Resource Sharing Efficiency in Network Slicing

Cristina Marquez, Marco Gramaglia, Marco Fiore, Albert Banchs, Xavier Costa-Perez

Resource Sharing Efficiency in Network Slicing

Cristina Marquez, Marco Gramaglia, Marco Fiore, *Senior Member, IEEE,*
Albert Banchs, *Senior Member, IEEE,* and Xavier Costa-Pérez, *Senior Member, IEEE*

***Abstract*—The economic sustainability of future mobile networks will largely depend on the strong specialization of its offered services. Network operators will need to provide added value to their tenants, by moving from the traditional *one-size-fits-all* strategy to a set of virtual end-to-end instances of a common physical infrastructure, named *network slices*, which are especially tailored to the requirements of each application. Implementing network slicing has significant consequences in terms of resource management: service customization entails assigning to each slice fully dedicated resources, which may also be dynamically reassigned and overbooked in order to increase the cost-efficiency of the system. In this paper, we adopt a data-driven approach to quantify the efficiency of resource sharing in future sliced networks. Building on metropolitan-scale real-world traffic measurements, we carry out an extensive parametric analysis that highlights how diverse performance guarantees, technological settings, and slice configurations impact the resource utilization at different levels of the infrastructure in presence of network slicing. Our results provide insights on the achievable efficiency of network slicing architectures, their dimensioning, and their interplay with resource management algorithms at different locations and reconfiguration timescales.**

***Index Terms*—Network slicing, resource management, NFV.**

I. Introduction

THE next generation of mobile networks is expected to become a dominant General Purpose Technology (GPT) that will generate trillions of global economic output [1] by enabling increasingly diversified mobile services. As a consequence, network operators will be demanded to support traffic characterized by steadily more heterogeneous Key Performance Indicator (KPI) and Quality of Service (QoS) requirements. These trends are driving the design of 5G networks towards a strong differentiation of guarantees, as well exemplified by recent ITU specifications that separate Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC) and Massive Machine Type Communication (mMTC) as macroscopic categories [2].

In fact, clear needs for tailored KPI and QoS requirements are already evident in today's mobile services, which encompass, *e.g.*, high-quality video streaming, connected objects communication, low-latency mobile gaming, jointly with best effort traffic. Current state-of-the-art LTE mobile networks use a QoS approach to traffic differentiation [3], with several QoS Class Identifier (QCI) levels that map to delay and error rate levels [4], or especially tailored Medium Access Control (MAC) mechanisms that co-exists with the legacy general-purpose MAC layer [5]. However, these approaches are still insufficient to accommodate the diversity of requirements of modern applications, which makes alternative network deployments emerge. For instance, mobile communications in industrial environments rely on proprietary architectures that ensure reliability levels not attainable with public mobile networks [6]; or, an incumbent provider like Google started deploying its own radio access infrastructure and transit network to run its many services under hard QoS guarantees [7].

Network virtualization and slicing. In this context, a key item in the agenda for 5G networks is to achieve much improved service differentiation. Promises of attaining this objective heavily rely on the diffusion of software-based networking solutions, which enable *network virtualization*: they allow evolving the traditional hardbox-based infrastructure into a cloudified architecture where once hardware-only network functions (*e.g.*, for spectrum management, baseband processing, mobility management) are implemented as software Virtual Network Functions (VNFs) running on a general-purpose *telco-cloud* [8]. Network virtualization enables the deployment of multiple virtual instances of the complete network, named *network slices*. Slices are then easily customized, and create on top of the physical infrastructure a set of logical networks, each tailored to accommodate fine-tuned Service Level Agreements (SLAs) reflecting the needs of different service providers.

Network slicing and resource management. Network slicing has profound implications on resource management. When instantiating a slice, the operator needs to allocate sufficient computational and communication resources to the associated VNFs. In some cases, these resources may be dedicated, becoming inaccessible to other slices [9]. Alternatively, smart assignment algorithms can be employed to dynamically allocate resources to slices based on the time-varying demands of tenants [10], [11]. This grants the flexibility to modify the share of resources assigned to each tenant, multiplexing logical slices into the software or hardware assets while trying to abide by tenant requirements. However, it also adds complexity, and may hinder resource isolation, the corresponding guarantees to tenants, or the ability to deploy fully customized slices.

The above shows that there is an inherent trade-off among: (*i*) *service customization*, which favours the deployment of specialized slices with tailored functions for each service and, possibly, dedicated and guaranteed resources; (*ii*) *resource management efficiency*, which increases by dynamically sharing the resources of the common infrastructure among the different services and slices; and, (*iii*) *system complexity*, resulting from deploying more dynamic resource allocation mechanisms that provide higher efficiency at the cost of em-

C. Marquez, M. Gramaglia are with Universidad Carlos III Madrid, 28911 Leganés, Spain. e-mail: mcmarque@pa.uc3m.es, mgramagl@it.uc3m.es.

M. Fiore is with CNR, 10129 Turin, Italy. e-mail: marco.fiore@ieiit.cnr.it.

A. Banchs is with Universidad Carlos III Madrid and IMDEA Networks Institute, 28911 Leganés, Spain. e-mail: banchs@it.uc3m.es.

X. Costa-Pérez is with NEC Laboratories Europe, 69115 Heidelberg, Germany. e-mail: xavier.costa@neclab.eu.

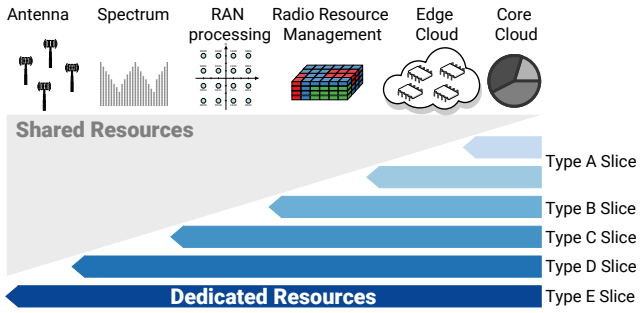


Fig. 1: Network slicing types. Deeper strategies use dedicated resources customized to services across a wider portion of the end-to-end network architecture.

ploying elaborated operation and maintenance functions [12].

Slicing strategies and trade-offs. This trade-off is fundamentally affected by the strategy adopted to implement network slicing, as illustrated in Figure 1. In its simplest realization, slices are limited to the core network (*type-A* slice in Figure 1): the allocation of resources to slices only involves cloud resources, and mostly becomes a Virtual Machine (VM) or container resource assignment problem [13]. In this case, the level of service customization granted by slices is low, since it is restricted to core network functions; yet, high efficiency can be achieved at low complexity, as a large portion of the network remains shared among all services and tenants.

More dependable slicing would offer customized functions, possibly involving dedicated resources, also at the radio access, through, *e.g.*, cloud RAN (C-RAN) paradigms. Here, basic radio-access slices allow for tailored MAC-layer scheduling [14] across a large number of antennas (*type-B* slice). Moving down the protocol stack, advanced slices implement customized baseband processing (*i.e.*, encoding and decoding operations) in the Base Band Units (BBUs), possibly providing tenants with a guaranteed bandwidth at the air interface (*type-C* slice). These approaches provide the ability to customize scheduling strategies, but they also reduce the possibility of radio resource sharing and/or increase the system complexity.

At fronthaul, resource isolation becomes a hardware problem [15]. A first case for slicing is one where tenants share antenna sites but are granted their own dedicated spectrum (*type-D* slice); we have virtually independent protocol stacks and full isolation, and sharing is limited to the physical hardware. Otherwise, tenants may require dedicated end-to-end resources down to the antennas (*type-E* slice); this results into slices that tell apart full, end-to-end virtual networks.

All slicing strategies described above may be applied independently of the kind of deployed network slice (*e.g.*, eMBB, URLLC or mMTC), as the latter maps to the orthogonal measures taken in each slice to fulfill specific tenant requirements (*e.g.*, guaranteed band for an URLLC slice, or shared core functions for a mMTC slice). In general, slicing strategies at the higher network layers provide a lower level of customization, yet they retain higher opportunities for resource sharing without additional complexity. Indeed, when slicing occurs at high layers (*e.g.*, *type-A*), the operator cannot offer full customization, but it can easily employ highly dynamic allocation schemes for the lower layers; in contrast, achieving such an efficient resource allocation is much more challenging when considering network slicing schemes with stringent requirements (*i.e.*, strategies involving the lower layers down to *type-E* slicing). For instance, when all slices have a common MAC layer, an efficient sharing of radio resources is easy, yet MAC functions are alike across services; conversely, if each slice implements a customized MAC protocol, efficiently sharing radio resources among services is more difficult.

Our contributions. From a system standpoint, the technology supporting different types of slices is well understood or even already available: there exist several cloud resource orchestrators for both commercial and open-source telco-cloud platforms [16]; and, a variety of solutions have been proposed to dynamically allocate resources across network slices [13].

However, the implications of network slicing in terms of efficient network resource utilization are still not well understood. Efficiency intuitively grows as one moves away from the radio access infrastructure (*type-E* slicing) towards the network core (*type-A* slicing); but we lack any more detailed characterization of the aforementioned trade-offs between customization, efficiency, and complexity. This is an important gap, since insights on the efficiency gains in network slicing are crucial to take informed decisions on resource configuration strategies: if efficiency is preserved with solutions that assign resources to slices more or less statically, high customization levels can be achieved at a reduced complexity; however, if the price in efficiency is high, more elaborate (and expensive) solutions may be desirable.

Our aim is to shed light on the trade-offs between customization, efficiency, and complexity in network slicing, by evaluating the impact of resource allocation dynamics at different network points. Our analysis offers insights that help determining in which cases the gain in efficiency is worth the sacrifice in customization/isolation and/or the extra complexity, and when a specific resource assignment algorithm pays off. Since resource management efficiency in network slicing highly depends on the traffic patterns of different services supported by the various slices, we build on real-world service-level measurement data collected by a major operator in a production mobile network, and:

(*i*) quantify the price paid to guarantee resource isolation to diverse types of network slices, under different QoS requirements and for alternative resource allocation policies;

(*ii*) unveil the effect of dynamically applying such policies, which allows the operator to periodically re-orchestrate resources or aggregate more than one service into the same slice;

(*iii*) study a number of specific use cases for slice deployment, so as to gain insight about the efficiency of network slicing deployments in scenarios of practical interest.

The results included in this paper can be used as insights for rule of thumb calibrations of network slicing deployments, and to evaluate the solution space for smart resource assignment algorithms under dynamically changing conditions.

II. Network Scenario and Metrics

In the following we expose our network model, as well as our representation of the slice QoS requirements and their

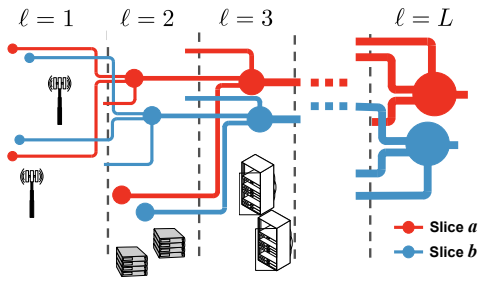


Fig. 2: Mobile network architecture. Nodes map to different equipment depending on the level ℓ, and form a hierarchy. The mobile traffic in each slice (*e.g.*, *a* or *b*) is increasingly aggregated as it flows from radio access to network core.

associated resource allocation strategy. We also introduce the metrics we adopt to evaluate the resource sharing performance.

A. Network slicing scenario

Let us consider a mobile network providing coverage to a geographical region where mobile subscribers consume a variety of services. The network operator implements slices $s \in \mathcal{S}$, each dedicated to a different subset of services.

We assume that each slice can be implemented according to any of the strategies in Figure 1. To capture such a general scenario, we model the mobile network architecture as a hierarchy composed by a fixed number of levels ($\ell = 1, \ldots, L$) ordered from the most distributed ($\ell = 1$) to the most centralized ($\ell = L$), as illustrated in Figure 2. Every network level ℓ is composed by a set C_ℓ of network nodes, each serving a given number of base stations. In the two extremes, we have $\ell = 1$, where network nodes in C_1 have a bijective mapping to individual antennas, and $\ell = L$, where C_L contains a single network node controlling all antennas in the whole target region. In between, for $1 < \ell < L$, the number of network nodes in C_ℓ decreases with ℓ, whereas that of base stations served by each such node increases accordingly. Note that, in general, a node $c \in C_\ell$ will operate on data flows that are increasingly aggregated with ℓ, which, as we will see, has a significant impact on resource management.

This hierarchical representation allows considering a variety of node types, along with their associated (possibly virtual) network functions. At the most distributed level ($\ell = 1$), each node runs functions that operate at the antenna level, *e.g.*, concern spectrum or airtime resources. In intermediate cases ($1 < \ell < L$), nodes are at first in charge of a small number of antenna sites, *e.g.*, C-RAN datacenters running VNFs such as dedicated baseband processing or radio resource management. As ℓ grows, VNFs are pushed further towards the network core, into telco-cloud datacenters that tunnel traffic to and from large sets of antenna sites: there, VNFs customize VM resources for large traffic volumes associated to the services delivered by each tenant to subscribers in wide geographical areas. In the limit case ($\ell = L$), all traffic in the target region is managed in a fully-centralized fashion at a single datacenter.

Ultimately, the layered network model allows generalizing our analysis to diverse VNFs, by studying the system performance at different network levels. This also implicitly accommodates all of the network slicing strategies outlined in Figure 1. Slices of *type-D* and *type-E* deal with the lowest network layers that are implemented at the antennas, hence correspond to $\ell = 1$. Slices of *type-A* refer to VNFs operating at higher network layers that are deployed at centralized cloud datacenters, hence correspond to high values of the network level ℓ. Slices of *type-B* and *type-C* are concerned with VNFs at radio access, *i.e.*, at base stations ($\ell = 1$), or at higher architectural levels ($1 < \ell < L$) in a C-RAN implementation.

Note that we do not require that a single network deploys virtualization technologies at all network levels. Instead, by taking a large number of levels and considering each of them in isolation, this approach lets us cover a wide range of deployment options and provide insights for all of them.

B. Slice specifications

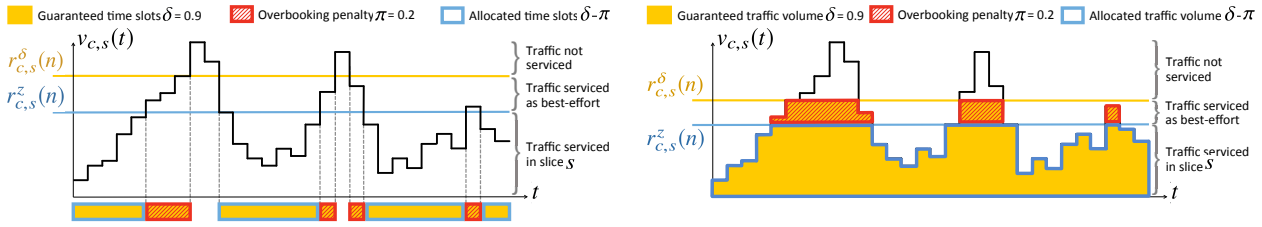
Network slicing primarily aims at letting the operator fulfill the QoS requirements requested by each tenant. To model such requirements, we consider discrete-time demands associated to slices, by averaging traffic over *time slots* denoted by t. Let $v_{c,s}(t)$ be the traffic demand associated to slice s at node c during slot t, as in Figure 3. We capture the QoS requirements of s as a *slice specification* z defined by two features.

1) Guaranteed demand δ: The operator engages to guarantee that the total traffic demand of the slice is fully serviced for a portion at least $\delta \in [0, 1]$, which can be expressed in terms of time or traffic. In the first case, the operator assures that the slice demand is fulfilled during a fraction δ of time slots, as in Figure 3a. In the second case, the slice demand is serviced for a fraction at least δ of its volume, as in Figure 3b.

2) Overbooking penalty π: The operator can decide to overbook network resources to multiple slices, transparently to the tenants [17]. Similar to common practices in the airline or hotel industries, this management model allocates the same resources to multiple tenants, expecting that some will ultimately not use all of their booked capacity; if this is not the case, and services actually require all of the reserved capacity, overbooking leads to violations of the guaranteed demand δ. Through overbooking, the operator can maximize its revenues by properly balancing the cost of allocated resources and the penalty associated with violations [17]. In our model, we do not adopt a specific overbooking strategy; instead, we consider that the strategy selected by the operator produces violations for a portion $\pi \leq \delta$ of the total traffic demand. This implies that only a fraction of traffic $\delta - \pi$ is actually serviced by the slice, while the portion π of violated demand is treated as best-effort traffic by the operator. This approach can capture any overbooking strategy, and lets us investigate how violations of δ affect savings in allocated resources. We remark that π may be a fraction of time slots or a fraction of traffic volume, consistently with the representation of δ: the two situations are illustrated in Figure 3a and Figure 3b, respectively.

C. Resource allocation to one slice

We denote a slice specification characterized by a guaranteed demand δ and an overbooking penalty π as $z = (\delta, \pi)$, which becomes more stringent for higher values of δ and



(a) Time slot δ (b) Traffic volume δ

Fig. 3: Example of resource allocation to a slice s at node c, under guaranteed demand $\delta = 0.9$ and overbooking penalty $\pi = 0.2$, during one reconfiguration period n. (a) The slice specification is expressed in terms of time slots, hence the discrete-time traffic of the slice, $v_{c,s}(t)$, is serviced for 90% of the time slots, denoted by the filled (yellow) temporal interval below the abscissa. Due to overbooking, demand δ is violated in 20% of the total time slots, highlighted by the (red) pattern intervals below the abscissa. (b) The slice specification is expressed in terms of traffic, hence $v_{c,s}(t)$ is serviced for 90% of its volume, denoted by the filled (yellow) area under the time series. Due to overbooking, demand δ is violated for 20% of the total volume, highlighted by the (red) pattern region. In both (a) and (b), horizontal solid lines denote the minimum allocated capacity satisfying the guaranteed demand constraint, $r^{\delta}_{c,s}(t)$, and the capacity actually allocated based on the operator's overbooking strategy, $r^{z}_{c,s}(t)$.

smaller π. The operator shall then ensure that enough resources are dedicated to the slice so as to meet z. We now expound the expression of the resources allocated to a slice s by the mobile network operator under a generic $z = (\delta, \pi)$.

In presence of algorithms that enable a dynamic reconfiguration of VNFs, the resource allocation can be re-modulated over time. In practice, however, the periodicity of reconfiguration is limited by the technological constraints of the slicing strategy adopted (see Figure 1). For instance, when network slicing is performed at the antenna level, times in the order of minutes are needed to turn on and off the radio-frequency front-end and reset the transport network. When dealing with radio resource management algorithms (*i.e.*, dynamic spectrum or multi-provider scheduling), re-assignments are constrained by signalling overhead. Or, in the case of VM orchestration, the timescale is limited by instantiation and migration delays [18].

Let us assume that τ is the minimum amount of time steps needed for resource reallocation, which we refer to as a *reconfiguration period*. We denote by $n \in \mathcal{T}$ the n^{th} reconfiguration period within the set $\mathcal{T}$ of all reconfiguration periods that compose the system observation time; n can be then seen as the set of τ time steps it encompasses, *i.e.*, $n = \{t, \ldots, t + \tau - 1\}$. During period n, we name $r^{\delta}_{c,s}(n)$ the minimum amount of resources that allow meeting the guaranteed demand δ for slice s at node c. Equivalently, $r^{z}_{c,s}(n)$ is the amount of resources that fulfill z, accounting for both δ and the overbooking penalty π. The formalism is the same when δ is a fraction of time or traffic, as shown in Figure 3. Then, our objective is the computation of $r^{z}_{c,s}(n)$, which represents the resources actually allocated by the operator to slice s at node c, based on $v_{c,s}(t)$ and z. Since calculations are different depending on whether δ (hence π) is expressed in terms of time or traffic, below we discuss these two instances separately. For the sake of readability, in the following we drop the c, s, and n notation, and refer to a generic slice, network node, and reconfiguration interval; hence $v(t)$ and r^{z} stand for $v_{c,s}(t)$ and $r^{z}_{c,s}(n)$, respectively.

1) Time slot fraction δ: In this case, the allocation of resources in the target reconfiguration period is such that the offered load $v(t)$ exceeds r^{z} for a fraction $\delta - \pi$ of the time slots in the reconfiguration period, as shown in Figure 3a. This can be formalized as $P\big(v(t) \leq r^{z}\big) = \delta - \pi$, $\forall t$, where $P(\cdot)$ denotes the probability of the argument. Let f_v be the Probability Density Function (PDF) of the demand, *i.e.*, $f_v(x) = P\big(v(t) = x\big)$. Then, the Cumulative Distribution Function (CDF) of the demand $v(t)$ in the reconfiguration period is $F_v(x) = \sum_{y=0}^{x} f_v(y) = P\big(v(t) \leq x\big)$. Therefore, the original condition above is $F_v\big(r^{z}\big) = \delta - \pi$, and the minimum r satisfying the actual guaranteed demand is $r^{z} = F_v^{-1}(\delta - \pi)$. Figure 4a illustrates this concept in a practical example[1].

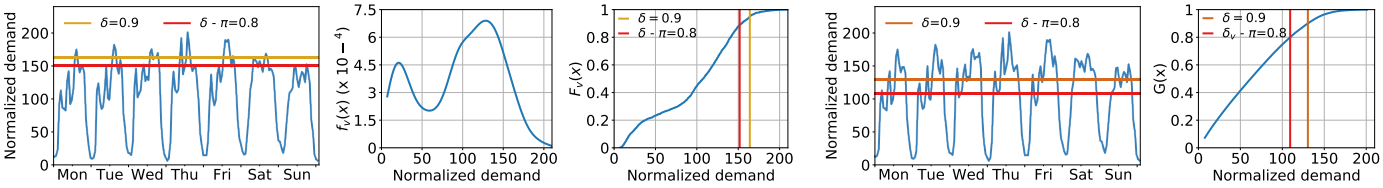
2) Traffic volume fraction δ: When the operator guarantees (and overbooks) a fraction of traffic, we do not reason in time slots anymore, but account for the effective demand volume associated to each time slot. For this purpose, we introduce a water-filling function, that computes the overall fraction of served traffic as a function of the assigned resources r. Specifically, we define $G(x) = \sum_t \left(\min\left(v(t), x\right)\right) / \sum_t v(t)$, for all time slots t in the target reconfiguration period. Through the above expression of $G(x) \in [0, 1]$, the value of x maps to the upper limit of a water-filling algorithm. The minimum r^{z} satisfying the actual guaranteed demand is then $r^{z} = G_v^{-1}(\delta - \pi)$. Figure 4b illustrates this concept in a practical case.

Note that in both cases above, the expressions of r assume that the amount resources needed to serve a given slice is directly proportional to the mobile traffic demand in that slice. While this clearly holds for some types of resources (*e.g.*, radio), we acknowledge that it may be a strong simplification in other settings. We argue, however, that it is a reasonable assumption for many practical VNFs. Moreover, this choice allows us to investigate through a unified framework different network levels ℓ, where resources map to diverse physical assets (such as spectrum, airtime, CPU time, computational power, or memory) depending on ℓ.

D. Multiplexing efficiency

Having computed $r^{z}_{c,s}(n)$ according to either model in Section II-C, we can define the amount of dedicated resources that

[1]Traffic volumes in Figure 4 as well as in the rest of the result reported in the paper are normalized with respect to the minimum average traffic recorded at a 4G antenna sector in our reference scenarios presented in Section III.



(a) Time slot (b) Traffic volume

Fig. 4: Example of resource allocation to a slice with specification $z = (\delta, \pi) = (0.9, 0.1)$ when δ is a fraction of time slots (a) or of traffic volume (b). Left (a,b): weekly time series of the mobile traffic demand for a slice s at a network node c. The horizontal lines denote the minimum resources r^{δ} and r^{z} to be allocated when $\tau = 1$ week. Central (a): representation of $f_v(x)$. Right (a,b): $F_v(x)$ and $G(x)$, with cuts at δ and $\delta - \pi$ that identify the needed resource r^{δ} and r^{z}, respectively.

the operator allocates to network slices at network level ℓ, over the entire system observation period, as

$$\mathbb{D}^{z}_{\ell,\tau} = \sum_{s\in\mathcal{S}} \sum_{c\in\mathcal{C}_\ell} \sum_{n\in\mathcal{T}} \tau \cdot r^{z}_{c,s}(n). \quad (1)$$

Equation (1) covers the demand that receives dedicated resources within slices. However, under overbooking, a fraction π of traffic is penalized, *i.e.*, is treated as best-effort. Such traffic is not isolated anymore, and can be aggregated into a single time series described, at node c and during period n, as

$$v_c(t) = \sum_{s\in\mathcal{S}} \max\left\{0, \min\left\{r^{\delta}_{c,s}(n), v_{c,s}(t)\right\} - r^{z}_{c,s}(n)\right\}. \; t \in n, \quad (2)$$

This equation computes the penalized traffic in a slice s as the difference between the resources dedicated to the slice, $r^{z}_{c,s}(n)$, and those that would be actually needed to accommodate the guaranteed demand, $r^{\delta}_{c,s}(n)$. As exemplified in Figures 4a and 4b, $r^{\delta}_{c,s}(n)$ can be computed by the strategy in Section II-C, as $F^{-1}_{v}(\delta)$ or $G^{-1}_{v}(\delta)$, for the cases where δ is a fraction of time slots or traffic volume, respectively. Then, the shared resources required to serve all traffic penalized by overbooking are, trivially, $r_c(n) = \max_{t\in n} v_c(t)$. Finally, we can calculate the total amount of resources that the operator needs to allocate at network level ℓ, in order to meet specifications z, as

$$\mathbb{R}^{z}_{\ell,\tau} = \mathbb{D}^{z}_{\ell,\tau} + \sum_{c\in\mathcal{C}_\ell} \sum_{n\in\mathcal{T}} \tau \cdot r_c(n). \quad (3)$$

Equation (3) returns the total amount of resources that the operator needs to provision at network level ℓ in order to satisfy its commitments with all tenants, when dynamically reconfiguring[2] the allocation with periodicity τ, and according to its designated overbooking strategy. In order to unveil the implications of this value, we compare it against a *perfect sharing* benchmark. In perfect sharing, the allocated resources correspond to those required when there is no isolation among different services, hence traffic multiplexing is maximum.

Let $u_c(t) = \sum_{s\in\mathcal{S}} v_{c,s}(t)$ be the total demand for mobile data traffic at node c, summed over all slices. We then denote by $\hat{r}^{\delta}_{c}(n)$ the resources needed to accommodate $u_c(t)$ during reconfiguration period n. For the sake of fairness, the same requirement δ on guaranteed demand is enforced here as well[3].

[2]Equation (3) maps to the special case where no reconfiguration is possible at level ℓ, when τ is the total system observation time, *i.e.*, $|\mathcal{T}| = 1$.

[3]We remark that the notion of overbooking penalty is meaningless under perfect sharing, as all traffic is aggregated and treated as best-effort already.

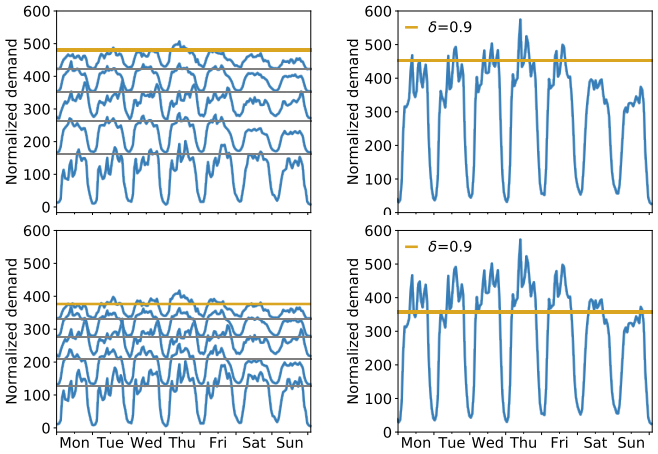


Fig. 5: Examples of multiplexing efficiency, when $\delta = 0.9$ expressed in time slots (top) and traffic volume (bottom).

Thus, adopting the methodology presented in Section II-C, $\hat{r}^{\delta}_{c}(n)$ can be computed as $F^{-1}_{u}(\delta)$ or $G^{-1}_{u}(\delta)$, where $F_u(x)$ and $G_u(x)$ are the CDF of the total demand $u(t)$, $t \in n$, expressed in time slots and traffic volume, respectively. The resources allocated under perfect sharing are then computed as

$$\mathbb{P}^{\delta}_{\ell,\tau} = \sum_{c\in\mathcal{C}_\ell} \sum_{n\in\mathcal{T}} \tau \cdot \hat{r}^{\delta}_{c}(n). \quad (4)$$

Taking the above benchmark, we define the *multiplexing efficiency* as the ratio between the resources required with perfect sharing and those needed under network slicing, *i.e.*,

$$\mathbb{E}^{z}_{\ell,\tau} = \mathbb{P}^{\delta}_{\ell,\tau} \;/\; \mathbb{R}^{z}_{\ell,\tau}. \quad (5)$$

In summary, $\mathbb{E}^{z}_{\ell,\tau}$ quantifies the efficiency of network slicing in terms of resource management at network level ℓ, under resource reconfiguration intervals of duration τ, and with slice specification $z = (\delta, \pi)$. As $\mathbb{E}^{z}_{\ell,\tau}$ approaches one, the total amount of slice-isolated resources tends to that assured by a perfect sharing. As slicing the network becomes increasingly capacity-demanding, the efficiency drops instead towards zero.

Let us illustrate the operation of multiplexing efficiency in Figure 5, when δ is expressed as a fraction of time slot (top) or of traffic volume (bottom). The left column depicts the time series of the mobile traffic demand for a set $\mathcal{S}$ of five slices, observed at a single network node c, during one reconfiguration interval n ($\tau = 1$ week). A slice specification $z = (\delta, \pi) = (0.9, 0)$ commits the operator to allocate, for each

slice s, at least the capacity marked by the grey horizontal lines, which are computed as discussed in Section II-C. Their sum, in thick gold, denotes $\sum_{s\in\mathcal{S}} r^z_{c,s}(n) + r_c(n)$, *i.e.*, the value that, once multiplied by τ, returns the resources specified by Equation (3), at a single node c and during reconfiguration interval n. The right column, instead, shows the time series of the traffic demand aggregated over all slices in $\mathcal{S}$. By applying the specification z, we get a value $\hat{r}^z_c(n)$, highlighted by the horizontal thick gold line. Its multiplication by τ gives the equivalent capacity needed under *perfect sharing* as per Equation (4). Then, the multiplexing efficiency is the ratio between the values highlighted by the thick gold lines on the right and left plots, respectively. In this toy example, the value on the left is only slightly higher than that on the right, hence $\mathbb{E} \sim 1$ and resource isolation is efficient. This is not necessarily the case in practical scenarios, as we will detail later.

III. Reference scenarios

We evaluate the efficiency of resource management in a sliced network by considering two modern metropolitan-scale network scenarios. As mentioned in Section I, today's mobile services already offer a variety of requirements that makes it meaningful to investigate the impact of slice isolation on network efficiency with present traffic.

Our two reference urban regions are a large metropolis of several millions of inhabitants, and a typical medium-sized city with a population of around 500,000, both situated in Europe. Service-level measurement data was collected in the target areas by a major operator with a national market share of around 30%. Details are in Section III-A. On top of this, we model the hierarchical network infrastructures in the target regions by assuming a deployment of nodes that balances load and reduces latency. This is discussed in Section III-B.

A. Mobile service demands

The real-world demands generated by individual mobile services in the two reference regions were collected during three months in late 2016. The information was gathered by monitoring individual IP data sessions over the GPRS Tunneling Protocol User plane (GTP-U), and running Deep Packet Inspection (DPI) and proprietary fingerprinting algorithms to infer the mobile service associated to each 2G/3G/4G data session. The data was aggregated geographically (per antenna sector) and temporally (over 5-minute time intervals) by the operator, so as to make the data non-personal and to preserve user privacy; all operations were carried out within the operator premises, under control of the local Data Privacy Officer (DPO), and in compliance with applicable regulations.

The resulting measurement data describe downlink and uplink traffic for hundreds of prominent mobile services consumed in the target regions. Building on such information, we define potential slices by identifying mobile services that meet two requirements: (*i*) they generate a substantial offered load (above 0.1% of the total network traffic), sufficient to justify a dedicated network slice; and (*ii*) they have clear KPIs and QoS requirements. We identify 38 services that meet the criteria above, and associate them to a different network slice each.

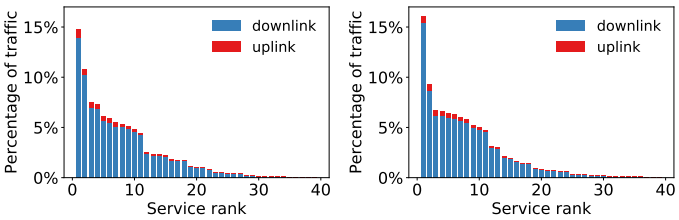


Fig. 6: Percentage of the mobile traffic generated by the selected services. Different colors denote downlink and uplink traffic. Left: large metropolis. Right: medium-sized city.

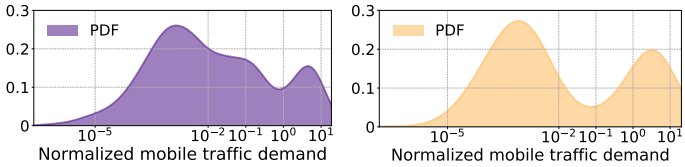


Fig. 7: PDF of the traffic demands across all antenna sectors. Left: large metropolis. Right: medium-sized city.

Our choice of services represents well the heterogeneous nature of today's mobile traffic. It encompasses many popular services, such as YouTube, Netflix, Snapchat, Pokemon Go, Facebook or Instagram, and covers a wide range of classes with diverse network requirements, including mobile broadband (*e.g.*, long-lived and short-lived video streaming), low-latency (*e.g.*, gaming, messaging), and best effort (*e.g.*, web browsing, social media), which are representative forerunners of 5G services [19]. Figure 6 provides basic information on our selection of services. It outlines the downlink-dominated, highly skewed traffic split among the services, whose percent traffic can differ of more than two orders of magnitude.

A strong diversity also emerges in the way the selected services are consumed across the geographical space within the two urban regions. Figure 7 portrays the PDF of the total offered load at individual antenna sectors, which again spans several orders of magnitude. The main cause of heterogeneity is the radio access technology: our measurement data captures 2G, 3G, and 4G access, and 4G antennas accommodate much larger fractions of the demand and generate the rightmost bell-shaped lob of the distributions. Still, 10-time differences in the traffic volume appear even across 4G antenna sectors, implying substantial location-based demand variability.

B. Hierarchical network structure

The deployment of antennas in the target regions is shown in Figure 8, which highlights the diversity of the case studies in terms of network infrastructure, owing to the different geographical span and user population density of the two areas. While we do not have information on the architecture of the mobile networks beyond the radio access, we model the hierarchical structure exemplified in Figure 2 after current proposals for cloudified network slicing [20], as follows.

At the generic level ℓ, the operator deploys a number $N_\ell = |C_\ell|$ of nodes, each responsible for a subset of the antenna sites at the radio access level. Every node will thus run VNFs (whose nature will depend on ℓ) on the mobile data traffic incoming from or outgoing to its associated antennas.

Fig. 8: Antenna deployments in the target regions. Left: large metropolis. Right: medium-sized city.

We assume that the operator deploys generic level-ℓ nodes and links based on two criteria: (*i*) the offered load shall be similar at all nodes; (*ii*) the subset of antennas served by a same node shall be geographically contiguous. The first criterion ensures load balancing, and the second reduces latency between antenna sites and nodes. Jointly, these criteria represent a plausible strategy that aims at maximizing the performance of network slicing. We remark that the resulting node deployment is static and does not change during our experiments; instead, the node resources allocated to each slice may change under dynamic resource allocation schemes.

Under these criteria, the problem of associating the level-ℓ nodes with the original antenna sites in Figure 8 is a special case of *balanced graph k-partitioning*. Let us consider a graph where each vertex $v \in V$ maps to one antenna site, and has an associated cost $c(v)$ equal to the mobile traffic demand recorded at the site; also, let an edge $e = \{u, v\} \in E$ connect vertices u and v only if the corresponding antenna sites are geographically adjacent[4]. The problem of level-ℓ node-to-antenna site association translates into dividing the graph into N_ℓ sub-graphs, such that the sum of costs of nodes in each partition is balanced. We introduce decisions variables

$$e_{uv} = \begin{cases} 1 & \text{if } e \text{ is a cut edge} \\ 0 & \text{otherwise} \end{cases} \qquad \forall e \in E, \tag{6}$$

$$x_{v,k} = \begin{cases} 1 & \text{if } v \text{ is in partition } k \\ 0 & \text{otherwise} \end{cases} \qquad \forall v \in V, \forall k, \tag{7}$$

and formulate an Integer Linear Programming (ILP) problem:

$$\min \sum_{e_{uv} \in E} e_{uv} \tag{8}$$

$$\text{s.t.} \sum_{v \in V} x_{v,k} \cdot c(v) \leq (1+\epsilon) \cdot \frac{\sum_{v \in V} c(v)}{N_\ell}, \ \forall k \tag{9}$$

$$\sum_{v \in V} x_{v,k} \cdot c(v) \geq (1-\epsilon) \cdot \frac{\sum_{v \in V} c(v)}{N_\ell}, \ \forall k \tag{10}$$

$$\sum_{k} x_{v,k} = 1, \ \forall v \in V. \tag{11}$$

$$e_{uv} \geq x_{u,k} - x_{v,k}, \ \forall e \in E, \forall k \tag{12}$$

$$e_{uv} \geq x_{v,k} - x_{u,k}, \ \forall e \in E, \forall k \tag{13}$$

The objective function given by Equation (8) aims at minimizing the number of cut edges that join vertices in separate partitions, so as to generate graph subsets that are as compact as possible. Our goal in terms of load balancing is ensured by the constraints given by Equations (9) and (10), which bound the load difference among the various subsets of antennas: each partition is forced to have a total cost that is within a fraction ϵ from the ideal case of a perfectly even cost $\sum_{v \in V} c(v)/N_\ell$. The constraint given by Equation (11) ensures that each vertex is in exactly one partition, while those given by Equations (12) and (13) determine the value of decision variables e_{uv} based on whether vertices u and v belong to a same partition as defined by $x_{u,k}$ and $x_{v,k}$.

The resulting optimization problem is NP-hard. We use a suitably configured version of the Karlsruhe Fast Flow Partitioner (KaFFPa) heuristic [21] to solve it. In doing so, we allow for ±10% imbalance among the load served by nodes at every level ℓ, *i.e.*, $\epsilon = 0.1$ in Equations (9) and (10). Figure 9 shows three examples of antenna site partitioning among network nodes, for a selection of levels ℓ in the large metropolis scenario. Table I summarizes instead the main features of the partitions obtained in our two urban scenarios.

[4]Multiple notions of adjacency are possible. We opt for one that leverages the common practice of approximating antenna coverage areas via a Voronoi tessellation: two sites are then adjacent if they share one Voronoi cell side.

Fig. 9: Association of antenna sites to level-ℓ nodes in the large metropolis scenario. The plots refer to $\ell = 8$ (16 nodes, left), $\ell = 9$ (8 nodes, middle) and $\ell = 10$ (4 nodes, right).

ℓ		1	2	3	4	5	6	7	8	9	10	11	12
Traffic per node		5	10	15	30	60	75	100	150	300	600	1167	2334
N_ℓ	Metropolis	422	230	160	80	40	32	23	16	8	4	2	1
	City	122	60	40	20	10	8	6	4	2	1		

TABLE I: Hierarchical network structures. Rows are (*i*) the level $\ell \in \{1, \ldots, 12\}$, (*ii*) the corresponding normalized mobile traffic per node, and *(iii)-(iv)* the number of nodes N_ℓ serving each urban region at network level ℓ. At $\ell = 1$, nodes map to 4G antenna sectors, and the traffic per node is an average. From $\ell = 2$ to $\ell = L$, we consider the partitions obtained by solving the optimization problem given by Equation (8).

IV. Data-driven evaluation

Our performance evaluation is organized as follows. First, we investigate worst-case settings where very stringent slice specifications are enforced and no reconfiguration is possible (Section IV-A). We then relax these constraints, and assess efficiency as slice specifications are moderated (Section IV-B), as well as under a dynamic orchestration of network resources (Section IV-C). We then investigate the impact of a varying number of slices on efficiency (Section IV-D), and finally explore a number of meaningful, specific case studies among all possible system configurations (Section IV-E).

A. Slicing efficiency in worst-case settings

The least efficient sliced network scenario implies (*i*) strict slice specifications, where the mobile network operator com-

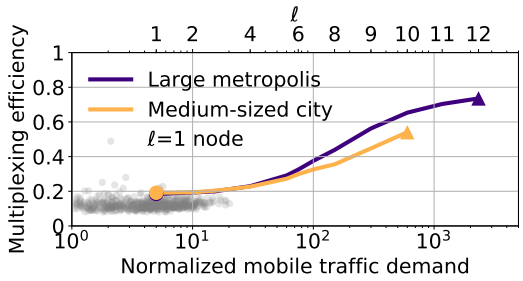


Fig. 10: Efficiency of slice multiplexing versus the normalized mobile traffic served by one node (bottom x axis) at level ℓ (top x axis) in the two reference urban scenarios. Results are for a static resource assignment, *i.e.*, $|\mathcal{T}_\tau| = 1$, and slice specification $z = (\delta, \pi) = (1, 0)$. Dots denote $\ell = 1$ and triangles $\ell = L$, for each scenario. Scattered grey points around $\ell = 1$ denote the efficiency and traffic measured at all level-1 nodes (*i.e.*, individual 4G antenna sectors) separately.

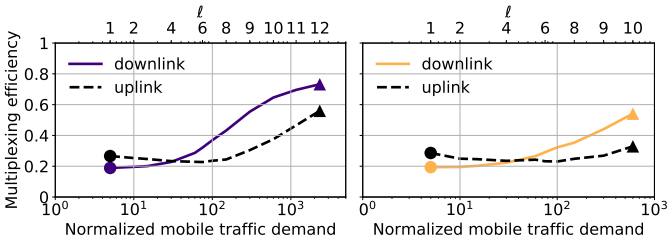


Fig. 11: Efficiency of slice multiplexing, in the same settings of Figure 10, separating downlink and uplink. Left: large metropolis. Right: medium-sized city.

mits to guarantee the whole traffic demand ($\delta = 1$) for all slices, (*ii*) no possibility of overbooking ($\pi = 0$), and (*iii*) a static allocation of resources without option for reconfiguration over time (τ spans the whole three-month observation time in our measurement dataset, and $|\mathcal{T}_\tau| = 1$). With this configuration, the operator trades efficiency for simplicity: it replicates physical resources for different slices, and statically allocates to each slice the resources needed to meet the associated offered load. This strategy yields the lowest efficiency in terms of occupied physical resources, but does not require any advanced solution for dynamic resource management to be implemented in the network. It could be a pragmatic approach to practical network slicing, if the loss of efficiency is small.

Figure 10 portrays the multiplexing efficiency of slicing as a function of the network hierarchy level ℓ (top x axis); for the sake of clarity, the latter is mapped to the normalized mobile traffic demand observed by a level-ℓ node (bottom x axis), as per Table I. The two curves refer to our two reference urban scenarios, and outline the fluctuation of the efficiency as one moves from resources at the antenna level (dot on the left) to those in a fully centralized cloud (triangle on the right). These results, and all others unless stated otherwise, refer to the case where the 16 mobile services that generate the most network traffic are allocated to independent slices each; the rationale for this choice will be apparent when discussing the effect of a varied number of slices, in Section IV-D.

The curves in Figure 10 confirm the intuition that the efficiency grows as one moves from very distributed resources at the antenna level to more centralized ones. This trend roots in the temporal dynamics of traffic in the difference slices: the demands for each slice are typically very bursty at individual antenna sectors, whereas aggregating demands over a growing number of base stations results in increasingly smoother time series. The coefficients of variation of the traffic time series substantiate this conjecture: their values range in $[1.487, 2.363]$ for $\ell = 1$ and in $[0.511, 0.587]$ for $\ell = L$, with intermediate levels resulting in midway ranges. The erratic, high activity peaks that occur at the antenna level ($\ell = 1$) force the allocation of substantial static resources in order to accommodate the per-slice traffic. For higher ℓ values, peak-to-average ratios are instead substantially reduced, mitigating these effects and increasing the overall efficiency.

In addition to the qualitative trend with ℓ, Figure 10 lets us appreciate the following quantitative results on the efficiency.

- The efficiency is very low (~0.19) at the antenna level: ensuring physical resource isolation across slices in absence of dynamic reconfiguration capabilities would require more than 5 times the capacity of a legacy architecture where no network slicing is implemented. The grey points highlight that such a poor efficiency affects all 4G antenna sectors, independently of their specific offered load.
- The efficiency grows slowly when aggregating traffic at the network edge ($\ell = 2$ to $\ell = 6$). Some gain starts to be appreciable as one moves above $\ell = 7$ in our reference scenarios, *i.e.*, at network nodes that accommodate the demands from many tens of antenna sectors at least.
- However, in absolute terms, even when considering that all traffic generated in each of our two urban scenarios is aggregated at a single level-L node ($L = \{12, 10\}$ in the large metropolis and medium-sized city, respectively, see Table I), the efficiency stays fairly low, at 0.54–0.74.

We note that, although the method presented in Section II-C operates on individual levels separately, Figure 10 offers a complete view of end-to-end efficiency across the network, and the result covers all of the different types of slices presented in Section I. For instance, a *type-A* slicing in Figure 1 limits the analysis to the rightmost part of the plot: implementing the most basic form of slicing requires roughly doubling the resources deployed in the network core cloud with respect to a legacy non-sliced case. More complicated slices that reach deeper into the network architecture encompass larger portions of the curves in Figure 10. As an example, let us image that a *type-C* slice in Figure 1 corresponds to a network level $\ell = 6$ in a specific infrastructure layout: then, the plot details the loss of efficiency that the operator can expect at all intermediate nodes, down to a threefold increase of required resources at the C-RAN datacenters that lie at the very edge of the slice. Furthermore, when considering an end-to-end network slice, we have that the slice can be associated to resources located at different levels of the network infrastructure (e.g., some resources at the antenna $\ell = 1$ and others at the core $\ell = L$). In this case, the resulting overall efficiency of the network slice is the combination of the individual efficiencies of the resources deployed at different levels.

The results can be further disaggregated for the downlink and uplink directions, as shown in Figure 11. Downlink traffic

dominates the total demand, as previously seen in Figure 6: therefore, the associated efficiency curves are very close to those in Figure 10. However, the trend of efficiency during uploads is sensibly different from the global one: slicing traffic in uplink tends to become remarkably (40% to 60%) less efficient as one moves towards more centralized network levels. We argue that the reason lies in the small uplink traffic volume, which results in bursty time series with high peak-to-average ratios, even upon aggregation over multiple antennas.

The distinct trends for downlink and uplink are especially important in the light of the different costs associated to the demands in the two directions. By looking at the sheer traffic load, the overall resource assignment should be driven by the downlink behavior, since it currently dominates the aggregate data volumes, as per Figure 6. However, specific applications, hence slices, heavily rely on uplink traffic: for instance, the fact that efficiency at the antenna level is also low in uplink means that services with strong requirements on access network latency (*e.g.*, mobile gaming) are as hard to accommodate as downlink bandwidth-eager ones (*e.g.*, video streaming). As another example, baseband processing at a virtualized radio access is remarkably more CPU-intensive for uplink traffic [22]: the very low efficiency recorded in uplink at the network edge can make resources assignment challenging when dealing with *type-C*, *type-D* or *type-E* slices in Figure 1.

An interesting final remark on the results in Figures 10 and 11 is that we do not observe substantial differences between the two reference cities. Minor discrepancies only emerge for high values of ℓ, and can be easily imputed to the intrinsic topological and demographic differences that characterize the two scenarios. The affinity of results in the two different urban regions is in fact a constant across all results, as it will be observed in the remainder of this Section.

B. Configuring slice specifications

Severe slice specifications may represent a major cause for the poor efficiency recorded in Section IV-A. To gain insight on this, we investigate the impact that the QoS requirements imposed on each slice have on the opportunities for multiplexing slice demands. Note that here we still consider a static allocation of resources, and no possibility of reconfiguration.

Figure 12 offers a complete overview of sensible resource configuration schemes, in which we vary the overall QoS that each tenant is provided by the operator. Consistently with our system model, different QoS levels are reflected by diverse values of δ and π, hence we explore the impact of those two parameters on the multiplexing efficiency. The first configuration is depicted in the top-left pair of plots in Figure 12, which portray efficiency as a function of the guaranteed demand δ expressed as a time slot fraction, with no overbooking ($\pi = 0$). As one would expect, efficiency grows when not all the traffic demand for each slice has to be served. The increase is much more evident in the case of antenna-level resources ($\ell = 1$) than in the network core ($\ell = L$). The good news is that a large fraction of the gain is achieved close to $\delta = 1$, *i.e.*, a slight reduction from a fully guaranteed demand may yield a large gain: in the best case, reducing δ from 1 to 0.99 raises efficiency from 0.35 to 0.6 (a 71% increase) when $\ell = 7$ in the medium-sized city scenario. The bad news is instead that efficiency values that are actually serviceable for the operator are only reached when significant amounts of traffic are not accommodated: figures above 0.8 (implying that implementing network slicing requires no more than 25% additional resources) are achieved in all configurations only when $\delta = 0.9$, and 10% of the demand is denied.

Trends are similar when the same slice specification parameters (varying δ, and $\pi = 0$) are defined as a traffic volume fraction, in the top-right pairs of plots in Figure 12. The major differences are at the antenna level ($\ell = 1$), where the multiplexing efficiency is substantially lower than in the case of δ and π expressed as time slot fractions. Indeed, imposing QoS constraints in terms of time slots or traffic volume leads to comparable efficiency when all time slots contribute a similar amount of traffic volume, and the demand is even over time. Centralized cases with high ℓ are closer to this situation. However, we already noted in Section IV-A that demands are much more irregular close to the radio access: here, most of the traffic volume is contributed by high activity peaks, and volume-based thresholds must still accommodate a significant portion of such peaks, instead of ignoring them completely as in the time slot-based case. Thus, volume-based service specifications at the antenna level force the operator to deploy a substantial amount of resources per slice even under more relaxed guaranteed demands.

Statistics are very different when including overbooking in the picture. The bottom part of Figure 12 illustrates the impact of the overbooking penalty (π), when the full demand is guaranteed, *i.e.*, $\delta = 1$. The plots refer again to pairs of scenarios, under slice specifications expressed in terms of time slots (bottom-left pair) and traffic volume (bottom-right pair). In almost all settings, the multiplexing efficiency quickly rises beyond 0.8 by just having 3% of the slice traffic not served in isolation. The only exception occurs for traffic volume-based guarantees at the antenna level: in this case, the efficiency gain with π is lower, yet the improvement with respect to the case where δ is varied (top-right pair) is dramatic. These results let us conclude that an overbooking that leads to serving a small portion of traffic peaks in a best-effort fashion is an interesting strategy for the operators, maintaining high standards ($\delta = 1$) with a reasonable increment of resources ($\leq 25\%$).

C. Slicing under dynamic resource orchestration

All previous results refer to cases where resources are statically allocated. We now investigate the multiplexing efficiency of network slicing when the operator can orchestrate network resources in an adaptive way, by re-allocating them to different slices over time. As discussed in Section II-C, this is equivalent to considering a resource reconfiguration interval τ that is shorter than the system observation time in our system model. Specifically, we assume that the operator can reconfigure the resources at each network level ℓ with a fixed periodicity τ which depends on the capabilities of the underlying virtualization technology. In our study, the operator allocates resources optimally to meet all slice specifications in

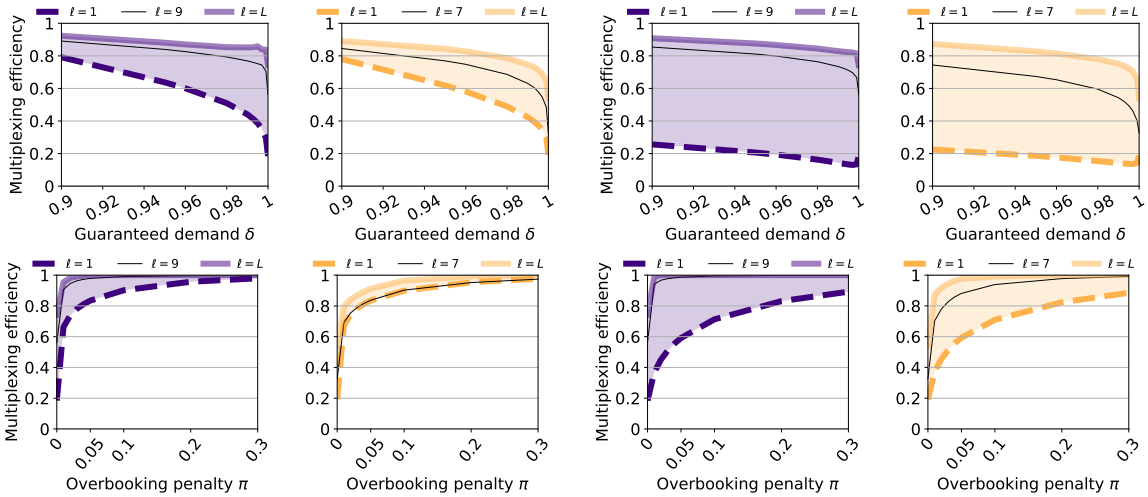


Fig. 12: Efficiency of slice multiplexing versus slice specifications. Top: efficiency versus guaranteed fraction δ of time slots (left pair) and traffic volume (right pair), with $\pi = 0$. Bottom, efficiency versus overbooking penalty π and $\delta = 1$ in time slots (left pair) and traffic volume (right pair). Thick dashed and solid lines represent the extreme network levels $\ell = 1$ and $\ell = L$, while thin solid lines are for an intermediate network level, for the large metropolis (purple) and medium-sized city (gold).

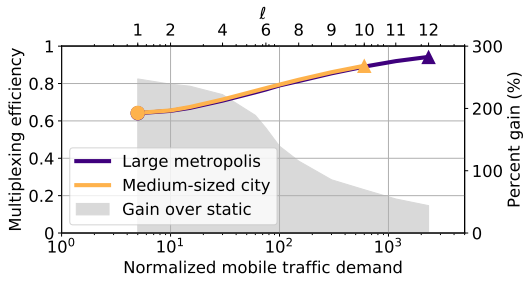


Fig. 13: Efficiency of slice multiplexing (left y axis) and percent gain over static assignment (right y axis) versus the normalized mobile traffic served by one node (bottom x axis) at level ℓ (top x axis) in the two reference urban scenarios. Results are for a dynamic resource assignment where reconfigurations occur with periodicity $\tau = 30$ minutes, under slice specification z with $\delta = 1$ and $\pi = 0$.

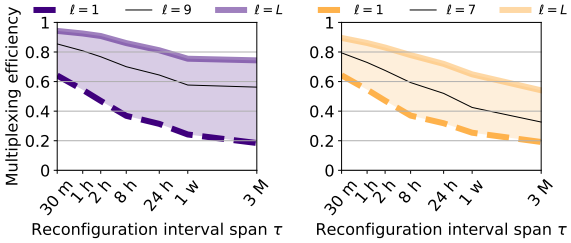


Fig. 14: Efficiency of slice multiplexing versus the resource reconfiguration periodicity τ. Thick dashed and solid lines denote the extreme network levels $\ell = 1$ and $\ell = L$, the thin solid line follows an intermediate network level. Left: large metropolis. Right: medium-sized city.

each reconfiguration interval of duration τ. This is equivalent to assuming the availability of an oracle algorithm that, at the beginning of a reconfiguration interval, has perfect knowledge of the future demand for each service over the rest of the interval. Then, the operator can reserve for each slice the minimum amount of resources to abide by the requirements, as detailed in Section II-C and exemplified in Figure 3.

Our baseline result, in Figure 13, refers to $\tau = 30$ minutes, which can be regarded as a fairly high resource reconfiguration frequency for several scenarios. For instance, VNF management in the network core cloud has typically larger time scales of hours or even days [23]. At radio access, instead, faster dynamic reassignments are technically possible; however, forecasting the demand over short time scales of minutes is challenging and easily leads to slice specification violations, hence reconfiguration intervals in the order of hours are more credible [14]. In these settings, dynamic allocation mechanisms and a perfect prediction of the demand over the future 30 minutes can substantially improve the efficiency of slice multiplexing. Indeed, when comparing the curves in Figure 13 with their equivalent in Figure 10, the gain is evident. We explicitly portray the benefit as the grey region in Figure 13: it ranges between 90% ($\ell = L$) and 250% ($\ell = 1$). The cause of such a significant advantage roots in that different mobile services allocated to separate slices tend to peak at different times of the day, as discussed in details in recent analyses of mobile service dynamics [24]. The temporal diversity of peaks across slices lets a perfect orchestrator reuse the same resources to cover time-disjoint high-activity periods in multiple slices, hence increasing the system efficiency.

Despite the much higher gain at the antenna level, there is still a large gap between the efficiency at the radio access and in the network core. An order-of-minute dynamic orchestration of resources allows for near-perfect slice multiplexing at a datacenter that fully centralizes the traffic in our large metropolis scenario. In contrast, efficiency is bounded at around 0.6 for levels close to $\ell = 1$, *i.e.*, at individual antenna sectors or at nodes serving small groups of a few antennas each. This implies that the operator still has to nearly double the capacity to isolate slices at network levels close to the radio access.

A more comprehensive picture is provided by Figure 14, which encompasses a wide set of reconfiguration intervals τ, from the 30-minutes case we just analyzed in detail up

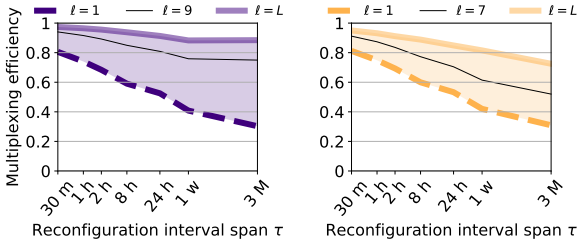


Fig. 15: Efficiency of slice multiplexing with per-category slicing. The plot semantics are the same as in Figure 14.

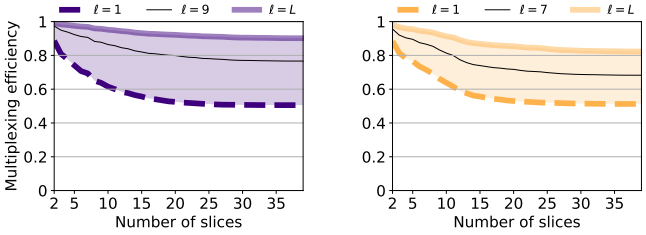


Fig. 16: Efficiency of slice multiplexing as a function of the number of slices $k+1$ (on the x axis), when the k services with the highest traffic load have a dedicated slice and the remaining services are aggregated into a common slice. Thick dashed and solid lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the thin solid line follows an intermediate network level. Left: large metropolis. Right: medium-sized city.

TABLE II: Case studies. Each row maps to one configuration. Columns report the corresponding service category, number of slices, network level, reconfiguration interval, slice specification parameters, and multiplexing efficiency in the large metropolis (LM) and medium-sized city (MC) scenarios.

Category	Slices	Network level	τ	Slice specification			Efficiency	
				Guarantee	δ	π	LM	MC
Streaming	**8**	**Antenna** ($\ell = 1$)	**1h**	**Volume**	**1**	**0**	**0.35**	**0.35**
Streaming	8	MEC ($\ell = 9, 7$)	4h	Volume	1	0	0.74	0.59
Streaming	8	Core ($\ell = L$)	10h	Volume	1	0	0.84	0.73
Web	**6**	**Antenna** ($\ell = 1$)	**1h**	**Volume**	**1**	**0.02**	**0.71**	**0.71**
Web	6	MEC ($\ell = 9, 7$)	4h	Volume	1	0.02	0.97	0.85
Web	6	Core ($\ell = L$)	10h	Volume	1	0.02	1	0.96
Social media	**4**	**Core** ($\ell = L$)	**10h**	**Time slot**	**0.99**	**0.05**	**0.90**	**0.96**
Social media	4	Antenna ($\ell = 1$)	1h	Time slot	0.99	0.05	0.81	0.81
Social media	4	MEC ($\ell = 9, 7$)	4h	Time slot	0.99	0.05	0.92	0.89
Gaming	**6**	**MEC** ($\ell = 9, 7$)	**4h**	**Volume**	**1**	**0**	**0.57**	**0.59**
Gaming	6	Antenna ($\ell = 1$)	1h	Volume	1	0	0.58	0.59
Gaming	6	Core ($\ell = L$)	10h	Volume	1	0	0.57	0.65
Messaging	**5**	**MEC** ($\ell = 9, 7$)	**4h**	**Volume**	**0.99**	**0.03**	**0.68**	**0.80**
Messaging	5	Antenna ($\ell = 1$)	1h	Volume	0.99	0.03	0.5	0.45
Messaging	5	Core ($\ell = L$)	10h	Volume	0.99	0.03	0.91	0.89

to 3 months, *i.e.*, the entire timespan of the dataset, which maps to the static resource configuration case considered in Section IV-A. As one could expect, the multiplexing efficiency of slices is decreased as τ grows, since the system becomes less flexible. Interestingly, the loss of efficiency is steeper at lower values of τ: reducing the frequency of reallocation from once every 30 minutes to once every day yields an efficiency loss comparable to that caused by increasing τ from one day to 3 months. This is consistent with the typical duration of human activities, in the order of tens of minutes, which reflects on similar timescales of mobile service demand fluctuations [24]. Therefore, predicting traffic and allocating resources at longer periodicity rapidly reduces the system efficiency: either the operator is able to deploy virtualization technologies that enable such a reconfiguration frequency, or it is probably not worth considering dynamic resource allocation at all.

D. Varying number of slices

Up to now, we assumed that the slicing strategy adopted by the operator involved assigning one slice to each of the 16 services that generate the most traffic. In fact, the mapping of services into specific network slice instances is a business-driven choice that is based on several factors, such as the requirements of the services in terms of isolation, the specific policies implemented by the operator [25], or the practice of the tenants, which may decide to group multiple services into a same slice for economic reasons. The number of slices and the demands associated to each will have an impact on the overall multiplexing efficiency, which we investigate next.

We first analyze a business-driven scenario where network slices are dedicated to sets of services of a same category, *i.e.*, streaming, social media, web, cloud, gaming, messaging and miscellanea, respectively. Here, we set $\delta = 1$ and $\pi = 0$ for all slices. In this scenario, we study the impact of the system reconfiguration dynamics, as displayed in Figure 15. Trends are similar to those observed for a per-service slicing in Figure 14. Despite a higher efficiency in general, the fractional gain brought by increasingly faster resource orchestration is comparable under the two different slicing policies.

We then explore different slicing strategies according to a hierarchical scheme where the k services that generate the highest traffic loads acquire a dedicated slice each. The demands for all remaining services are instead aggregated into a common, non-customized, slice. Figure 16 shows the resulting multiplexing efficiency as a function of the total number $k+1$ of slices in the network, when the reconfiguration period τ is set to 1 hour, $\delta = 1$ and $\pi = 0$ for all slices. Increasing the number k of isolated mobile services entails a reduction of efficiency: this is expected, since a larger k moves traffic from the common slice, within which multiplexing is perfect, to dedicated slices that require isolated resources. Interestingly, however, the loss of efficiency is accumulated in the first half of the plots, *i.e.*, considering a number of slices larger than 16 does not affect efficiency anymore. Therefore, most of the resource utilization cost for the operator comes from the very few mobile services that generate the largest demands, and multiplexing efficiency is only increased when such services are treated as best-effort traffic. Incidentally, these results also motivate our choice of focusing on 16 slices in previous experiments: this setting maps to a lower bound on performance in terms of efficiency with our dataset.

E. Case studies

To conclude our evaluation, we investigate the multiplexing efficiency under network slicing in a number of specific case studies. This analysis lets us detail particular settings of practical interests, and complements the previous results where each system parameter was studied in isolation. Each

case study focuses on a specific service category (*e.g.*, video streaming), where we assume that different applications (*e.g.*, YouTube, iTunes, DailyMotion, Netflix, etc.) are allocated to isolated slices. The detailed configurations and the associated efficiency results are provided in Table II, for both the large metropolis and medium-sized city scenarios. Our analysis below addresses one network level in each case study, highlighted in bold in Table II. For the sake of completeness, the table also includes additional levels for each scenario, which allow appreciating, for each case study, the efficiency of end-to-end slicing across the network architecture.

Case study #1 – High QoS at the access network. The first case study focuses on slicing at the antenna level, and on capacity-demanding services such as video streaming and web access. These are challenging settings for the operator, who must provide high-quality support for a large volume of bursty traffic; a quite fast reconfiguration (τ = 1 h) is thus a reasonable relief. The efficiency is nonetheless low if hard-QoS requirements (δ = 1) are to be met, *e.g.*, for video streaming slices: the operator shall commit up to threefold the resources needed in a non-sliced scenario – a high cost considering that radio access resources such as spectrum or RAN processing capacity can be very expensive. In less strict slices like those dedicated to web access, paying minimal overbooking penalties (π = 0.02) is an appealing option, as it may reduce costs considerably by raising efficiency to 0.71.

Case study #2 – Large traffic flows in the core. This case study shifts the focus to datacenters in the network core, and targets social media services that generate high demands but are less latency-dependent. The large traffic volumes observed at this level allow achieving high efficiency (above 0.90) under loose QoS (δ = 0.99, π = 0.05), and with limited reconfiguration possibilities (τ = 10 h). These results further prove the benefits of centralization for the effective implementation of network slicing.

Case study #3 – Computing at the edge. Gaming services with strong QoS requirements are likely candidates to be among the first services to be delivered over edge deployments [26], hence they represent a sensible target for MEC-level slicing. We consider 6 popular mobile games, to which we allocate dedicated slices with firm specifications (δ = 1, π = 0). Although we allow for quite fast reconfiguration (τ = 4h), the price that the operator has to pay is high in both urban scenarios: the required resources are almost doubled with respect to a non-sliced network. Similar considerations hold for messagging services, although in this case QoS requirements can be moderated to δ = 0.99 π = 0.03, with a 20-30% efficiency gain with respect to the gaming case.

V. Related work

Multi-service networks [27] are a key building block for the implementation of the network slicing paradigm [28] that, in turn, will enable new business models such as multi-tenancy [12] and finally pave the way to 5G. At this stage, the bulk of the work on next generation network sharing architectures is already available, ranging from novel visions of the network [20] to specific architectures proposals [29]. More specifically, research work already addressed the extension to multi-service settings of fundamental parts of the 5G system, such as the Radio Access Network (RAN) [30], [31], the core network [32], or the management and orchestration components [33]. Such research effort is already making its way into standardization: 3GPP considered multi-service and network slicing aspects for the next Release 15 [34].

On top of the architectural research work, enabling multi-service networks has also been considered from an algorithmic point of view. The focal point of research in this area has been RAN resource allocation [35], [11], [36], as oversubscribing spectrum is especially difficult. However, resource sharing has also been tackled for other kinds of virtualized functions [13].

Despite the attention that multi-service networks, network slicing and multi-tenant networks have been receiving for the last few years, little attention has been paid to how such network slices will behave in practical scenarios. Understanding the system efficiency *in the wild* has only been possible in reduced scenarios involving very few devices [11], or by making assumption on the real patterns, modelling user movements and service requests with random processes [37]. The only works that employ a data source comparable to ours are the one in [38] and our seminal work in [39].

VI. Discussion and conclusions

Our data-driven analysis unveils how real-world service usage patterns may affect the deployment of a key paradigm for future-generation mobile networks such as network slicing, and the impact it has on resource management. Specifically, we retain the following main takeaway messages.

Multi-service requires more resources. Building a network that is capable of providing different services (possibly associated to several tenants) will necessarily reduce efficiency in resource utilization. We quantify this loss in almost one order of magnitude if considering distributed resources (such as spectrum), yet the efficiency loss stays as high as 20% even in large datacenters in the core network. These figures translate into high costs for the infrastructure provider, who must compensate for them by aggressively monetizing on the new business models enabled by a multi-service scenario (*e.g.*, Network Slice as a Service, Infrastructure as a Service).

Traffic direction is a factor. Uplink and downlink traffic exhibit similar efficiency trends across network levels, but uplink exacts a much higher efficiency degradation to meet equivalent QoS requirements. Although uploads account for a small fraction of the overall load, the lower efficiency of uplink may entail additional challenges for the operators. Indeed, uplink QoS requirements are key to specific services such as mobile gaming, and it is likely that multiple instances of such services belonging to different tenants have to be served in a resource-isolated fashion in parallel.

Loose service level agreements may not help. Although slice specifications may be moderated, the overall efficiency grows only when guarantees on the serviced demand are very much lowered, up to a point that they may not be suitable for certain services (needing, *e.g.*, "5 nines reliability", or strict bandwidth requirements over very short time slots).

Overbooking is a key strategy. While downgrading the requirements in terms of served fraction of traffic only helps when brought to extreme levels, flexibly serving small portions of the individual slice demands via a non-customized common slice provides high benefits. Therefore, overbooking solutions that only marginally underserve slices may yield substantial economic gains for the operators, as they allow trading off substantial resource deployment costs with negligible penalty fees due to slight SLA violations. This corroborates the importance of recent approaches for practical end-to-end resource overbooking in sliced 5G networks [17].

Guaranteeing traffic volumes at the antenna is costly. If operators define SLAs in terms of assured traffic volumes, they shall note that meeting the QoS requirements will need substantial additional resources at the radio access, even if guarantees are loose and overbooking is in place. SLAs defined in terms of guaranteed time slots allow much more flexibility in balancing efficiency and QoS for each network slice.

Dynamic resource assignment must be rapid. The design of dynamic resource allocation algorithms is crucial to increase the efficiency of future sliced networks. However, substantial gains will only be attained if the virtualization technologies enable a fast enough re-orchestration of network resources. While current Management and Orchestration (M&O) frameworks provide such capabilities, intelligent algorithms able to forecast mobile service demands and anticipate resource reconfiguration are also required, Artificial intelligence and machine learning are promising techniques to accomplish this [40], and are also being brought into the network management landscape by standards [41].

Aggregating services is beneficial. Aggregating similar services into a same slice increases the system efficiency significantly, yet it comes at the price of losing the ability to customize treatment to each service. This implies that operators may face a business trade-off between providing dedicated support to highly remunerative, popular services, and incurring high management costs to implement the associated slices.

Urban topography has limited impact. The fact that all of our results are very consistent in two urban areas with a quite different nature lets us provide general insights that hold beyond one particular scenario. More precisely, as usage demands are eventually driven by human factors, we expect that our considerations might apply to other metropolitan regions in (and possibly beyond) Europe.

Efficiency under uncertain load demands. Our analysis concerns resource management efficiency under known loads, as slices are allocated the exact resources needed to meet the corresponding service demands. This lets us investigate the impact of the limited reconfigurability of resources, which forces the operator to provision a constant amount of resources during the following reconfiguration period. In a real system, however, the network slices demands are not known *a priori*, and resources have to be allocated based on a forecast of the expected demand during the next re-orchestration interval. This introduces a second source of inefficiency, *i.e.*, the inaccuracy of traffic predictions, which imposes some overprovisioning in the allocated capacity to combat the uncertainty associated with the future load information. This second aspect has been recently analyzed by the authors in [42], where an approach is developed that forecasts the capacity needed to accommodate the traffic of a slice. Figures about the expected global performance of a practical system can then be obtained by summing the effects of both sources of inaccuracy. For instance, if the resource reconfiguration periodicity imposes allocating 100% extra resources (which is a typical case according to the results in the previous sections), and capacity predictors entail 10% overprovisioning (a likely number according to [42]), then the overall additional resources required will amount to 110%. This extra capacity can then be served with a mixture of guaranteed demand and overbooking, as discussed in Section II. While a thorough analysis of the overall efficiency resulting from considering both effects is left as future work, it is worth mentioning that, according to the results presented in this paper and in [42], it is expected that the overall efficiency will be dominated by the resource allocation dynamics analyzed in this paper.

To conclude, ours does not pretend to be a fully comprehensive analysis, rather one that lays the foundations to a better understanding of the new trade-offs introduced by network slicing in terms of resource management efficiency. The empirical bounds we derived represent a starting point for deeper investigations of an unexplored subject with strong implications for the future generations of mobile networks.

ACKNOWLEDGMENTS

The work of University Carlos III of Madrid was supported by the H2020 5G-MoNArch project (Grant Agreement No. 761445) and the work of NEC Laboratories Europe by the 5G-Transformer project (Grant Agreement No. 761536). The work of CNR-IEIIT was partially supported by the ANR CANCAN project (ANR-18-CE25-0011).

REFERENCES

[1] K. Campbell *et al.*, "The 5G economy: How 5G technology will contribute to the global economy," IHS Economics & IHS Technology Economic Impact Analysis, Jan. 2017.

[2] ITU, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," Report ITU-R M.2410-0, Nov. 2017.

[3] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-QoS-Aware Fair Scheduling for LTE," in *IEEE 73rd Vehicular Technology Conference (IEEE VTC 2011 Spring)*, May 2011.

[4] 3GPP, "Policy and charging control architecture," TS 23.203 version 12.6.0 Release 12.

[5] 3GPP, "Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT)," 3GPP Technical Report (TR) 45.820, Aug. 2015.

[6] X. Li, *et al.*, "A review of industrial wireless networks in the context of industry 4.0," *Wireless Networks*, vol. 23, no. 1, pp. 23–41, Jan. 2017.

[7] Google, "Google project fi." [Online]. Available: https://fi.google.com/about/

[8] P. Rost *et al.*, "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, May 2016.

[9] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, Oct. 2017.

[10] A. Ksentini and N. Nikaein, "Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102–108, Jun. 2017.

[11] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture," in *23rd Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2017)*, Oct. 2017.

[12] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[13] J. G. Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, Sep. 2016.

[14] V. Sciancalepore *et al.*, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*, May 2017.

[15] S. Sharma *et al.*, "Dynamic Spectrum Sharing in 5G Wireless Networks With Full-Duplex Technology: Recent Advances and Research Challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 674–707, Feb. 2018.

[16] M. Odini, "OpenSource MANO," IEEE Softwarization: A Collection of Short Technical Articles, Jul. 2016. [Online]. Available: https://sdn.ieee.org/newsletter/july-2016/opensource-mano

[17] J.X. Salvat *et al.*, "Overbooking Network Slices through Yield-driven End-to-End Orchestration," in *ACM 14th International Conference on emerging Networking EXperiments and Technologies (ACM CoNEXT 2018)*, Dec. 2018.

[18] T. L. Nguyen and A. Lebre, "Virtual Machine Boot Time Model," in *25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2017)*, Mar. 2017.

[19] 5th Generation Public Private Partnership (5G-PPP), "View on 5G architecture (v. 2.0)," 5G-PPP Architecture WG White Paper, Dec. 2017.

[20] P. Rost *et al.*, "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 84–91, May 2016.

[21] P. Sanders and C. Schulz, "Think Locally, Act Globally: Highly Balanced Graph Partitioning," in *International Symposium Experimental Algorithms (SEA 2013)*, Jun. 2013.

[22] S. Bhaumik *et al.*, "CloudIQ: a framework for processing base stations in a data center," in *18th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2012)*, Aug. 2012.

[23] F. Z. Yousaf and T. Taleb, "Fine-grained resource-aware virtual network function management for 5G carrier cloud," *IEEE Network*, vol. 30, no. 2, pp. 110–115, Mar. 2016.

[24] C. Marquez *et al.*, "Not All Apps Are Created Equal: Analysis of Spatiotemporal Heterogeneity in Nationwide Mobile Service Usage," in *13th International Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT 2017)*, Dec. 2017.

[25] 3GPP, "Telecommunication management; study on management and orchestration of network slicing for next generation network (release 15)," 3GPP Technical Report (TR) 28.801, Jan. 2018.

[26] D. Telekom, "Deutsche telekom, niantic and mobiledgex partnership." [Online]. Available: http://bit.ly/dt_niantic

[27] P. Rost *et al.*, "Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72–79, May 2017.

[28] Next Generation Mobile Networks (NGMN) Alliance, "Description of network slicing concept," NGMN White Paper, Feb. 2015.

[29] N. Nikaein *et al.*, "Network Store: Exploring Slicing in Future 5G Networks," in *10th International Workshop on Mobility in the Evolving Internet Architecture (ACM MobiArch 2015)*, Sep. 2015.

[30] I. F. Akyildiz, P. Wang, and S. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1–18, Jul. 2015.

[31] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks," in *12th International on Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT 2016)*, Dec. 2016.

[32] M. R. Sama, X. An, Q. Wei, and S. Beker, "Reshaping the mobile core network via function decomposition and network slicing for the 5G Era," in *2016 IEEE Wireless Communications and Networking Conference (IEEE WCNC 2016)*, Apr. 2016.

[33] A. Mayoral *et al.*, "Multi-tenant 5G Network Slicing Architecture with Dynamic Deployment of Virtualized Tenant Management and Orchestration (MANO) Instances," in *42nd European Conference and exhibition on Optical Communication (ECOC 2016)*, Sep. 2016.

[34] 3GPP, "NR and NG-RAN Overall Description, Stage-2 (Release 15)," 3GPP Technical Specification (TS) 38.300, Jan. 2018.

[35] P. Caballero *et al.*, "Network slicing games: Enabling customization in multi-tenant networks," in *IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*, May 2017.

[36] Y. L. Lee *et al.*, "Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[37] D. Bega *et al.*, "Optimising 5G infrastructure markets: The business of network slicing," in *IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*, May 2017.

[38] A. Okic *et al.*, "Analyzing Different Mobile Applications in Time and Space: a City-Wide Scenario," in *2019 IEEE Wireless Communications and Networking Conference (IEEE WCNC)*, Apr. 2019.

[39] C. Marquez *et al.*, "How should i slice my network?: A multi-service empirical evaluation of resource sharing efficiency," in *24th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2018)*. ACM, 2018, pp. 191–206.

[40] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Comm. Surveys Tutorials*, 2019.

[41] European Telecommunications Standards Institute (ETSI), "Improved operator experience through Experiential Networked Intelligence (ENI)," ETSI White Paper No. 22, Oct. 2017.

[42] D. Bega *et al.*, "DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning," in *IEEE International Conference on Computer Communications (IEEE INFOCOM 2019)*, Apr. 2019.

Cristina Marquez is a Ph.D. student at Universidad Carlos III of Madrid (UC3M) under the supervision of Prof. A. Banchs. She obtained her Double Master Degree at the same University, completing the M.Sc. degree in Telematics Engineering in 2017 and in Telecommunication Engineering in 2018, while pursuing a Ph.D. in Telematics Engineering. She has been accepted as visiting student at MIT before finishing the PhD. Her areas of research are Big Data Analytics, Resource Management, Wireless Networks and Mobile Networks (5G).

Marco Gramaglia is a post-doc researcher at University Carlos III of Madrid (UC3M), where he received M.Sc (2009) and Ph.D (2012) degrees in Telematics Engineering. He held post-doctoral research positions at ISMB (Italy), the CNR-IEIIT (Italy) and IMDEA Networks (Spain). He was involved in EU projects and authored more than 40 papers appeared in international conference and journals.

Marco Fiore (S'05, M'09, SM'17) is a researcher at CNR-IEIIT (Italy) a Royal Society visiting research fellow, and a Marie Curie fellow. He received a PhD degree from Politecnico di Torino (Italy) and a HDR degree from Univeristé de Lyon (France). He was associate professor at INSA Lyon, (France), associate researcher at Inria (France), visiting researcher at Rice University (USA) and UPC, (Spain), and visiting research fellow at UCL (UK). His current research interests are in the fields of mobile networks, network traffic analytics and privacy.

Albert Banchs (M'04-SM'12) received the M.Sc. and Ph.D. degrees from the Polytechnic University of Catalonia (UPC-BarcelonaTech) in 1997 and 2002, respectively. He is currently a Full Professor with the University Carlos III of Madrid (UC3M), and has a double affiliation as Deputy Director of the IMDEA Networks institute. Prof. Banchs has served in many conference TPCs and journal editorial boards, and is currently Editor of IEEE Transactions on Wireless Communications and IEEE/ACM Transactions on Networking.

Xavier Costa-Pérez (M'06-SM'18) is Head of 5G Networks R&D and Deputy General Manager of the Security and Networking Research Division at NEC Laboratories Europe. His team contributes to products roadmap evolution as well as to European Commission projects and received several awards for successful technology transfers. Dr. Costa is a 5GPPP Technology Board member, has served on the Program Committees of several conferences (including IEEE Greencom, WCNC, and INFOCOM), published at top research venues and holds several patents. He received both his M.Sc. and Ph.D. degrees in Telecommunications from the Polytechnic University of Catalonia (UPC) in Barcelona and was the recipient of a national award for his Ph.D. thesis.