Malandrino, F., Chiasserini, C. F., Landi, G. (2019). Service Shifting: a Paradigm for Service Resilience in 5G. *IEEE Communications Magazine,*57(9), pp. 120-125.

DOI: https://doi.org/10.1109/MCOM.2019.1800986

Service Shifting: a Paradigm for Service Resilience in 5G

Francesco Malandrino*†‡, Carla Fabiana Chiasserini†*‡, Giada Landi§
*CNR-IEIIT, Italy; †Politecnico di Torino, Italy; ‡CNIT, Italy; §NextWorks s.r.l., Italy
francesco.malandrino@ieiit.cnr.it, chiasserini@polito.it, g.landi@nextworks.it

***Abstract*—Many real-world services can be provided through multiple virtual network function (VNF) graphs, corresponding, e.g., to high- and low-complexity variants of the service itself. Based on this observation, we extend the concept of service scaling in network orchestration to *service shifting*, i.e., upgrading or downgrading the VNF graph to use among those implementing the same service. Service shifting can serve multiple goals, from reducing operational costs to reacting to infrastructure problems. Furthermore, it enhances the flexibility of service-level agreements between network operators and third party content providers ("verticals"). In this paper, we introduce and describe the service shifting concept, its benefits, and the associated challenges, with special reference to how service shifting can be integrated within real-world 5G architectures and implementations. We conclude that existing network orchestration frameworks can be easily extended to support service shifting, and its adoption has the potential to make 5G network slices easier for the operators to manage under high-load conditions, while still meeting the verticals' requirements.**

I. Introduction

5G networks are built *for services*, not merely for connectivity. Third-party providers, called *verticals* (e.g., automotive industries, e-health companies, and media content providers), will purchase from mobile operators the networking and processing capabilities necessary to provide their services. Such services will concurrently run on the mobile operator's infrastructure, which will support their diverse requirements under the so-called *network slicing* paradigm [1], [2].

Additionally, services that have especially tight latency requirements and/or need to process extremely large amounts of traffic can leverage the so-called multi-access edge computing (MEC) paradigm. Under the MEC paradigm, computation entities (e.g., servers) are placed at the edge of the network, thus complementing Internet-based datacenters and reducing network congestion and the associated latency. By doing so, not only does MEC improve the performance of existing services, but also it enables entirely new services, including [3] virtual and augmented reality. On the negative side, MEC servers have a limited computational and memory capabilities, which shall be shared among all deployed services.

According to the network function virtualization (NFV) technology, services are specified by verticals [1], [2], [4] as a set of virtual network functions (VNFs) connected to form a VNF graph, along with the needed target Key Performance Indicators (KPIs), e.g., maximum delay or minimum reliability. Operators will host the VNFs on their own infrastructure, ensuring that they are assigned enough resources for the service to meet the target KPIs while keeping operator costs as low as possible. Such a problem is known as network orchestration [5] or VNF placement [6], and has been widely researched in the literature. Popular approaches and tools include queuing theory [6], game theory [7], and graph theory [8].

It is a natural and often unspoken assumption that every vertical service is associated with *one* VNF graph: either the service can be provided through the specified VNFs with the target KPIs, or the service deployment fails. In some cases, resource shortages are managed by limiting the damage, e.g., getting as close as possible to the target KPIs [6] or enforcing different priorities among services; however, it is typically assumed that VNFs composing a service requested by a vertical are not changed.

On the contrary, in many real-world cases, such as those discussed in Sec. II, the same vertical service can be provided through a full-fledged, *primary* VNF graph, and also in a suboptimal yet useful fashion through a different, *secondary* graph. The mobile operator can thus perform two additional operations when matching the services to provide with the available resources: it can *shift down* a certain service, dropping its primary VNF graph and deploying the secondary one in case of resource shortage, or *shift up* that service performing the opposite operation.

It is important to point out how service shifting is profoundly different from the familiar experience of trying to use a service, e.g., a video call, and then, if

the bandwidth is insufficient, switching to a similar one, e.g., an ordinary voice call. The fundamental difference is that shifting happens within *the same* service, which in turn implies that:

- shifting is initiated and performed by the network, and is seamless for the user;
- the vertical is aware of shifting decisions, and can take care of the associated non-technical aspects (e.g., discounts at billing time) with no action on the user's part.

Service shifting is also deeply different from service scaling: service scaling aims at finding enough resources to run the current VNF graph, by means of assigning more resources to currently-active servers (scale-up) or finding new servers to use (scale-out). On the other hand, service shifting is about choosing the most appropriate VNF graph to use in order to provide a given service, also considering the resources available.

In this paper, we discuss the role the service shifting operation in 5G networks, as well as the opportunities and challenges it brings. Specifically, Sec. II discusses the relevance of shifting operations, presenting several examples of services that can benefit from them. Sec. III deals with the role of service shifting decisions in a comprehensive network orchestration strategy, and Sec. IV describes how it can be implemented in practice, taking a real-world 5G architecture as a reference. Finally, Sec. V summarizes the results of our performance evaluation, carried out through a small, yet representative, reference scenario, and Sec. VI concludes the paper.

II. Shifting services

Shifting mostly benefits the services that leverage the MEC paradigm, i.e., services with (i) strict latency requirements, or (ii) very significant amounts of data to process as locally as possible. Different VNF graphs represent the fact that the same goal can be pursued through different strategies, associated with different resource requirements.

A good example is the sensor monitoring service depicted in Fig. 1(a), presenting a power grid monitoring service [9, Sec. 3.4]: in ordinary conditions, sensor readings are checked against static thresholds and used for prediction. An alarm is generated if current values exceeded the static threshold, or the predicted values are detected as anomalous. However, if a resource shortage prevents the primary VNF graph from being deployed, there is a benefit in *at least* being able to raise an alarm if thresholds are exceeded, by implementing the bottom VNF graph in Fig. 1(a). Implementing such a *secondary* graph is preferable, for both the vertical and the mobile operator, to not implementing the service at all.

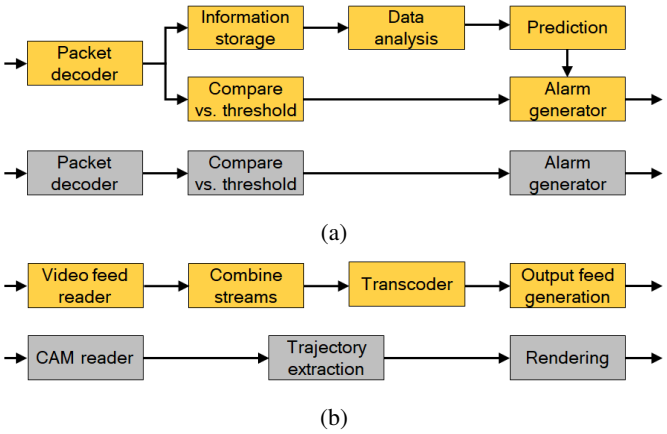


Fig. 1. Primary (gold background) and secondary (silver background) VNF graphs associated with a grid monitoring (a) and a bird's eye view (b) service. Note that some VNFs may be common to both graphs, as in (a).

Another relevant example is the *Bird's eye view* service [9, Sec. 3.1.4], which provides drivers (and autonomous vehicles) with a stream of real-time information about the current road conditions, including approaching vehicles/pedestrians. As depicted in Fig. 1(b), such information can be obtained from cameras mounted on vehicles or along the roads (top graph). If resources are insufficient, cooperative awareness messages (CAMs) can instead be leveraged to construct a schematic view of the positions of the nearby vehicles (bottom graph).

Note that the secondary graph is either a subset of the primary one, or it includes a (smaller) number of VNFs, each of which characterized by lower requirements. It follows that deploying the secondary graph of a certain service *in lieu* of the primary one will always lead to shorter delays and a reduced resource consumption, at the price of a lesser quality of experience for the user.

III. Applications and decision-making approaches

Here we describe two of the main applications of service shifting, namely, reacting to resource shortage situations (Sec. III-A) and extending the expressiveness of Service Level Agreements (SLAs) (Sec. III-B). For each application, we discuss the decision-making entities that are involved and the approaches they can take.

A. Reaction to resource shortage

As mentioned earlier, in 5G networks operator-owned resources (e.g., servers) are used to run vertical-specified services, i.e., the VNFs composing their VNF graph. A *resource shortage* situation happens when the quantity of

available resources drops unexpectedly, or the traffic load grows suddenly. This can be caused by several different conditions, including:

- problems in the operator infrastructure, e.g., servers breaking down or data centers becoming inaccessible due to link failures;
- sudden increases in traffic, including mass events ("flash crowds");
- emergency situations and natural disasters, whereby parts the network infrastructure can be destroyed and network demand, by both victims and responders, increases.

In resource shortage conditions, the operator is unable to meet all target KPIs for all services. The traditional approach is to *re-orchestrate* [10] the affected services, which include (i) moving VNFs from unavailable servers to operating ones, and (ii) scaling down the resources they are assigned. This unavoidably results in KPI targets being violated, which, in turn, may jeopardize the usefulness of the service itself, e.g., lagging video for the see-through service discussed in Sec. II.

In this context, service shifting represents a very attractive alternative to scaling down. Instead of trying to implement the primary VNF graph of a service while missing the associated KPI targets, the operator can shift down that service and provide it through its secondary VNF graph. As for choosing *which* services to shift down, the operator can follow several approaches, including:

- revenue maximization: down-shifted services bring a reduced revenue, hence, shift down the services associated with the lowest revenue loss;
- minimization of the user QoE degradation: down-shifted services result in a lower user satisfaction as the quality of experience users perceive may be severely impacted, hence shift down the less popular services;
- minimization of the service reaction time: re-orchestration, e.g., instantiating new VNF instances and updating routing tables, takes a non-negligible time, hence, shift down the services requiring the fewest such operations.

B. Extending SLAs

The possibility of service shifting can be leveraged during the SLA negotiation between verticals and operators. As an example, a vertical may accept that the secondary VNF graph is used for its service for a certain fraction of requests and/or in certain times of the day, in exchange of a reduced fee. Similarly, the semantics of service priorities can be extended to mandate that a service can be shifted down only if all lower-priority services (by the same vertical) have already been shifted down.

For operators, service shifting means extending the orchestration options: in addition to VNF placement and resource assignment [6], operators will be able to use shifting decisions to pursue their high-level objective to meet the SLA commitments while minimizing costs. For verticals, service shifting is an additional way to express their needs when negotiating SLAs, thus avoiding paying for unnecessary resources or features.

On the negative side, orchestration decisions are bound to become more complex, from several viewpoints, including the identifying the decision-making entities, provide them with the information they need, and designing swift, yet effective, algorithms for them to run. All such aspects are discussed in Sec. IV next.

IV. Service shifting in practice

We now describe how service shifting can be implemented in real-world 5G networks. Specifically, we describe which entities will be in charge of making and enacting shifting decisions (Sec. IV-A), how they will interact (Sec. IV-B), and the associated challenges (Sec. IV-C).

A. Shifting in 5G architectures

Service shifting can be viewed as an extension to traditional network orchestration, which makes orchestration decisions even more complex to handle. This further strengthens the need, recently emerged in the 5G research community, to distribute the burden of network orchestration decisions across multiple decision-making entities, working at different abstraction layers.

In the network management and orchestration (MANO) framework, standardized by ETSI in standard GS NFV MANO 001, virtually all network orchestration decisions are made by the NFV Orchestrator (NFVO). The NFVO takes as an input the service graphs and KPIs specified by verticals through the Operation and Business Support Services (OSS/BSS). Its output is represented by VNF instantiation and placement decisions, which are subsequently enacted by lower-level entities like the VNF manager (VNFM).

Several 5G-related research efforts envision alternative solutions, advocating to split the tasks assigned to the NFVO in the MANO framework between two entities: a higher-level one, making decisions on a per-service basis, and a lower-level one, working with individual VNFs with decisions more oriented to resource-based criteria. Taking the architecture proposed by the H2020

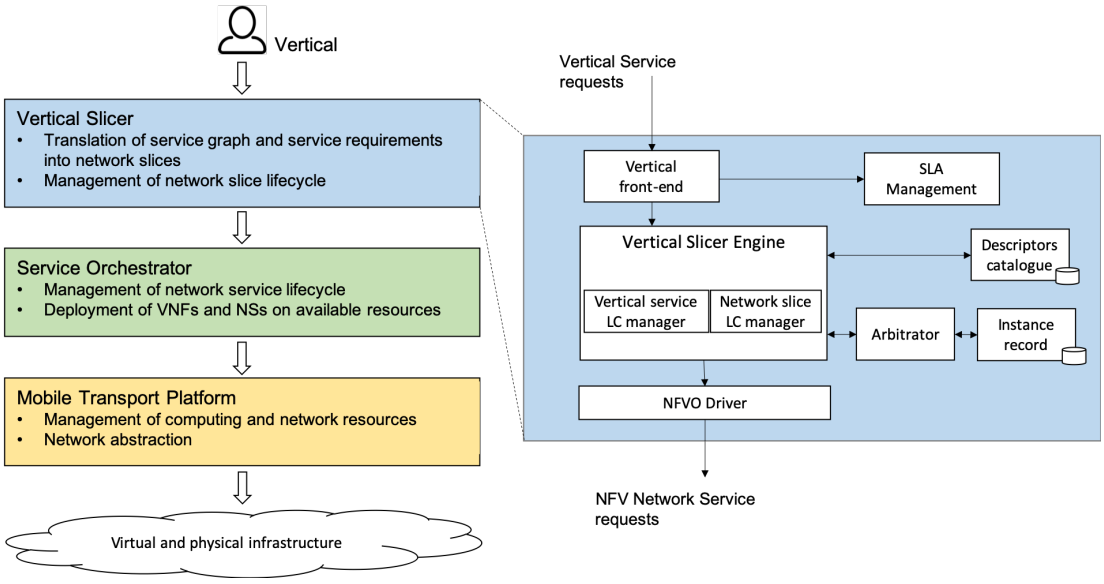


Fig. 2. The high-level architecture of the 5G-TRANSFORMER project, the interaction between decision-making entities therein, and the internal architecture of the vertical slicer.

project 5G-TRANSFORMER in [11], and represented in Fig. 2, we can identify:

- the vertical slicer (VS), translating the verticals' requirements into service graphs, also accounting for the service-level agreements (SLAs) in place;
- the service orchestrator (SO), taking the service graph as an input and using the network, computing and storage resources available in the infrastructure to build the network slice that will run the service.

In such a context, service shifting decisions can be made by higher-level, service-aware entities such as the VS. This avoids further increasing the burden on lower-level entities like the SO, which are already in charge of VNF placement and resource assignment.

B. Making and implementing the decisions

As discussed earlier, in the 5G-TRANSFORMER architecture the VS will be in charge of shifting decisions. In the following, we discuss its internal architecture, also summarized in Fig. 2, and how it will interact with other 5G-TRANSFORMER entities in order to make and implement the shifting decisions.

The internal architecture of the VS, summarized in the right part of Fig. 2, includes multiple sub-entities, including:

- the *arbitrator*, in charge of actually making the decisions;
- a SLA manager, storing information on SLA resources and tracking how they are used;
- catalogs and record managers, storing the information needed by the arbitrator;
- an engine, in charge of coordinating the work of all other VS sub-entities, as well as the life cycle (LC) of network slices;
- font-ends and drivers, implementing the interfaces between the VS and other 5G-TRANSFORMER entities, as well as with verticals.

Within such an architecture, adding service shifting capabilities to the VS would require four main actions. First, the vertical front-end shall be extended, in order to allow verticals to indicate multiple requirements (hence, multiple VNF graphs) for the same service. Furthermore, in order to store such VNF graphs, a new catalog shall be added, called *VNF graph catalog*. Additionally, the actual shifting algorithms must be implemented at the arbitrator. Finally, the NFVO driver shall be updated to convey shifting decisions from the VS to the SO.

Fig. 3 presents a simplified vision of how 5G-TRANSFORMER entities interact when making and enacting service shifting decisions. In steps 1–2, the vertical informs the VS of the requirements associated with the different versions of its services. In steps 3–8, the vertical requests to the VS the deployment of services s_1 and s_2. After checking the available SLA resources, the VS decides to deploy the primary graph of s_1 and the secondary one of s_2, and instructs the SO accordingly.

In step 9, the monitoring platform detects a resource shortage situation and informs the SO, which relays the warning to the VS. Such a situation requires to shift down a service, and the VS decides to shift s_1 from primary to secondary. The decision is then notified to the

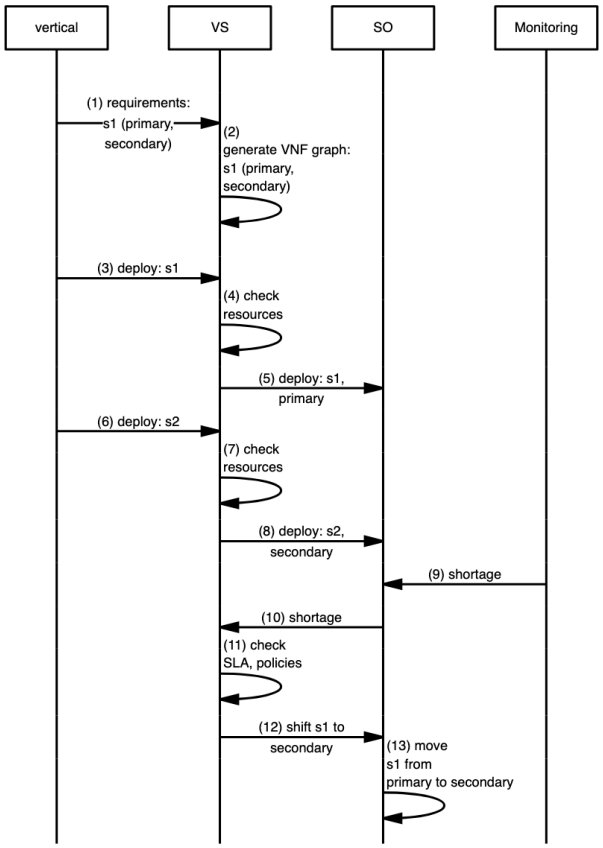


Fig. 3. Making and enacting shifting decisions: interaction between entities of the 5G network architecture.

SO, which enacts it by removing the VNFs associated with the primary graph of s_1 and deploying those of the secondary graph.

C. Challenges

There are several challenges to tackle order to make effective service shifting decisions. Among the most significant, we discuss gathering and collecting input information, timing the decisions, and managing the transition between VNF graphs.

Input information and monitoring. As recalled in Sec. IV-A, the VS and SO decision entities run algorithms that need to receive as input different kinds of monitoring data, related to a variety of physical and virtual components and resources, from physical infrastructures to virtual resources, up to application and service level data. The monitoring platform should be flexible enough to support different types of customizable data sources in a distributed environment. They should also implement preliminary data elaboration tasks to efficiently deliver aggregated monitoring parameters and produce automated notifications, based on simple thresholds or more complex strategies for anomaly detection.

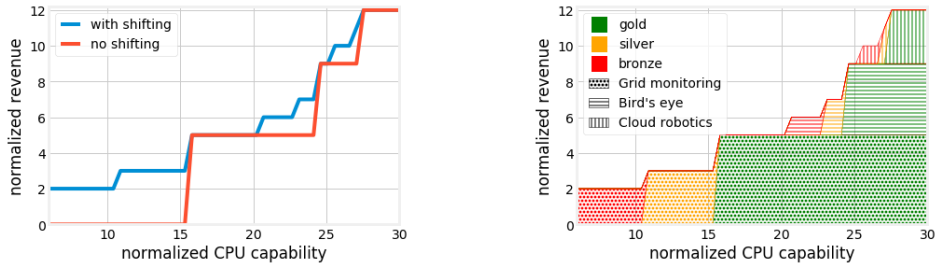
The complexity of aggregation and elaboration of the raw monitoring data, as collected by the elementary monitoring sources, is centralized at the monitoring platform. Such processing is driven by the rules that are dynamically configured according to the network service specification, in order to detect the particular conditions triggering scaling or shifting actions. Whenever a target pattern is detected in the aggregated monitoring data, automated alerts are notified to the monitoring consumers (VS or SO) that have an active subscription for the given pattern. Notifications may be managed either through explicit messages addressed to the target entities or through a message bus approach. Starting from the received alerts, the VS or the SO will make a decision about the need of a service shifting and will trigger the required actions.

Decision timing. Indeed, shifting decisions are often made in resource shortage conditions, where KPI targets are being or may be violated. Therefore, service (re)deployment decisions must be made *and enacted* swiftly. The first requirement, i.e., that decisions be made quickly, is at odds with the complexity of the decisions to make, which include placing multiple VNFs throughout the network infrastructure. The second requirement, i.e., that decisions be enacted swiftly, is often overlooked but very important: indeed, real-world 5G deployments show VNF instantiation times of several tens of seconds [12].

Moreover, a full operation service also needs applications completely up and running in the new VNFs; this requires additional time due to the starting procedures of the processes and the initial configuration of the applications running in Virtual Machines (VMs) or Containers. Live migration of, e.g., VMs also brings a certain degree of delay, which may impact the services that do not need to be shifted, but just moved to different servers. A report about live migration in OpenStack Ocata[1] shows average measurements from nearly 50 seconds up to 270 seconds for the time required to migrate "heavy" VMs, depending on the VMs' storage strategy (i.e., local vs. shared storage) and tunneling activation. Such delays can result in non-negligible service outage times, and substantial penalties for the mobile operator.

Intuitively, taking action as early as possible is a promising way out of such a conundrum. However, early actions may turn out to be unnecessary (e.g., the traffic of a certain service did not grow as much as anticipated), or even wrong. To minimize such mishaps, several *traffic prediction* [13] techniques have been developed, typically leveraging machine learning techniques to ac-

[1] http://superuser.openstack.org/wp-content/uploads/2017/06/ha-livemigrate-whitepaper.pdf



(a) (b)

Fig. 4. Small-scale example scenario: revenue obtained with and without service shifting (a) and revenue breakdown (b).

curately detect relevant trends.

Managing the transition between graphs. Shifting decisions, e.g., moving from the primary VNF graph of a service to the secondary one, require several operations on individual VNFs, e.g., deactivating those of the primary graph and activating the additional ones (if any) needed by the secondary one. The order in which such operations are performed has a significant impact on the effect of the shifting decision, and must therefore be taken into account.

One possible approach is make-before-break, i.e., first all VNFs of the secondary graph are deployed, and then those of the primary ones are removed. The main advantage of this approach is service continuity, i.e., there is no point in time at which the service is not provided. On the negative side, make-before-break means that, for a short time, both the VNF graphs will be active, and so even a shifting-down action ends up temporarily consuming more resources. This is acceptable if the action is taken early enough (e.g., thanks to effective forecast), but often infeasible in resource scarcity conditions.

The alternative approach is break-before-make, i.e., first remove the VNFs of the primary graph (that are not used by the secondary one), and then deploy those of the secondary graph (that were not already used by the primary one). This approach requires the smallest possible amount of resources, but it implies the possibility that, albeit for a limited amount of time, the service will be interrupted. Intermediate approaches, whereby deactivation and deployment operations are interleaved, are also possible: in the 5G-TRANSFORMER architecture, it is the SO's task to decide the exact sequence of operation to perform in order to implement the shifting decisions made by the VS.

V. Performance Evaluation

We now quantify the benefits of service shifting by implementing the following algorithm at the VS, based on [14], operating as follows:

1) the VS sorts the services in decreasing priority order;
2) for every service:
   a) start from the primary VNF graph;
   b) instruct the SO to deploy such a graph;
   c) if resources are insufficient, move to the next graph.

We consider a simple, yet representative, scenario, including the two services in Fig. 1 and the cloud robotics service described in [9, Sec. 2.4.1]. Grid monitoring has the highest priority, followed by bird's eye, and then by cloud robotics. As highlighted in [9], all services belong to the mission critical cluster and all require low latency and high reliability. Denoting, for simplicity, the different graphs associated to services as good, silver, and bronze, we set their the normalized requirements to (respectively) to 20, 10, and 5. Furthermore, the normalized revenues associated to the gold, silver, and bronze graphs are: $[5, 3, 2]$ for grid monitoring, $[4, 2, 1]$ for bird's eye, and $[3, 2, 1]$ for industrial robotics. The physical infrastructure is composed of six hosts (servers) connected in a two-layer topology reflecting the core network organization used in [15]. For simplicity, we focus on computational capabilities alone and vary the normalized CPU available at each host between 1 and 5.

What we seek to assess is how much network shifting, i.e., the possibility to deploy silver or bronze graphs *in lieu* of gold ones, improves the revenue. Fig. 4(a) provides a quite clear answer to our question: revenue increases as the available CPU grows, and service shifting is always associated with a higher revenue. It is even more interesting to observe, in Fig. 4(b), the services and graphs generating such a revenue, represented by patterns and colors respectively.

In the baseline case, all revenue comes from gold graphs (green areas in the plots): when the network capacity is very low, it is impossible to deploy anything; as it, the VS deploys first the gold graph of the grid monitoring service, then the gold graph of the bird's

eye service, and so on. If, on the other hand, service shifting is possible, we can see that the VS is able to deploy bronze and silver graphs of different services (yellow and read areas in the plot), even when there are not enough resources for the corresponding gold graph, thereby guaranteeing a higher revenue.

VI. Conclusion

We showed how service shifting can be beneficial in resource shortage situations, which may arise as a consequence of 5G network infrastructure issues, sudden increases in traffic demand, or emergency situations. We also identified the main challenges associated with service shifting; then, taking real-world 5G implementations as a reference, we highlighted how such challenges can be tackled without major changes to their architecture, thus making it easy to reap the benefits of service shifting. As confirmed by our performance evaluation, service shifting yields a threefold benefit: vertical requirements are satisfied in a wider range of cases, network infrastructure is better utilized, and mobile operators are able to obtain a higher revenue.

References

[1] H. Zhang, N. Liu *et al.*, "Network slicing based 5G and future mobile networks: mobility, resource management, and challenges," *IEEE Comm. Mag.*, 2017.

[2] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori *et al.*, "Network slicing to enable scalability and flexibility in 5G mobile networks," *IEEE Communicarions Magazine*, 2017.

[3] ETSI. (2017) Multi-access Edge Computing (MEC). https://www.etsi.org/technologies/multi-access-edge-computing (Accessed May 2019).

[4] L. M. Contreras and D. R. López, "A network service provider perspective on network slicing," *IEEE Softwarization*, 2017.

[5] S. Vassilaras, L. Gkatzikis, N. Liakopoulos *et al.*, "The algorithmic aspects of network slicing," *IEEE Comm. Mag.*, 2017.

[6] S. Agarwal, F. Malandrino, C. F. Chiasserini, and S. De, "VNF Placement and Resource Allocation for the Support of Vertical Services in 5G Networks," *IEEE/ACM Transactions on Networking*, 2019.

[7] M. Bagaa, T. Taleb, A. Laghrissi, A. Ksentini, and H. Flinck, "Coalitional Game for the Creation of Efficient Virtual Core Network Slices in 5G Mobile Systems," *IEEE Journal on Selected Areas in Communications*, 2018.

[8] S. Dräxler, H. Karl, and Z. Á. Mann, "Jasper: Joint optimization of scaling, placement, and routing of virtual network services," *IEEE Transactions on Network and Service Management*, 2018.

[9] 5G-TRANSFORMER project. (2017) D1.1: Report on vertical requirements and use cases. http://5g-transformer.eu/wp-content/uploads/2017/12/Report_on_vertical_requirements_and_use_cases.pdf (Accessed May 2019).

[10] D. M. Gutierrez-Estevez, M. Gramaglia, A. de Domenico, N. di Pietro, S. Khatibi, K. Shah, D. Tsolkas, P. Arnold, and P. Serrano, "The path towards resource elasticity for 5g network architecture," in *IEEE WCNC Workshops*, 2018.

[11] A. De la Oliva, X. Li, X. Costa-Perez *et al.*, "5g-transformer: Slicing and orchestrating transport networks for industry verticals," *IEEE Communications Magazine*, 2018.

[12] 5G-Crosshaul project. (2017) D5.2: Report on validation and demonstration results. http://5g-crosshaul.eu/wp-content/uploads/2018/01/5G-CROSSHAUL_D5.2.pdf (Accessed May 2019).

[13] V. Sciancalepore, K. Samdanis *et al.*, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *IEEE INFOCOM*, 2017.

[14] C. Casetti, C. F. Chiasserini *et al.*, "Arbitration among vertical services," in *IEEE PIMRC*, 2018.

[15] L. Cominardi, L. M. Contreras, C. J. Bcrnardos, and I. Berberana, "Understanding QoS Applicability in 5G Transport Networks," in *IEEE BMSB*, 2018.

Acknowledgment

This work is supported by the European Commission through the H2020 projects 5G-TRANSFORMER (Project ID 761536) and 5G-GROWTH (Project ID 856709).