such as URLLC services, may prefer a reservation-based approach that ensures that its requirements will surely be met at all times. Indeed, while share-based approaches may be able to provide statistical guarantees, this comes at a price in terms of efficiency as a tenant needs to acquire an overprovisioned share and/or apply very strict policy to limit the number of users. In contrast, a tenant with more elastic demands may benefit from a share-based approach, which provides a better overall performance yet may punctually fail to meet performance demands.
- In terms of fairness, a large tenant competing for resources with other large tenants will want to ensure it is allocated resources fairly and not subject to fluctuations due to changing demands. A share-based scheme provides this type of fairness guarantee. By contrast, in a reservation-based scheme, a tenant may find its requests blocked by an admission control mechanism or find the current price out of line with the expected costs, yielding unfairness between tenants.

TABLE 2. Analysis of resource allocation approaches. Strengths and weaknesses of each approach are highlighted in green and red, respectively.

	Share-based	Reservation-based
Complexity	Simple for the infrastructure, which does not need to perform complex operations. Part of the complexity is brought to the tenants, which need to decide their share and budget allocation	Simple for tenants, which only need to issue reservations. Infrastructure needs to run potentially complex algorithms
Overhead	Signaling is required to bring the load information to a centralized location, and to bring the tenants' budget allocation back to the nodes	Beyond the signaling involved in issuing the requests, realizing the reservations also involves potentially heavy signaling
Efficiency	Dynamic resource sharing provides a high level of efficiency by adapting to varying tenants' loads	When relying on long-term reservations, efficiency is harmed
Protection and isolation	Tenants are guaranteed to perform better than static slicing which implies some form of protection	Protection is provided by the nature of these scheme, as service guarantees are ensured for a given reservation independent of the behavior of the other tenants
Guarantees	A tenant is guaranteed a share of the overall resources but receives no guarantees on a per-node basis; however, by limiting the number of customers, it can provide statistical guarantees	Tenants of admitted network slice requests are given absolute guarantees regarding the service they receive
Stability	In some cases, the budget allocations of the tenants may fluctuate when responding to each others' allocations	Admitted tenants see stable allocations
Privacy	Information is only disclosed about the total load of the nodes, but not on the individual load of each tenant	Tenants may infer very limited information on the overall load from admitted/rejected requests
Predictability of costs	The cost incurred by a tenant will typically be highly predictable, as it depends only on the share	The predictability of the cost highly depends on the pricing model applied by the infrastructure provider

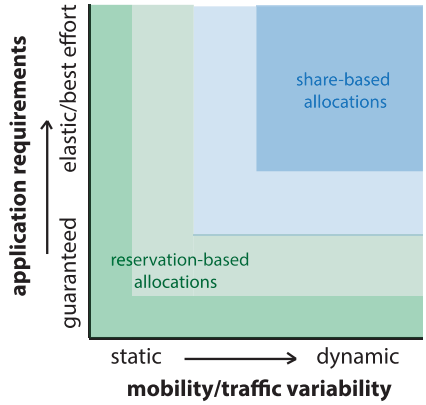


FIGURE 5. Suitability of share-based (blue) versus reservation-based (green) when considering traffic variability and application requirements. Darker areas refer to improved suitability of the approach.

Fig. 5 illustrates the suitability of reservation-based and share-based approaches along two of the dimensions discussed above, namely: (*i*) the variability of the traffic generated by a tenant, and (*ii*) the requirements of the tenants' applications, showing the region where each approach may be more suitable.

VII. REALIZING RESOURCE ALLOCATIONS WITH RESERVATION AND SHARE-BASED APPROACHES

While in this paper we have focused on the problem of deciding the resources to allocate to each slice, additional components need to be applied in conjunction with the techniques described here to implement the resources allocations while meeting the specific requirements of each slice. In the following, we present the key components of a network slicing architecture and discuss their relationship with the schemes presented in this paper.

DATA ANALYTICS AND FORECASTING

Since both the reservation-based and share-based approaches cannot reallocate resources at very short timescales, the allocations need to be performed some time in advance, which calls for forecasting algorithms that predict future demands based on the past load. 3GPP has included the data analytics modules needed to this end in its 5G architecture [24], [25] and a number of algorithms based on machine learning have been proposed in the literature [18], [19], [26]. Based on the load predicted by such a forecasting algorithm, network slices can issue the corresponding reservation requests under the reservation-based resource allocations, and the corresponding shares can be acquired under the share-based approach.

ADMISSION CONTROL FOR NETWORK SLICES

Under the reservation-based approach, the network infrastructure needs to determine whether a certain reservation of a network slice can be admitted while meeting the Service Level Agreement of the new slice as well as of the other slices being served. A number of algorithms have been proposed in the literature to this end (see, e.g., [17], [27]). Since slices may not always use their allocated resources, it may be possible to exploit multiplexing gains in order to improve the overall efficiency. Based on the outcome of such an admission control algorithm, the request of a network slice will be admitted or rejected in the reservation-based approach.

ADMISSION CONTROL FOR SLICE'S CUSTOMERS

A network slice aims at serving its customers, which may be, e.g., cars in a vehicular slice, sensors in an mMTC slice,

end-users in an eMBB slice, etc. It could happen that in some cases the resources of the slice, either in the share-based or the reservation-based approaches, do not suffice to satisfy the demands of all the slice's customers. To avoid this, a network slice tenant may opt for applying admission control to its customers, to ensure that the service received by admitted customers satisfies their demands and the required quality of service. In [9] an algorithm is proposed for admission control of end-users to network slices under the share-based approach.

DEALING WITH USER MOBILITY

Since the allocation of resources in different nodes may involve longer timescales than those corresponding to the mobility of users across nodes, the allocations need to account for user mobility. This affects both the reservation-based and the share-based approaches: when issuing a reservation request or when dividing the shares across nodes, a slice not only needs to account for the current distribution of its customers across nodes but also must account for the mobility of their customers. This has been studied in [17] for the reservation-based approach and in [12] for the share-based approach.

PLACEMENT OF VNFS

Network slices rely on virtualized network functions (VNFs) that may potentially be placed in different nodes depending on the slice's needs and the availability of resources at each node. Both for the reservation and for the share-based approaches, slices need to take into account the location of their VNFs when issuing the corresponding resource reservations and share allocations requests, respectively. In [28], this problem has been studied for the reservation-based approach.

ALLOCATION OF COMPUTATIONAL RESOURCES

As mentioned in the introduction, the resource allocation approaches discussed in this paper may be applied to computational resources as well as to radio resources. For the allocation of computations resources, there are a number of technologies ranging from virtual machines to containers which provide different features and also involve different timelines for the set up and reallocation of resources (which is referred to as scaling [29]). Both in the reservation and share-based approaches, the allocation of computational resources would need to be implemented with one of these technologies.

SCHEDULING OF RADIO RESOURCES

In addition to computational resources, the approaches discussed in this paper also deal with the radio resources. To this end, the high-level allocations resulting from our reservation and share-based approaches need to be mapped to the scheduling of radio frames. The scheduling needs to be performed such that, in addition to meeting the desired overall resource allocations, we also meet the specific requirements of each slice in terms of latency and reliability. Some scheduling algorithms in the context of network slicing have been proposed in the literature (see, e.g., [30]). In [31], the authors advocate for the usage of AI for RAN slicing.

VIII. OPEN ISSUES AND FUTURE WORK

In the following, we outline some open issues that need to be addressed in order to implement the approaches discussed in this paper. We further identify some potential lines of future work to address these open issues.

INTERFACES WITH NETWORK SLICE TENANTS

Slice tenants need to be able to convey to the infrastructure provider their demands and preferences in a simple way. Indeed, in order to satisfy their requirements, tenants need resources at different nodes of the network and this information needs to be provided to the infrastructure provider. At the same time, slice tenants may not have specific expertise on network operations and hence they require simple and intuitive interfaces. To the best of our knowledge, the definition of such an interface remains an open challenge both for reservation-based and share-based approaches.

ALGORITHMS TO ESTIMATE THE NEEDS OF A NETWORK SLICE

In order to determine the amount of resources required over a certain time period, network slices need to forecast their future demands (both for the reservation-based and the share-based approaches). While plenty of work has been conducted to forecast future Internet traffic [32], network slices' traffic typically comes from very specific applications with unique features and thus an analysis tailored to each particular traffic type is required. Such an analysis does not exist for many network slice types and is an open challenge.

END-TO-END RESOURCE ALLOCATION TIMELINES

As discussed throughout this paper, the time required to re-allocate resources at different nodes is crucial to the overall efficiency of network slicing. These timescales are constrained by the ability to perform up and down scaling for VNFs [33], the overhead associated to end-to-end resource allocation [34] and the capacity of network slice tenants to determine their needs at a fine time granularity. As shown in previous works [23], notable improvements in performance can be achieved by reducing the timescales of resource re-allocations.

MEETING EXTREME RELIABILITY AND LATENCY REQUIREMENTS

In 5G networks, some network slices may have extreme requirements in terms of reliability and/or latency [35]. This ultimately requires that sufficient resources be allocated to such network slices, so as to ensure that resources will suffice to (*i*) cover the demands at all times with a very high probability and (*ii*) schedule the frames of such slices providing very low latencies. To handle this in a reasonably efficient manner, we need forecasting schemes that can estimate the demands of such slices with great accuracy.

This is very challenging; indeed, even the most advanced schemes available in the literature cannot meet such extreme requirements [26], [36].

COMBINATION OF SHARE-BASED AND RESERVATION-BASED APPROACHES

This paper has shown that share-based approaches have important advantages in terms of efficiency while reservation-based approaches perform better in terms of guarantees and isolation. The design of a resource allocation solution that combines the advantages of both approaches is a matter for future research. The approach proposed in [37], which allocates a fraction of the resources on a reservation basis and shares the remaining resources following a share-based approach, is a first step towards this end.

IX. CONCLUSION

This paper has studied resource allocation in context of network slicing. While traditional reservation-based approaches provide resource guarantees to network slices based on explicit requests, another approach considered in the standards and the literature involves allocating the resources based on fixed shares associated to the tenants. The decision of which of the two approaches is the more appropriate one will depend on the nature and requirements of the tenants. Share-based approaches are suitable for tenants that have a continued demand of resources over time and want to ensure that they will always have their share of resources available. In contrast, reservation-based approaches fit the needs of tenants with punctual needs and/or requiring hard guarantees. In this paper, we have analyzed the challenges and solutions involved with each of these approaches and have compared them against the requirements of infrastructure providers and tenants. While the focus of these approaches in on deciding the amount of resources to be allocated to each tenant, complementary mechanisms are required to handle these resources while meeting the specific service requirements; in this paper we have presented an overview of such mechanisms, giving a broad view of the components needed to perform resource allocation in network slicing and identifying some open issues.

ACKNOWLEDGMENT

Part of the work was conducted during a visit of A. Banchs at the University of Texas at Austin funded by a Salvador de Madariaga grant of the Spanish Ministry of Education, Culture, and Sports.

REFERENCES

[1] *Description of Network Slicing Concept*, NGMN 5G document P1, NGMN Alliance, Jan. 2016.

[2] *Study on Architecture for Next Generation System*, 3GPP, document TR 23.799, v0.5.0, May 2016.

[3] *5G Empowering Vertical Industries*, 5GPPP, White Paper, 2016.

[4] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[5] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[6] R. Su, D. Zhang, R. Venkatesan, Z. Gong, C. Li, F. Ding, F. Jiang, and Z. Zhu, "Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models," *IEEE Netw.*, vol. 33, no. 6, pp. 172–179, Nov. 2019.

[7] A. Antonopoulos, "Bankruptcy problem in network sharing: Fundamentals, applications and challenges," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 81–87, Aug. 2020.

[8] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, and G. S. Paschos, "The algorithmic aspects of network slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, Aug. 2017.

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

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

[11] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[12] J. Zheng, P. Caballero, G. de Veciana, S. J. Baek, and A. Banchs, "Statistical multiplexing and traffic shaping games for network slicing," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2528–2541, Dec. 2018.

[13] M. Leconte, G. S. Paschos, P. Mertikopoulos, and U. C. Kozat, "A resource allocation framework for network slicing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 2177–2185.

[14] M. Vincenzi, A. Antonopoulos, E. Kartsakli, J. Vardakas, L. Alonso, and C. Verikoukis, "Multi-tenant slicing for spectrum management on the road to 5G," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 118–125, Oct. 2017.

[15] H.-T. Chien, Y.-D. Lin, C.-L. Lai, and C.-T. Wang, "End-to-End slicing as a service with computing and communication resource allocation for multi-tenant 5G systems," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 104–112, Oct. 2019.

[16] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[17] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, and X. Costa-Perez, "A machine learning approach to 5G infrastructure market optimization," *IEEE Trans. Mobile Comput.*, vol. 19, no. 3, pp. 498–512, Mar. 2020.

[18] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Cognitive network management in sliced 5G networks with deep learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 280–288.

[19] F. Wei, G. Feng, Y. Sun, Y. Wang, S. Qin, and Y.-C. Liang, "Network slice reconfiguration by exploiting deep reinforcement learning with large action space," *IEEE Trans. Netw. Service Manage.*, early access, Aug. 25, 2020, doi: 10.1109/TNSM.2020.3019248.

[20] *Self-Organizing Networks (SON) Policy Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)*, 3GPP, document TS 28.628, Jun. 2013.

[21] *Aspects; Management and Orchestration; Concepts, Use Cases and Requirements*, 3GPP, document TS 28.530, Sep. 2019.

[22] L. Zhang, "Proportional response dynamics in the Fisher market," *Theor. Comput. Sci.*, vol. 412, no. 24, pp. 2691–2698, May 2011.

[23] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should I slice my network? A multi-service empirical evaluation of resource sharing efficiency," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw. (MOBICOM)*, Oct. 2018, pp. 191–206.

[24] *Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services (Release 16)*, 3GPP, document TS 23.288 v16.1.0, Jun. 2019.

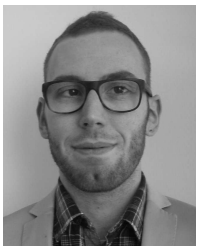
[25] *Management and Orchestration of Networks and Network Slicing; Management and Orchestration Architecture (Release 16)*, 3GPP, document TS 28.533 v16.0.0, Jun. 2019.

[26] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "AZTEC: Anticipatory capacity allocation for zero-touch network slicing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Jul. 2020, pp. 794–803.

[27] M. O. Ojijo and O. E. Falowo, "A survey on slice admission control strategies and optimization schemes in 5G network," *IEEE Access*, vol. 8, pp. 14977–14990, 2020.
[28] W. Guan, X. Wen, L. Wang, Z. Lu, and Y. Shen, "A service-oriented deployment policy of End-to-End network slicing based on complex network theory," *IEEE Access*, vol. 6, pp. 19691–19701, 2018.
[29] J. Gil Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 3, pp. 518–532, Sep. 2016.
[30] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.
[31] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 45–66, 2020.
[32] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.
[33] O. Houidi, O. Soualah, W. Louati, M. Mechtri, D. Zeghlache, and F. Kamoun, "An efficient algorithm for virtual network function scaling," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–7.
[34] G. Wang, G. Feng, T. Q. S. Quek, S. Shuang Qin, R. Wen, and W. Tan, "Reconfiguration in network slicing—Optimizing the profit and performance," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 2, pp. 591–605, Jun. 2019.
[35] A. Banchs, D. M. Gutierrez-Estevez, M. Fuentes, M. Boldi, and S. Provvedi, "A 5G mobile network architecture to support vertical industries," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 38–44, Dec. 2019.
[36] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 361–376, Feb. 2020.
[37] J. Zheng, G. de Veciana, and A. Banchs, "Constrained network slicing games: Achieving service guarantees and network efficiency," in *Proc. WiOpt*, Jun. 2020, pp. 1–8.

ALBERT BANCHS (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the UPC-BarcelonaTech, in 1997 and 2002, respectively. He has a double affiliation as a Professor with the University Carlos III of Madrid and the Deputy Director of the IMDEA Networks Institute. Before joining UC3M, he was with ICSI Berkeley, in 1997, Telefonica I+D, in 1998, and NEC Europe Ltd., from 1998 to 2003. He has participated in many European projects and industry contracts. He is currently the Technical Manager Deputy of the European Project 5G-TOURS. He has served in many TPCs and has also served in the editorial board of a number of journals. He is currently an Editor of the IEEE/ACM TRANSACTIONS ON NETWORKING.

GUSTAVO DE VECIANA (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California at Berkeley, in 1987, 1990, and 1993, respectively. He joined the Department of Electrical and Computer Engineering, where he is currently an Associate Chair and a Recipient of the Cockrell Family Regents Chair in Engineering. He served as an Editor and is currently serving as an Editor-at-Large for the IEEE/ACM TRANSACTIONS ON NETWORKING. In 2009, he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He also serves on the board of trustees of IMDEA Networks Madrid.



VINCENZO SCIANCALEPORE (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications and telematics engineering, in 2012 and 2015, respectively. He is currently a Senior 5G Researcher with NEC Laboratories Europe, focusing on network virtualization and network slicing challenges. He was a recipient of the National Award for the Best Ph.D. Thesis in the area of communication technologies (wireless and networking) issued by GTTI, in 2015. He is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

XAVIER COSTA-PEREZ (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications from the Polytechnic University of Catalonia (UPC), Barcelona. He is currently the Head of Beyond 5G Networks Research and Development, NEC Laboratories Europe, a Scientific Director of the i2Cat Research and Development Center, and a Research Professor with ICREA. His team contributes to products roadmap evolution as well as to European Commission Research and Development collaborative projects. He has published at top research venues and holds several patents. He served on the Program Committee of several conferences, including the IEEE Greencom, WCNC, and INFOCOM. He received several awards for successful technology transfers.