How Should I Slice My Network? A Multi-Service Empirical Evaluation of Resource Sharing Efficiency

Cristina Marquez
Universidad Carlos III Madrid
Spain
mcmarque@pa.uc3m.es

Marco Gramaglia
Universidad Carlos III Madrid
Spain
mgramagl@it.uc3m.es

Marco Fiore
CNR-IEIIT
Italy
marco.fiore@ieiit.cnr.it

Albert Banchs
Universidad Carlos III Madrid &
IMDEA Networks Institute
Spain
banchs@it.uc3m.es

Xavier Costa-Perez
NEC Laboratories Europe
Germany
xavier.costa@neclab.eu

ABSTRACT

By providing especially tailored instances of a virtual network, *network slicing* allows for a strong specialization of the offered services on the same shared infrastructure. Network slicing has profound implications on resource management, as it entails an inherent trade-off between: *(i)* the need for fully dedicated resources to support *service customization*, and *(ii)* the dynamic resource sharing among services to increase *resource efficiency* and cost-effectiveness of the system. In this paper, we provide a first investigation of this trade-off via an empirical study of resource management efficiency in network slicing. Building on substantial measurement data collected in an operational mobile network *(i)* we quantify the efficiency gap introduced by non-reconfigurable allocation strategies of different kinds of resources, from radio access to the core of the network, and *(ii)* we quantify the advantages of their dynamic orchestration at different timescales. Our results provide insights on the achievable efficiency of network slicing architectures, their dimensioning, and their interplay with resource management algorithms.

CCS CONCEPTS

• **Networks → Mobile networks**; *Network architectures*; *Network performance evaluation*; *Network management*;

MobiCom '18, October 29-November 2, 2018, New Delhi, India

ACM ISBN 978-1-4503-5903-0/18/10…$15.00
https://doi.org/10.1145/3241539.3241567

KEYWORDS

Network slicing; resource management; network efficiency

ACM Reference Format:
Cristina Marquez, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. 2018. How Should I Slice My Network? A Multi-Service Empirical Evaluation of Resource Sharing Efficiency. In *The 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18), October 29-November 2, 2018, New Delhi, India.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3241539.3241567

1 INTRODUCTION

Current trends in mobile networks point towards a strong diversification of services, which are characterized by increasingly heterogeneous Key Performance Indicator (KPI) and Quality of Service (QoS) requirements. This tendency is driving the design of 5G networks that will eventually have to support, *e.g.*, the Internet of Thing (IoT) with ultra-low rate communication from a massive number of devices, automotive and tactile applications with millisecond latencies, industrial communications with extreme reliability, and virtual/augmented reality services with very high data rates.

However, clear needs for tailored KPI and QoS requirements are already evident in today's mobile services, which encompass, *e.g.*, high-quality video streaming, machine-type communication, low-latency mobile gaming, jointly with best effort traffic. Unfortunately, current mobile network architectures [33] lack the necessary flexibility to meet the extreme requirements imposed by such services. This situation is pushing independent initiatives to address the problem. 3GPP has developed a IoT-specific MAC that co-exists with the legacy general-purpose MAC layer [1]. Network deployments in industrial environments rely on proprietary architectures that ensure reliability levels not attainable with public mobile networks [17]. Google started deploying its own

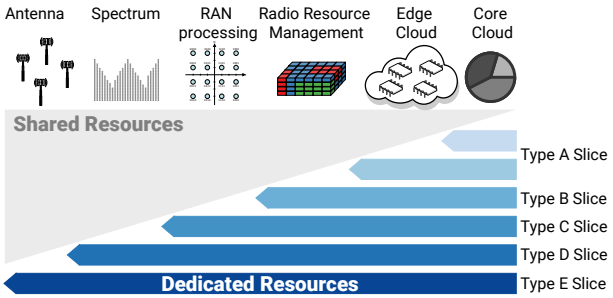

Figure 1: Network slicing types. Deeper strategies use dedicated resources customized to services across a wider portion of the end-to-end network architecture.

radio access infrastructure and proprietary transit network to run its many services under hard QoS guarantees [13].

Network virtualization and slicing. While their scope is clearly limited, these solutions do show the need for customized network support even with present-day traffic. They also substantiate the well-established vision that several network instances, each devoted to a specific set of services, have co-exist in the same infrastructure in order to satisfy the KPI and QoS requirements of current and future mobile applications. The agenda for 5G networks is to achieve this mainly via *network virtualization*, which evolves the traditional hardbox paradigm into a cloudified architecture where the once hardware-based network functions (*e.g.*, spectrum management, baseband processing, mobility management) are implemented as software Virtual Network Functions (VNFs) running on a general-purpose *telco-cloud*. Network virtualization enables the deployment of multiple virtual instances of the complete network, named *network slices*. Slices are then easily customized by tuning the functionality and location of VNFs. They thus create on top of the physical infrastructure a set of logical networks, each tailored to accommodate fine-tuned Service Level Agreements (SLA) reflecting the needs of different service providers.

Network slicing and resource management. Network slicing has profound implications on resource management. When instantiating a slice, an operator needs to allocate sufficient computational and communication resources to its VNFs. In some cases, these resources may be dedicated, becoming inaccessible to other slices [26]. Alternatively, smart assignment algorithms can be employed to dynamically allocate resources to slices based on the time-varying demands of tenants [11, 15]. 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, such algorithms introduce additional complexity, and may in some cases hinder resource isolation, the corresponding guarantees to tenants, and/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 employing elaborate operation and maintenance functions [28].

The above 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 [14]. 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 [30] 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 at the same time they reduce the possibility of radio resource sharing and/or increase the system complexity.

At fronthaul, resource isolation becomes a hardware problem [31]. 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.

In general, slicing strategies at the higher network layers provide a lower level of customization yet they can more easily achieve efficient 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 customization 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 is not tailored to their different needs; conversely, if each slice implements a different, customized MAC protocol, it is more difficult to efficiently share radio resources.

Contribution of this paper. From a system standpoint, the technology needed to support the different types of slices is well understood or even already available. For instance, there exist several cloud resource orchestrators for both commercial and open-source telco-cloud platforms [23]; similarly, a variety of solutions have been proposed for the dynamic allocation of resources across network slices [14].

However, the implications of network slicing in terms of efficiency of 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 decision 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. Based on our analysis, it is thus possible to determine in which cases the gains in efficiency are worth the sacrifice in customization/isolation and/or the extra complexity. Since resource management efficiency in network slicing highly depends on the traffic patterns of different services supported by the various slices, we build on substantial service-level measurement data collected by a major operator in a production mobile network, and:

(i) quantify the price paid in efficiency when suitable algorithms for dynamic resource allocation are not available, and the operator has to resort to physical network duplication;

(ii) evaluate the impact of sharing resources at different locations of the network, including the cloudified core, the virtualized radio access, or the individual antennas;

(iii) outline the benefit of dynamic resource allocation at different timescales, *i.e.*, allowing to reallocate resources across slices with different reconfiguration intervals.

To the best of our knowledge, this is the first work tackling the empirical assessment of network slicing in real-world networks. We believe that the insights it provides can be used as rule of thumb to evaluate the solution space for smart resource assignment algorithms and infrastructure dimensioning. For instance, our results show that efficiency gains are very high in the edge, where employing technologies that allow for dynamic resource allocation provides a high reward; in contrast, gains are much reduced in the core, where complex, highly flexible reconfiguration schemes may not always pay off. Mobile network operators should thus be aware that isolating slices at the radio access may have a high cost in terms of efficiency, and that network slicing should be combined with solutions for dynamic orchestration of resources, at least at the network edge.

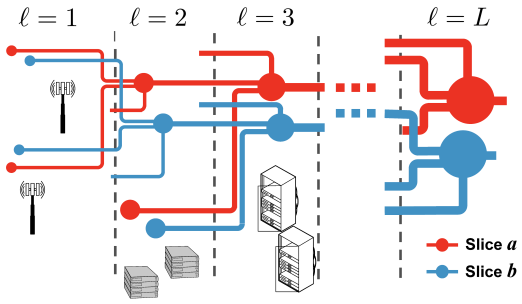

Figure 2: Hierarchical mobile network architecture. Nodes map to different equipment depending on the level ℓ, and form a hierarchy. The mobile traffic of services in each slice (*e.g.*, a or b) is increasingly aggregated as it flows from radio access to network core.

2 NETWORK SCENARIO AND METRICS

In the following we expose our network scenario, our representation of the slice QoS requirements and a consistent resource allocation strategy, and the metrics we adopt to evaluate the resource sharing performance.

2.1 Network slicing scenario

Let us consider a mobile network providing coverage to a generic geographical region, where mobile subscribers consume a variety of heterogeneous services. The operator owning the infrastructure 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 associate (possibly virtual) network functions. At the most distributed level ($\ell = 1$), each node runs functions that operate at the antenna level, *e.g.*, involving spectrum or airtime. In intermediate cases ($1 < \ell < L$), nodes are at first in charge 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, where the operator can tailor cloud resources to the whole demand for the services of a tenant. Note that, in the case of VNFs, this allows to evaluate the impact of instantiating or moving VNFs at different nodes.

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 for radio access resources, which may run at the base stations ($\ell = 1$) in a distributed implementation, or at higher architectural levels ($1 < \ell < L$) in a centralized 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.

2.2 Slice specifications

Network slicing allows the operator to fulfil minimum QoS requirements requested by each tenant. We capture such requirements as a *slice specification* z, which is established so as to ensure a sufficient service quality for the slice demands. More precisely, a slice specification involves:

(i) Guaranteed time fraction f: the operator engages to guarantee that the traffic demand of the slice is fully serviced during at least a fraction $f \in [0, 1]$ of time.

(ii) Averaging window length w: the operator commitment on fraction f above is intended on discrete-time demands of granularity w, with traffic averaged over the disjoint time windows of duration w.

We denote such a slice specification as $z = (f, w)$, which becomes more stringent for higher values of f and smaller w.

To ensure compliance with the requirements, the operator shall guarantee that enough resources are allocated to all slices $s \in \mathcal{S}$ at every node $c \in C_\ell$ of each network level ℓ. Formally, the required amount of resources needed to meet a slice specification $z = (f, w)$ is computed as follows. Let $o_{c,s}(t)$ denote the load offered by slice s at node c and time t; also, let $\overline{o}_{c,s}(k) = \frac{1}{w} \int_k o_{c,s}(t)\, \mathrm{d}t$ be the average load over window k covering a time interval of the same name with duration w. Let us also denote by $r^z_{c,s}(k)$ the amount of resources allocated to slice s at node c during window k. According to the above requirements, $r^z_{c,s}(k)$ has to be set such that the following inequality holds

$$P\Big(r^z_{c,s}(k) \geq \overline{o}_{c,s}(k)\Big) \geq f, \tag{1}$$

where $P(\cdot)$ denotes the probability of the argument. Basically, Equation (1) states that the resources allocated should meet the demand for at least a fraction f of averaging windows.

Note that the expression in Equation (1) assumes that the amount resources needed to serve a given slice, $r^z_{c,s}(k)$, is directly proportional to the mobile traffic demand in that slice, $\overline{o}_{c,s}(k)$. While this clearly holds for some types of resources (*e.g.*, radio), we acknowledge that it may be a strong simplification in other cases. 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.

2.3 Resource allocation to slices

In presence of algorithms that enable a dynamic reconfiguration of VNFs, the resource allocation can be re-modulated over time. If, at some node c, one could reallocate resources at every averaging window, it would be sufficient to assign to a slice s the resources it requires during that window with probability at least f, according to Equation (1).

However, in practice the periodicity of reconfiguration is limited by the adopted slicing strategy (see Figure 1) as well as by the constraints of the underlying technology. For instance, when network slicing is performed at the antenna level, non-negligible 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

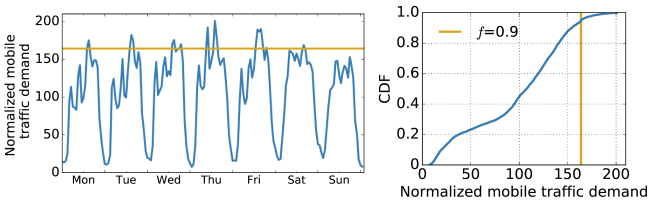

Figure 3: Example of resource allocation to a slice with specification $z = (f, w)$. Left: time series of the mobile traffic demand for a slice s dedicated to a popular video streaming service. The time series refers to traffic averaged over windows of length $w = 1$ hour, recorded at a datacenter c serving a medium-sized city ($\ell = L$) during one reconfiguration interval n ($\tau = 1$ week). Right: CDF $F^w_{s,c,n}$ of the demand in the left plot. The guaranteed time fraction is $f = 0.9$, hence the minimum resources $\hat{r}^z_{c,s}(n)$ to meet the service requirements under slice specification $z = (f, w) = (0.9, 1$ hour$)$ is the 90$^{\text{th}}$ percentile of the distribution, highlighted by the vertical line. The same value is shown in the left plot as a horizontal line: traffic above it is not guaranteed.

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 times [20].

Let us assume that $\tau \gg w$ is the minimum time needed for resource reallocation, which we refer to as a reconfiguration period. Let us denote by $n \in \mathcal{T}$ the n^{th} reconfiguration period within the set $\mathcal{T}$ of all the reconfiguration periods that compose the whole system observation time; also, $\hat{r}^z_{c,s}(n)$ is the amount of resources allocated to slice s in node c during the reconfiguration period n, under specification z. Since no reassignment is possible within a reconfiguration period, then $r^z_{c,s}(k) = \hat{r}^z_{c,s}(n)$, for all averaging windows k within reconfiguration period n. In compliance with Equation (1), the allocation of resources at reconfiguration period n shall be such that the offered load does not exceed $\hat{r}^z_{c,s}(n)$ for at least a fraction f averaging windows encompassed by n. Let $F^w_{s,c,n}$ be the Cumulative Distribution Function (CDF) of the demand for slice s at node c during reconfiguration period k, averaged over windows of length w: then, the minimum $\hat{r}^z_{c,s}(n)$ that satisfies Equation (1) can be computed as $\hat{r}^z_{c,s}(n) = (F^w_{s,c,n})^{-1}(f)$. Figure 3 illustrates this concept[1].

Once we have computed $\hat{r}^z_{c,s}(n)$, we can define the amount of resources that the operator will need to allocate at network level ℓ over the entire system observation period as

$$\mathbb{R}^z_{\ell,\tau} = \sum_{s\in\mathcal{S}} \sum_{c\in C_\ell} \sum_{n\in\mathcal{T}} \tau \cdot \hat{r}^z_{c,s}(n). \tag{2}$$

[1] All traffic volumes in the paper are normalized with respect to the minimum average traffic recorded at a 4G antenna sector in our reference scenarios.

The above equation represents the total amount of resources needed to meet slice specifications z, under the possibility of dynamically re-configuring the allocation with periodicity τ. Note that it can accommodate the special case where no reconfiguration is possible at level ℓ, by setting τ to the total system observation time, *i.e.*, $|\mathcal{T}| = 1$.

2.4 Multiplexing efficiency

Equation (2) provides the total amount of resources that the operator needs to provision in order to satisfy the commitments with all tenants. 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. Formally,

$$\mathbb{P}^z_{\ell,\tau} = \sum_{c\in C_\ell} \sum_{n\in\mathcal{T}} \tau \cdot \hat{r}^z_c(n), \tag{3}$$

where $\hat{r}^z_c(n)$ denotes the resources needed to accommodate the traffic demand at node c during reconfiguration period n, aggregated over all slices. For the sake of fairness, the same specification $z = (f, w)$ assumed for individual slices are enforced in the benchmark provided by Equation (3). Thus, $\hat{r}^z_c(n) = (F^w_{c,n})^{-1}(f)$, where $F^w_{c,n}$ is the CDF of the total demand for mobile data traffic at node c during reconfiguration period n, averaged over windows of length w.

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

$$\mathbb{E}^z_{\ell,\tau} = \mathbb{R}^z_{\ell,\tau} \;/\; \mathbb{P}^z_{\ell,\tau}. \tag{4}$$

Equation (4) refers to network level ℓ, resource reconfiguration intervals of duration τ, and slice specification z.

In summary, $\mathbb{E}^z_{\ell,\tau}$ quantifies the efficiency of the network slicing paradigm in terms of resource management: as $\mathbb{E}^z_{\ell,\tau}$ approaches 1, the total amount of slice-isolated resources tend to that assured by a perfect sharing. Indeed, with perfect sharing we can allocate resources at a given level according to the total peak demand over the reconfiguration period, while with network slicing we need to allocate resources according to the peak demand at each slice, which becomes inefficient when such peaks occur at different windows. Figure 4 illustrates the intuition behind multiplexing efficiency with an example.

3 CASE STUDIES

We evaluate the efficiency of resource allocation in a sliced network by considering two realistic case studies in modern metropolitan-scale mobile networks. As mentioned in the introduction, today's mobile services already offer a variety of requirements that makes it meaningful to investigate the impact of slice isolation on resource management.

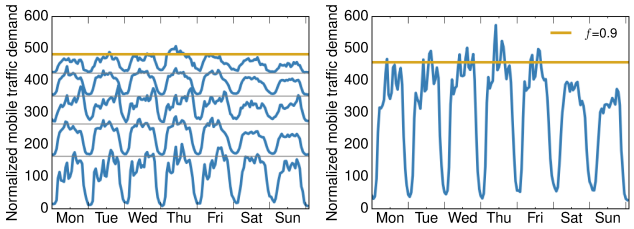

Figure 4: Example of multiplexing efficiency. Left: time series of the mobile traffic demands for a set $\mathcal{S}$ of five slices s, observed at a single datacenter c serving one medium-sized city ($\ell = L$), during one reconfiguration interval n ($\tau = 1$ week). The slice specification $z = (f, w) = (0.9, 1$ hour) commits the operator to serve, for each slice, at least the traffic volumes highlighted by the grey horizontal lines (computed for each time series as in Figure 3). The sum value of such lines, in thick gold, denotes $\sum_{s\in\mathcal{S}} \tau \cdot \hat{r}^z_{c,s}(n)$: as we are looking at a single node c and one specific reconfiguration interval n, this is the traffic volume that provides the needed resources according to Equation (2). Right: time series of the traffic demand aggregated over all services for the same set of slices. By applying an identical slice specification, we get the equivalent traffic volume $\hat{r}^z_c(n)$ to be served under *perfect sharing* as per Equation (3); this is highlighted by the horizontal thick gold line. The multiplexing efficiency is the ratio between the values highlighted by the thick gold lines on the right and left plots. In this toy example, the two values are close, hence resource isolation is efficient. In practical scenarios (Section 4.1), we find major differences between network slicing and perfect sharing, and resource isolation proves highly inefficient.

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%. We leverage these real-world traffic demands to define network slices. Details are in Section 3.1.

On top of this, we model the hierarchical network infrastructures in the target regions by assuming that the operator deploys level-ℓ nodes so as to balance the offered load among them. This is discussed in Section 3.2.

3.1 Mobile service demands

The real-world demands generated by individual mobile services in the two reference regions were collected by the operator 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

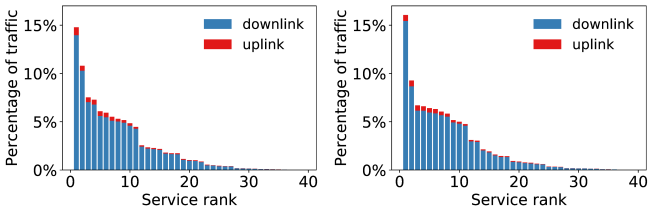

Figure 5: Percentage of the mobile traffic generated by the each service in our study. The fraction of downlink and uplink traffic is denoted by different colors. Left: large metropolis. Right: medium-sized city.

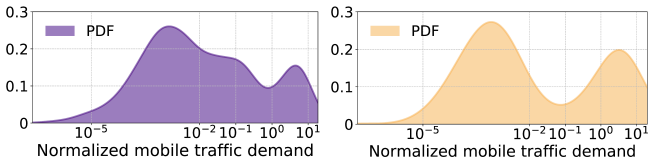

Figure 6: PDF of traffic demands across antenna sectors. Left: large metropolis. Right: medium-sized city.

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 the creation of a dedicated network slice; and *(ii)* they entail clearly distinguishable KPIs and QoS requirements. We identify 38 services that meet the criteria above, and associate them to a different network slice each.

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). We consider such service classes as representative forerunners of those expected for 5G services [4]. Figure 5 provides basic information on our selection of services. It outlines the downlink-dominated, highly skewed traffic split among the services: the percent traffic can differ of more than two orders of magnitude.

Figure 7: Antenna deployments in the target regions. Left: large metropolis. Right: medium-sized city.

A strong diversity also emerges in the way the selected services are consumed across the geographical space within the two urban regions. Figure 6 portrays the Probability Density Function (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: as our measurement data captures 2G, 3G, and 4G access, it is natural that 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.

3.2 Hierarchical network structure

The deployment of antennas in the target regions is illustrated in Figure 7, which highlights the different scales of the two case studies in terms of geographical span and density of user populations, and thus of the network infrastructures needed to support the local mobile service demands. 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 [24], 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. We assume that the operator deploys generic level-ℓ nodes and links based on two criteria: *(i)* the offered load should be similar at all nodes; and, *(ii)* the subset of antennas associated to a same node shall be geographically contiguous. The first criterion ensures basic load balancing, while the second reduces capital expenditures to connect (*e.g.*, via optics fibre) the antenna sites to the nodes. As these criteria aim at maximizing the performance of network slicing, we argue that they correspond to a plausible deployment strategy. 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 when employing dynamic resource allocation schemes.

Under these criteria, the problem of associating the level-ℓ nodes with the original antenna sites in Figure 7 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[2]. 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

$$e_{uv} = \begin{cases} 1 & \text{if } e \text{ is a cut edge} \\ 0 & \text{otherwise} \end{cases} \qquad \forall e \in E, \tag{5}$$

$$x_{v,k} = \begin{cases} 1 & \text{if } v \text{ is in partition } k \\ 0 & \text{otherwise} \end{cases} \qquad \forall v \in V, \forall k, \tag{6}$$

and formulate an Integer Linear Programming (ILP) problem:

$$\min \sum_{e_{uv} \in E} e_{uv} \tag{7}$$

$$\text{s.t.} \sum_{v \in V} x_{v,k} c(v) \leq (1+\epsilon) \frac{\sum_{v \in V} c(v)}{N_\ell}, \ \forall k \tag{8}$$

$$\sum_{v \in V} x_{v,k} c(v) \geq (1-\epsilon) \frac{\sum_{v \in V} c(v)}{N_\ell}, \ \forall k \tag{9}$$

$$\sum_{k} x_{v,k} = 1, \ \forall v \in V. \tag{10}$$

$$e_{uv} \geq x_{u,k} - x_{v,k}, \ \forall e \in E, \forall k \tag{11}$$

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

The objective function given by Equation (7) 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 (8) and (9), 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 (10) ensures that each vertex is in exactly one partition, while those given by Equations (11) and (12) 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 [29] to solve it. In doing so, we allow for a ±10% unbalance among the load served by nodes at every level ℓ, *i.e.*, $\epsilon = 0.1$ in Equations (8) and (9).

[2]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.

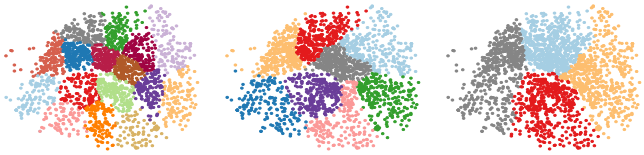

Figure 8: Association of antenna sites to level-ℓ nodes in the large metropolis scenario. The plots refer to $\ell = 8$ (16 nodes, left), $\ell = 9$ (8 nodes, middle) and $\ell = 10$ (4 nodes, right). Figure best viewed in colours.

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
$\mathcal{N}_\ell$	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 1: Hierarchical network deployments in our two urban case studies. 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 $\mathcal{N}_\ell$ serving a reference urban region at network level ℓ. At $\ell = 1$, nodes map to individual 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 (7).

Figure 8 shows three examples of antenna site partitioning among network nodes, for a selection of levels ℓ in the large metropolis scenario[3] Table 1 summarizes instead the main features of the partitions obtained in our two urban scenarios.

4 DATA-DRIVEN EVALUATION

We organise our evaluation as follows. First, we investigate worst-case settings where very stringent slice specifications are enforced, and no dynamic reconfiguration of resources is possible (Section 4.1). We then relax these constraints, and assess efficiency as slice specifications are softened (Section 4.2), or in presence of periodic resource orchestration (Section 4.3). Finally, we evaluate the impact of varied slice configurations (Section 4.4), and of a resource assignment accounting for instantaneous traffic demands (Section 4.5).

4.1 Slicing efficiency in worst-case settings

The least efficient sliced network scenario involves: *(i)* strict slice specifications, where the mobile network operator commits to guarantee the whole traffic demand ($f = 1$) averaged over short time periods ($w = 5$ minutes), for all slices; and, *(ii)* no possibility of resource reconfiguration over time, *i.e.*, τ spans the whole three-month observation time in our measurement data, and $|\mathcal{T}_\tau| = 1$. In these worst-case settings, the

[3]Note that graph partitioning is only used to outline plausible deployments where node load is reasonably balanced, yet, as we do not require a perfect balance, the specific partitioning algorithm is of no particular relevance.

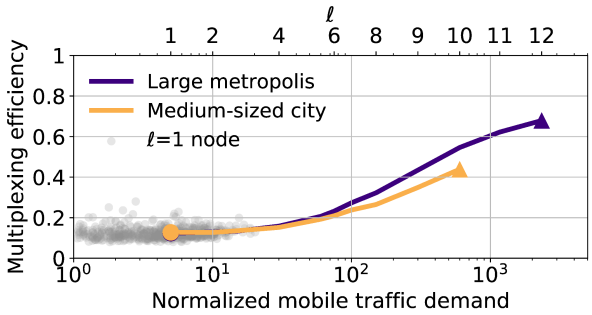

Figure 9: Efficiency of slice multiplexing versus the normalized mobile traffic served by one node (bottom x axis) at level ℓ (top x axis) in the two reference urban scenarios. Results are for a static resource assignment, *i.e.*, $|\mathcal{T}_\tau| = 1$, and slice specification $z = (f, w) = (1, 5$ minutes). 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.

operator is forced to replicate physical resources for different slices, statically allocating to each slice the resources needed to meet the associated offered load.

The multiplexing efficiency of slicing under these conditions is presented in Figure 9, which portrays it as a function of the network hierarchy level ℓ; for the sake of clarity, the latter is also mapped to the normalized mobile traffic demand observed by a level-ℓ node, as per Table 1. Each curve refers to a urban region, and confirms the intuition that the efficiency grows as one moves from the antenna level (dot on the left) to a fully centralized cloud (triangle on the right).

The underlying reason for this trend is that the traffic demands for each slice can be very bursty at individual antenna sectors; this forces the allocation of substantial resources in order to accommodate, for each slice, extemporaneous activity peaks that occur erratically in time. Aggregating demands over an increasing number of antennas results instead in growingly smoother time series. To substantiate this explanation, we look into *(i)* the timing behaviour of the different services, and *(ii)* the impact of aggregating traffic at different levels in the network, and observe the following:

(i) Different slices typically peak at different times, *e.g.*, some during work hours and others in the evening. This is exemplified by the time series in the left plot of Figure 4, and is in line with recent analyses of mobile service dynamics [18].

(ii) The burstiness of demands associated to each slice is significantly reduced as the network level grows. For instance, in the metropolis case study, the coefficients of variation of the traffic time series range in $[1.487, 2.363]$ for $\ell = 1$, in $[0.618, 0.758]$ for $\ell = 5$, and in $[0.511, 0.587]$ for $\ell = L$.

Ultimately, non-aligned and elevated traffic peaks make a static resource allocation inefficient at low network levels.

For higher values of ℓ the peak intensity is reduced, mitigating these effects and increasing multiplexing efficiency.

In addition to the general trend of efficiency with ℓ, Figure 9 allows appreciating the following quantitative results.

- The efficiency is extremely low (~0.15) at the antenna level: ensuring physical resource isolation across slices in absence of dynamic reconfiguration capabilities would require approximately 7 times the capacity of a legacy architecture where no network slicing is implemented. The grey dots in the figure highlight that such poor efficiency uniformly affects all 4G antenna sectors, independently of their offered load.
- The efficiency grows slowly when aggregating traffic at the network edge ($\ell = 2$ to $\ell = 6$). Instead, the multiplexing 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 target urban scenarios is aggregated at a single level-L node (recall that $\ell = L = 12$ in the large metropolis, and $\ell = L = 10$ in the medium-sized city, see Table 1), the efficiency remains fairly low, at 0.4–0.65. In other words, implementing the most basic form of slicing within the network core cloud (*type-A* slicing in Figure 1) would still double the amount of required resources with respect to a legacy non-sliced case.

Interestingly, differences are minimal between the two reference cities, and only emerge for high values of ℓ: we impute those to the intrinsic topological and demographic differences that characterize the two scenarios.

The results can be disaggregated for downlink and uplink traffic, as shown in Figure 10. The outcome is consistent in the two urban regions, and neatly tells apart the two directions. Downlink traffic dominates the total demand, as previously seen in Figure 5: therefore, the associated efficiency curves are very close to those in Figure 9. However, this is not the case for the uplink direction: slicing uploads tends to become remarkably (30% to 50%) less efficient as one moves towards more centralized network levels. We argue that the reason lies again 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 behaviour, since it currently dominates the aggregate data volumes, as per Figure 5. However, specific applications, hence slices, heavily rely on uplink traffic: for

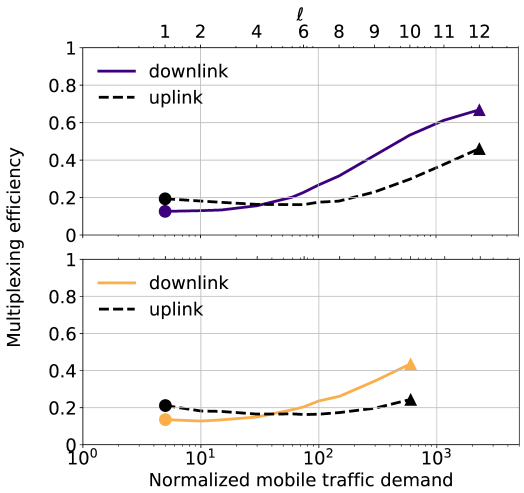

Figure 10: Efficiency of slice multiplexing, in the same settings of Figure 9, separating downlink and uplink. Top: large metropolis. Bottom: medium-sized city.

instance, the fact that efficiency at the antenna level is also low in uplink means that services that pose strong requirements on access network latency (*e.g.*, mobile gaming) are as hard to accommodate as the bandwidth-eager ones in downlink (*e.g.*, video streaming). As another example, baseband processing at a virtualized radio access is remarkably more CPU-intensive for uplink traffic [8]: the very low efficiency recorded in uplink at the network edge can make the resources assignment problem very challenging when dealing with *type-C*, *type-D* or *type-E* slices in Figure 1.

4.2 Moderating slice specifications

The poor efficiency found above is also caused by the very severe slice specifications we considered. To gain insight on this, we investigate the impact of the QoS requirements for each slice on the opportunities for multiplexing slice demands, still under a static allocation of resources.

We first relax the stringent requirement considered before in the fraction f of time during which the traffic demand for a slice must be guaranteed by the operator. The left plots in Figure 11 show how reducing f from 1 to 0.9 affects the efficiency of slice multiplexing, at different network levels ℓ and in the two reference scenarios. Decreasing f drastically improves the efficiency; for instance, by reducing the guaranteed time percentage from 100% to slightly lower values, such as 99.5%, we can nearly double the efficiency. On the downside, there exists a diminishing returns effect as f is lowered. Even allowing an overindulgent 90% guaranteed time percentage cannot bring efficiency above 0.8 for $\ell = 1$: the operator shall still increase its radio access capacity by 20% in order to isolate slices. These observations hold for all network levels ℓ and in both urban regions.

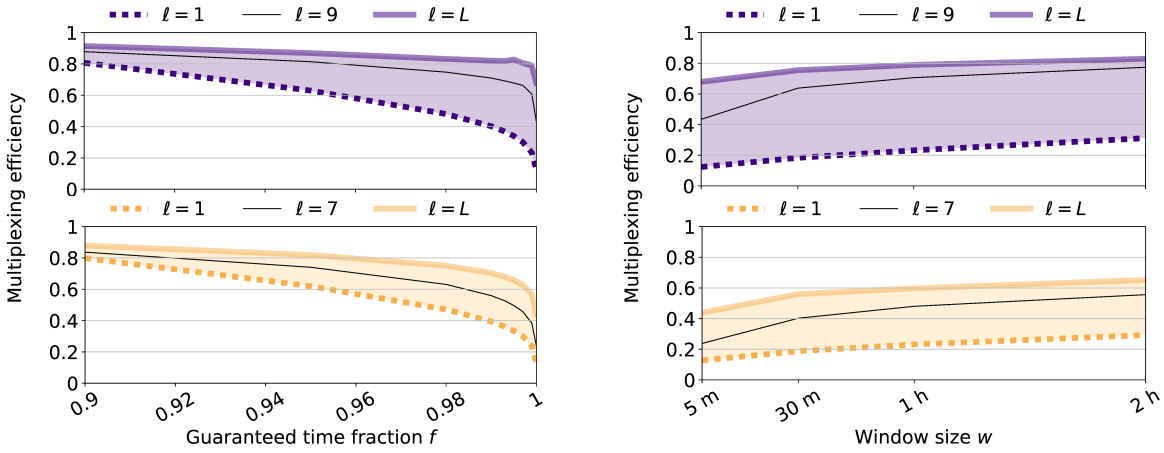

Figure 11: Efficiency of slice multiplexing versus slice specifications. Left: guaranteed time fraction f. Right: averaging window length w. Dashed and solid coloured lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city.

The other parameter governing our slice specifications is the time window length w over which traffic is averaged. We find that w has a less significant impact on efficiency than f. The exact figures are in the right plots of Figure 11: the gain is mild even for long 2-hour windows, and tuning w cannot reduce the large gap between the efficiency at the antenna level and in the network core cloud. Thus, a 3-fold capacity increase would be needed to implement slicing at physical level, even if w were set to tolerant order-of-hour values.

A final relevant aspect is that, with the proposed slice specification, it is possible that the slice demands are not satisfied over periods involving more than one consecutive time window. By appropriately setting the window size and the f parameter, we have some control over the duration of such periods. For instance, for the medium-sized city scenario and a window size of $w = 5$ m, the length of a period not fully meeting the demand is (on average) around 2 windows for $f = 0.99$, 2.5 windows for $f = 0.95$ and 3 windows for $f = 0.9$. Similar trends are observed for the large city and other window sizes. This shows that the efficiency gains resulting from decreasing f do not only involve a price in terms of the total time not satisfying the demand, but also in terms of the duration of the corresponding periods.

4.3 Orchestrating resources dynamically

We now relax the constraint on the fully static allocation of resources, and consider a network where resources can be dynamically re-allocated to VNFs over time. Such a system allows the operator to re-assign the amount of resources dedicated to each slice, adapting them to the actual time-varying demand for the services associated to the slice.

As discussed in Section 2.3, we consider that the operator can reconfigure the resources with a fixed periodicity τ

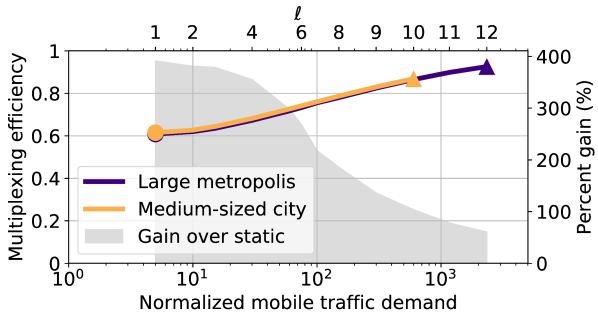

Figure 12: Efficiency of slice multiplexing (left y axis) and percent gain over static assignment (right y axis) versus the normalized mobile traffic served by one node (bottom x axis) at level ℓ (top x axis) in the two reference urban scenarios. Results are for a dynamic resource assignment where re-configurations occur with periodicity $\tau = 30$ minutes, under slice specification $z = (f, w) = (1, 5$ minutes). Dots denote $\ell = 1$ and triangles $\ell = L$ for each scenario.

which depends on the capabilities of the underlying virtualization technology. In our scenario, the operator allocates resources optimally with respect to the target slice specifications, for each reconfiguration interval of duration τ. This is equivalent to assuming availability of an oracle algorithm that, at the beginning of a reconfiguration interval, has perfect knowledge of the future time series of the demand for each service and for the rest of the interval. Then, exact information about the following timespan τ allows for an optimal matching of minimum resources to requirements, as detailed in Section 2.3 and exemplified in Figure 3.

Our baseline result, in Figure 12, refers to the case of $\tau = 30$ minutes. Note that this can be regarded as a fairly high

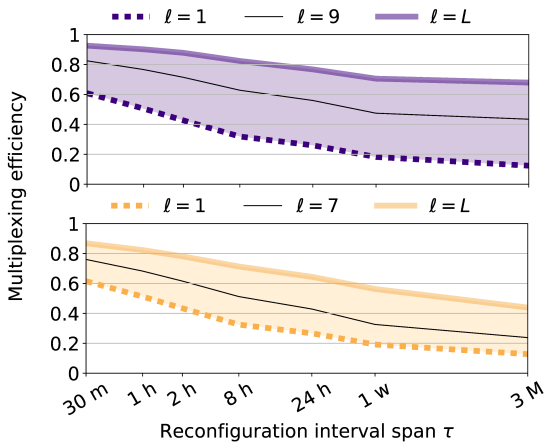

Figure 13: Efficiency of slice multiplexing versus the resource reconfiguration periodicity τ. Dashed and solid coloured lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city.

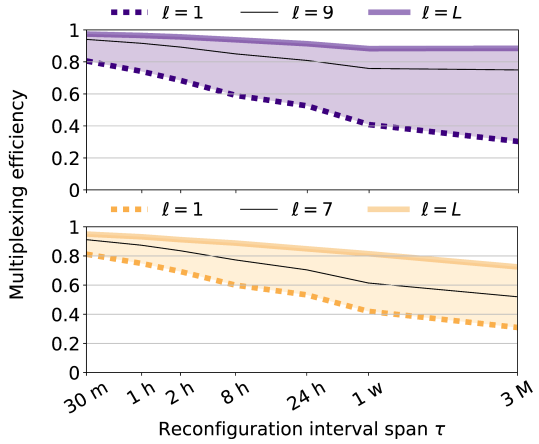

Figure 14: Efficiency of slice multiplexing in presence of 7 slices dedicated to specific service categories. Dashed and solid colored lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city.

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 [32]. 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 [30].

We can see from the results that 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 12 with their equivalent in Figure 9, the gain is evident. We made the benefit explicit as the grey region in Figure 12: it ranges between 60% and 400%, depending on the network level ℓ considered. We further observe that there is a very important difference between efficiency at the radio access and in the network core. A high-frequency dynamic orchestration of resources allows for near-perfect slice multiplexing at a cloud datacenter that fully centralizes the traffic in our large metropolis scenario. In contrast, efficiency is stuck at 0.6 (despite a much higher percent gain) 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 almost double the capacity to isolate slices at network hierarchy levels close to the radio access.

A more comprehensive picture is provided by Figure 13, which encompasses a wide set of reconfiguration intervals τ, from the 30 minutes case we just analysed in detail up to 3 months (*i.e.*, the entire timespan of the dataset, which maps to the static resource configuration case considered in Section 4.1). 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 most remarkable for low values of τ: reducing the frequency of reallocation from once every 30 minutes to once every 2 hours yields a high loss of efficiency (close to 0.2) comparable to that incurred, *e.g.*, by increasing τ from 2 to 8 hours. If we further constrain the frequency of resource reallocation to once per week or once every three months, the additional erosion of efficiency is much lower. The takeaway message is that either the operator is able to deploy virtualization technologies that allow for fast reconfiguration (in the order of a few hours at most), or it is probably not worth considering dynamic resource allocation at all.

4.4 Varying slice configurations

The mapping of services into specific network slice instances may be based on several factors, such as the requirements of the services or the specific policies implemented by each operator [3]. The number of slices and the resulting volume of traffic in each slice will have an impact on the overall multiplexing efficiency, which we investigate next.

We first study a slice configuration where the services of a similar type are aggregated together into the same slice, which allows to reduce the 38 slices that we had in the previous experiments down to 7 slices dedicated to streaming, social network, web, cloud, gaming, messaging and miscellaneous services, respectively. Figure 14 illustrates the multiplexing efficiency achieved by such a slice configuration as a function of the reconfiguration period τ. The values are substantially larger than those obtained with a larger number of

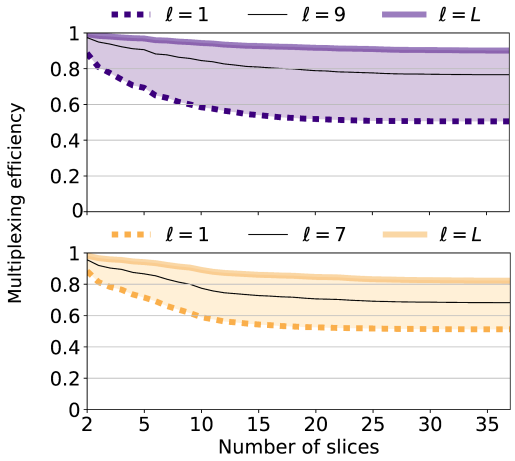

Figure 15: Efficiency of slice multiplexing as a function of the number of slices x, when the $x-1$ services with the highest traffic load have a dedicated slice and the remaining services are aggregated into a common slice. Dashed and solid colored lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city.

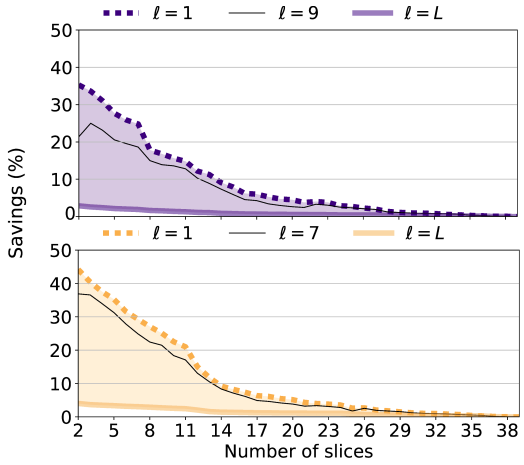

Figure 16: Savings obtained by relaxing the service guarantees of the common slice, corresponding to the difference between the resources required when $f = 1$ for the common slice, and those required when $f = 0.9$ for that slice. Dashed and solid colored lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city.

slices (see Figure 13): by aggregating service traffic we have significant gains in efficiency, yet we lose the ability to provide customized functions to each specific mobile service. It is worth highlighting, however, that multiplexing efficiency remains rather low for small l and large τ values.

A second sensible slice configuration assumes that the providers of the services that generate the highest traffic load acquire a dedicated slice tailored to their service, while the remaining services are aggregated into a common, non-customized, slice. In Figure 15, we analyze the multiplexing efficiency resulting from this configuration as a function of the total number of slices in the network (including the dedicated slices and the common one) when the reconfiguration period τ is of 1 hour and $f = 1$ for all slices. Results show that the trend becomes almost flat after 15 slices, which implies that efficiency is only improved when the services with the largest demands are brought into the common slice.

In the above slice configuration, it may be reasonable to expect that those tenants acquiring dedicated slices are provided a stricter guarantees than the ones in the common slice. In order to evaluate the benefits resulting from such a strategy, Figure 16 illustrates the resource savings resulting from providing the common slice with a guaranteed time fraction $f = 0.9$, computed as the relative percentage of resources spared with respect to those required in the configuration where all slices have $f = 1$. Results show that savings remain very low in the network core (when $\ell \sim L$), but can be significant for resources located close to the radio access (when $\ell \sim 1$). In the latter case, savings are important (up to 20-40%) when the top-10 services are included in the non-customized, low-QoS common slice. Indeed, as these account for 65% of the overall traffic (see Figure 5), they have a much higher incidence on the system performance.

4.5 Equipment deployment efficiency

To conclude our analysis, we look at the problem of resource multiplexing efficiency in a sliced network from a rather different perspective. Equations (2) and (3) derived in Section 2 assume that the relevant metric for the operator is the amount of resources utilized to accommodate the demand for mobile services aggregated over time. Therefore, the analysis carried out in Sections 4.1–4.4 is appropriate to evaluate operating expenses (OPEX), which increase when the available resources are used more intensively, and can be applied, *e.g.*, to electric power consumption, management overheads, or deterioration of assets with use.

However, another interesting viewpoint on efficiency is in terms of equipment to be deployed to meet the instantaneous demand. This relates to the capital expenditure (CAPEX) incurred by the mobile network operator, typically hardware and infrastructure costs. In this case, the expressions are slightly different, and capture the fact that the equipment must be dimensioned so as to match the peak demand. Formally, let $\hat{r}^z_{c,s}(n)$ be the resources needed to satisfy specifications z for slice $s \in \mathcal{S}$ at node $c \in \mathcal{C}_\ell$ during reconfiguration interval $n \in \mathcal{T}$, computed as indicated in Section 2.3. Then,

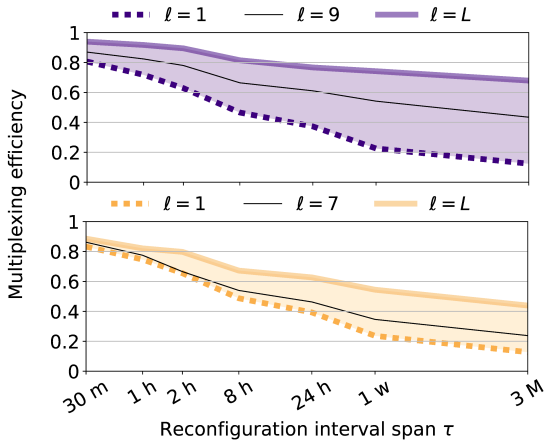

Figure 17: Efficiency of slice multiplexing from an equipment deployment perspective versus τ. Dashed and solid colored lines denote the extreme network levels $\ell = 1$ and $\ell = L$, while the black solid line follows an intermediate network level. Top: Large metropolis. Bottom: medium-sized city.

the equipment resources needed to accommodate the traffic activity peak in slice s at network level ℓ are computed as

$$\mathbb{R}^{\star z}_{\ell,\tau} = \sum_{s\in\mathcal{S}} \sum_{c\in C_\ell} \max_{n\in\mathcal{T}} \left(\hat{r}^z_{c,s}(n)\right). \tag{13}$$

Similarly, the equivalent resources needed under perfect sharing in the same settings are

$$\mathbb{P}^{\star z}_{\ell,\tau} = \sum_{c\in C_\ell} \max_{n\in\mathcal{T}} \left(\hat{r}^z_c(n)\right), \tag{14}$$

where $\hat{r}^z_c(n)$ is the amount of resources needed to accommodate the total demand aggregated over all slices in $\mathcal{S}$ at node c and reconfiguration interval n, under requirements z. The multiplexing efficiency for deployed equipment is then

$$\mathbb{E}^{\star z}_{\ell,\tau} = \mathbb{P}^{\star z}_{\ell,\tau} / \mathbb{R}^{\star z}_{\ell,\tau}. \tag{15}$$

The equipment deployment efficiency given by the above equation is shown in Figure 17. The figure summarizes results in our reference urban scenarios, under a wide range of reconfiguration time interval durations τ, and across all network architectural levels ℓ. We highlight the following aspects.

(i) In absence of mechanisms that allow for dynamic reconfiguration, the efficiency is very much comparable to that observed in the previous analysis, as shown by the values for $\tau = 3$ months in Figures 13 and 17. This is a clear indication that deploying hardware and infrastructure to provide resource isolation across slices risks to have an unbearable cost for operators if no dynamic resource reallocation is possible.

(ii) Flexibility in the orchestration of resources pays off also in terms of equipment deployment efficiency, which can be increased up to 0.8–0.95 when fast reconfiguration over 30-minute intervals is possible. These values correspond to an additional 5%–25% cost in terms of network infrastructure over the perfect sharing benchmark.

(iii) The main difference between efficiency of resource usage, given by Equation (4), and equipment deployment, given by Equation (15), is observed at architectural levels closer to radio access. When ℓ is close to 1, a dynamic reconfiguration of resources allows improving deployed infrastructure efficiency much faster than resource usage efficiency. In other words, resource isolation across slices has a sensibly lower impact on equipment installation costs than on operating expenses. For instance, at the antenna level ($\ell = 1$), efficiency is 0.6 in Figure 13 and 0.8 in Figure 17, implying that the extra cost over perfect sharing is high for resource utilization (over 60%) and much lower for equipment deployment (below 25%).

(iv) In contrast to the above, in the network core (*i.e.*, for ℓ that tends to L) trends are similar in Figure 13 and Figure 17.

Overall, our results stress how multiplexing efficiency of slice resources is largely consistent across the different perspectives entailed by the expressions of Equations (4) and (15). That is, the OPEX and CAPEX incurred by the operators to support network slicing have comparable trends with respect to the different system parameters, with the notable exception of lower deployment costs for a radio access infrastructure supporting high reconfigurability.

5 RELATED WORK

Multi-service networks [25] are the fundamental building block for the implementation of the network slicing paradigm [6] that, in turn, will enable new business models such as multi-tenancy [28] 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 [24] to specific architectures proposals [21, 35]. 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) [5, 12], the core network [27], or the management and orchestration components [19]. As a matter of fact, that research effort is already making its way into standardization: 3GPP is considering multi-service and network slicing aspects for the next Release 15, expected to deliver the first set of 5G standards [2].

On top of the architectural research work, enabling multi-service network has also been considered from an algorithmic point of view. The focal point of the research in the area has been the resource allocation in the RAN [9, 11, 16, 22] as the spectrum is the most difficult part of the network to oversubscribe. However, resource sharing in a virtualized network has also been tackled for other kinds of functions [14].

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 [7].

Our work sheds light on this overlooked aspect, by providing an empirical evaluation of slicing efficiency in large-scale scenarios, in presence of realistic multi-service demands.

6 TAKEAWAYS AND PERSPECTIVES

We analyzed, from an empirical perspective, the implications of real-world mobile service usage patterns on the network infrastructure. To the best of our knowledge, this is the first attempt at understanding the impact on resource management of network slices in a multi-service, multi-tenant network at scale. We retain a number of takeaways, listed next.

Multi-service requires more resources. Building a network that is capable of providing different services (possibly associated to several tenants) will necessarily introduce a decrease in the efficiency of the resource usage. 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 a fully centralized scenario (*i.e.*, a large datacenter 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 exhibits 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 further reduced efficiency of uplink may entail real challenges for the operators. Indeed, uplink QoS requirements are key to specific services with stringent network access needs (*e.g.*, mobile gaming). Even more so, it is likely that multiple instances of such services belonging to different tenants (*e.g.*, video-gaming platforms owned by different gaming providers) have to be served in a resource-isolated fashion in parallel.

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

Dynamic resource assignment must also 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, which may be challenging for short timescales. Underestimation of resources may lead to SLA violations, whereas over-provisioning may harm the economic feasibility of the system. Artificial intelligence and machine learning are promising techniques to accomplish this [10, 34] and are being brought into the network management landscape by standards [10].

Aggregating services is beneficial. Aggregating similar services into the same slice increases the system efficiency significantly, yet this comes at the price of losing the ability to provide a customized treatment to each service. In contrast, if the services with the highest traffic load acquire their own slice and the remaining ones are aggregated into a common slice, the resulting gains are limited unless the common slice includes services with significant load.

Deployment is slightly more efficient than operation. We analyzed the sharing efficiency from both a continuous resource usage and an infrastructure deployment perspective. While they have similar trends in the network core, the efficiency at the radio access is higher for installed hardware in presence of high-frequency resource reallocation.

Urban topography has limited impact. The fact that our results are very consistent in two urban areas of 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 may be extended to other regions and countries in (and possibly beyond) Europe.

There is room for improvement. As a final remark, we would like to stress that ours does not pretend to be a 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 a unexplored subject with strong implications for the future generations of mobile networks.

ACKNOWLEDGMENTS

We would like to thank the shepherd and reviewers for their valuable comments and feedback. 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 was supported by the H2020 5G-Transformer project (grant agreement no. 761536).

REFERENCES

[1] 3rd Generation Partnership Project (3GPP). 2015. Cellular system support for ultra-low complexity and low throughput Internet of Things (CIoT). 3GPP Technical Report (TR) 45.820.

[2] 3rd Generation Partnership Project (3GPP). 2018. NR and NG-RAN Overall Description, Stage-2 (Release 15). 3GPP Technical Specification (TS) 38.300.

[3] 3rd Generation Partnership Project (3GPP). 2018. Telecommunication management; Study on management and orchestration of network slicing for next generation network (Release 15). 3GPP Technical Report (TR) 28.801.

[4] 5th Generation Public Private Partnership (5G-PPP). 2017. View on 5G Architecture (version 2.0). 5G-PPP Architecture Working Group White Paper.

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

[6] Next Generation Mobile Networks (NGMN) Alliance. 2015. Description of network slicing concept. NGMN White Paper.

[7] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K Samdanis, and X. Costa-Perez. 2017. Optimising 5G infrastructure markets: The business of network slicing. In *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*. Atlanta, GA.

[8] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo. 2012. CloudIQ: a framework for processing base stations in a data center. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (ACM MobiCom 2012)*. Istanbul, Turkey.

[9] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez. 2017. Network slicing games: Enabling customization in multi-tenant networks. In *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*. Atlanta, GA.

[10] European Telecommunications Standards Institute (ETSI). 2017. Improved operator experience through Experiential Networked Intelligence (ENI) Introduction - Benefits - Enablers - Challenges - Call for Action. ETSI White Paper No. 22.

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

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

[13] Google. [n. d.]. Google Project Fi. https://fi.google.com/about/

[14] J. G. Herrera and J. F. Botero. 2016. Resource Allocation in NFV: A Comprehensive Survey. *IEEE Transactions on Network and Service Management* 13, 3 (Sept. 2016), 518–532.

[15] A. Ksentini and N. Nikaein. 2017. Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction. *IEEE Communications Magazine* 55, 6 (June 2017), 102–108.

[16] Y. L. Lee, J. Loo, T. C. Chuah, and L. C. Wang. 2018. Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks. *IEEE Transactions on Wireless Communications* 17, 4 (April 2018), 2146–2161.

[17] X. Li, D. Li, J. Wan, A. V. Vasilakos, C. Lai, and S. Wang. 2017. A review of industrial wireless networks in the context of Industry 4.0. *Wireless Networks* 23, 1 (Jan. 2017), 23–41.

[18] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, and Z. Smoreda. 2017. Not All Apps Are Created Equal: Analysis of Spatiotemporal Heterogeneity in Nationwide Mobile Service Usage. In *Proceedings of the 13th International Conference on Emerging Networking EXperiments and Technologies (ACM CoNEXT 2017)*. Incheon/Seoul, South Korea.

[19] A. Mayoral, R. Vilalta, R. Casellas, R. Martinez, and R. Munoz. 2016. Multi-tenant 5G Network Slicing Architecture with Dynamic Deployment of Virtualized Tenant Management and Orchestration (MANO) Instances. In *Proceedings of the 42nd European Conference and Exhibition on Optical Communication (ECOC 2016)*. Dusseldorf, Germany.

[20] T. L. Nguyen and A. Lebre. 2017. Virtual Machine Boot Time Model. In *Proceedings of the 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2017)*. St. Petersburg, Russia.

[21] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis. 2015. Network Store: Exploring Slicing in Future 5G Networks. In *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture (ACM MobiArch 2015)*. Paris, France.

[22] B. Niu, Y. Zhou, H. Shah-Mansouri, and V. W. S. Wong. 2016. A Dynamic Resource Sharing Mechanism for Cloud Radio Access Networks. *IEEE Transactions on Wireless Communications* 15, 12 (Dec. 2016), 8325–8338.

[23] M. Odini. 2016. OpenSource MANO. IEEE Softwarization: A Collection of Short Technical Articles. https://sdn.ieee.org/newsletter/july-2016/opensource-mano

[24] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi. 2016. Mobile network architecture evolution toward 5G. *IEEE Communications Magazine* 54, 5 (May 2016), 84–91.

[25] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker. 2017. Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks. *IEEE Communications Magazine* 55, 5 (May 2017), 72–79.

[26] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti. 2017. On Radio Access Network Slicing from a Radio Resource Management Perspective. *IEEE Wireless Communications* 24, 5 (Oct. 2017), 166–174.

[27] M. R. Sama, X. An, Q. Wei, and S. Beker. 2016. Reshaping the mobile core network via function decomposition and network slicing for the 5G Era. In *Proceedings of the 2016 IEEE Wireless Communications and Networking Conference (IEEE WCNC 2016)*. Doha, Qatar.

[28] K. Samdanis, X. Costa-Perez, and V. Sciancalepore. 2016. From network sharing to multi-tenancy: The 5G network slice broker. *IEEE Communications Magazine* 54, 7 (July 2016), 32–39.

[29] P. Sanders and C. Schulz. 2013. Think Locally, Act Globally: Highly Balanced Graph Partitioning. In *Proceedings of the International Symposium Experimental Algorithms (SEA 2013)*. Rome, Italy.

[30] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs. 2017. Mobile traffic forecasting for maximizing 5G network slicing resource utilization. In *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM 2017)*. Atlanta, GA.

[31] S. K. Sharma, T. E. Bogale, L. B. Le, S. Chatzinotas, X. Wang, and B. Ottersten. 2018. Dynamic Spectrum Sharing in 5G Wireless Networks With Full-Duplex Technology: Recent Advances and Research Challenges. *IEEE Communications Surveys & Tutorials* 20, 1 (Feb. 2018), 674–707.

[32] F. Z. Yousaf and T. Taleb. 2016. Fine-grained resource-aware virtual network function management for 5G carrier cloud. *IEEE Network* 30, 2 (March 2016), 110–115.

[33] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel. 2011. Multi-QoS-Aware Fair Scheduling for LTE. In *Proceedings of the IEEE 73rd Vehicular Technology Conference (IEEE VTC 2011 Spring)*. Budapest, Hungary.

[34] C. Zhang, P. Patras, and H. Haddadi. 2018. Deep Learning in Mobile and Wireless Networking: A Survey. (March 2018). arXiv:1803.04311 [cs.NI]

[35] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung. 2017. Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges. *IEEE Communications Magazine* 55, 8 (Aug. 2017), 138–145.