Key Performance Indicators for 5G network slicing

Sławomir Kukliński*◇
Lechosław Tomaszewski*
*Orange Polska, ◇Warsaw University of Technology
Warsaw, Poland

***Abstract*—Network slicing technology will influence the way in which new networking solutions will be designed and operated. So far, network slicing is often linked with 5G networks, but this approach can be used to deploy any communications network(s) over a common infrastructure. The concept is still a subject of intensive research and standardization. From the point of view of network or service operator, it is necessary to define fundamental qualitative indicators for performance evaluation of the network slicing. Such parameters are often called Key Performance Indicators (KPIs). Network slicing KPIs should deal with network slicing run-time and life-cycle management and orchestration. The paper proposes a set of KPIs for network slicing taking into account the 5G network specifics.**

I. Introduction

Network slicing enables the creation of parallel virtual telecommunication networks (which in some cases can be integrated with applications) over a common infrastructure. The main advantage of this approach is the ability of on-demand creation of isolated networking solutions, which are combined or tailored for specific applications and give slice management capabilities to slice tenants. Most of the approaches to network slicing follow the concept described by NGMN [1], [2]. The key enabler of network slicing is the ETSI Network Functions Virtualization (NFV) framework [3].

Network slicing is still not a mature technology and is a subject of many research projects as well as standardization efforts. The research concerns multiple issues: efficient resource allocation, isolation between slices, multi-domain slicing, slice orchestration, and management as well as slice description and selection issues. A comprehensive list of network slicing-related research topics can be found in [4]. So far, the performance issues related to both the run-time and the life-cycle operations of network slicing are not addressed well.

The set of the most important performance parameters of the telco system or subsystem is typically referred to as Key Performance Indicators (KPIs). In general, the set of KPIs can be huge, but using them and keeping conformance to them is a common practice of all telco operators. They use KPIs for (i) requirements for system definition and implementation; (ii) verification of proper functioning of installed networks, sub-networks, systems or sub-systems; (iii) definition of customer's requirements and provider's obligations as a part of the Service-Level Agreement (SLA) contract; (iv) comparison of performance provided by specific vendors or technologies. If a system is composed of multiple subsystems, fulfilling the subsystems' KPIs can provide the fulfillment of the overall system's KPIs. The definition of KPIs helps in the overall system engineering (dimensioning) in order to provide the requested end-to-end KPIs often linked with the Quality of Experience (QoE). For telco operators, verification of KPIs against their goals is a basic "health" indicator of solutions in service. Some KPIs can be standardized (as long as they are technology-specific), but some can be operator-specific. In general, standardization of KPIs is highly desirable, and Standards Developing Organizations (SDO) such as ITU-T, ETSI, 3GPP or GSMA provide referential KPIs for some of their standardized solutions.

In telco networks, KPIs may deal with different network segments, layers, mechanisms, aspects and also services or activities (e.g. fault handling time). Some of KPIs can be user-oriented, whereas others can be network-oriented. In mobile networks KPIs can be related to network transport, front-haul, radio link quality, data plane efficiency, and control plane operations as handover execution time, user attachment time, etc. Typically, KPIs calculation is done by the network or service management system. However, some KPIs for the network management system itself (for example fault handling time) can be also defined. KPIs are usually calculated by the network/service monitoring system, and in some cases, the management system may have the ability to take actions in order to guarantee the requested KPIs. One of the issues related to KPIs is their efficient monitoring.

In the era of the softwarized network, there are new operations whose performance should be evaluated. For example, the orchestration, as well as virtual infrastructure performance, have to be monitored. The network slicing is perceived as a key technology that will be used by telco operators, soon. However, the KPIs concerning this technology remain still undefined. Such KPIs can deal with operations related to both, slice run-time operations as well as slice life-cycle operations. It is worth noting that the definition of KPIs is also linked with their calculation methodology and may have an impact on the overall architecture of the developed system. The following arguments are the main motivation for the paper that deals with KPIs for network slicing. Our approach is linked with 5G (and 5G+) network slicing, but it can be used also in other networking solutions.

The structure of the paper is the following: Section II consists of the description of the related work, Section III consists of the list of the KPIs proposed for network slicing, Section IV discusses the calculation of the KPIs by ETSI MANO. Finally, Section V concludes the paper.

This work has been supported by the EU-Japan project 5G!Pagoda (under grant agreement 723172) and by the EU-China project 5G-DRIVE (under grant agreement 814956).

II. Related work

The issue of network performance and resulting service quality is fundamental for telco operators. The numerous and specific technology-related performance indicators, when tracked, provide quantitative insight into the behavior of equipment, sub-systems and entire systems. The group of higher-level abstraction performance indicators, giving a representative view of the end-to-end network, forms a list of Key Performance Indicators (KPIs), which at the level of underlying communication technology (network) contribute to the end-to-end communication service-level quality view, represented by Key Quality Indicators (KQIs). The referential approach to the topic is defined by several SDOs. It is worth emphasizing that the number of KPIs, in general, should be minimized, but at the same time, the KPIs should describe the most important system or network features. The currently used KQIs, introduced by ETSI [5] and profiled by 3GPP [6], [7] on the basis of fundamental definitions and concepts of ITU-T [8], offer a framework to assess objectively 2G/3G/4G services' performance and quality from the end-to-end perspective. The practical digest for operational use in telco operators, which integrates outputs of above-mentioned SDOs, is provided by GSMA [9] and its framework is structured in four layers:

- Network Availability;
- Network Accessibility;
- Generic service layer consisting of Service Accessibility, Service Integrity, and Service Retainability;
- Specific service layer consisting of groups of parameters typical for distinct communication services (e.g. SMS, voice call, web browsing, streaming etc.).

There is a lot of standardization efforts focused on defining of criteria of 5G network performance and quality assessment. The existing approaches are typically linked only with a partial view of the overall system. This is justified by the overall complexity of the 5G network and the way in which the 5G architecture is decomposed. The approach to quality assessment is usually focused on 5G services' requirements and characteristics, as defined by ITU-R (cf. [10], [11]) and 3GPP (cf. [12], [13]). NGMN has published its recommendations for KPIs and requirements for 5G [14]. 3GPP specification [15] has defined so far the following 5G KPIs related to network slicing:

- Accessibility KPIs: registered subscribers through AMF and UDM, registration success rate per network slice instance (NSI);
- Integrity KPIs: end-to-end latency of the 5G network, upstream/downstream throughput for Network Slice Instance (NSI) and at *N3* interface, RAN-UE throughput;
- Utilization KPIs: mean number of Protocol Data Unit sessions for NSI, virtualized resource utilization for NSI.

A long list of management and orchestration of 5G network performance measurements can be found in [16]. The specification includes performance measurements of individual network functions gNB, AMF, SMF, UDM, and PCF. The 3GPP plans to include KPIs related to mobility, retainability and availability in the future releases of the specification. In the context of network slicing the specification includes performance management related to virtualized resources called common performance management for network functions. The monitoring parameters include mean virtual CPU/memory/disk utilization. The measurements follow the ETSI specification [17].

In the context of network slicing, worth mentioning are also standardization activities of ETSI NFV Industry Specifications Group, which has published several specifications dealing in general with performance monitoring within the NFV framework. Additionally, the issues related to management of MANO itself, including its fault and performance, have been defined in [18]. Very detailed list of NFVI metrics has been developed by ATIS [19]. However, both activities are focused on detailed performance monitoring for the internal MANO purpose and less linked with the KPIs philosophy. As we want to reuse the ETSI NFV specifications as much as possible, we will discuss some details of the NFV specifications in the context of KPI calculations later on.

There are also research projects that are focused on 5G network KPIs. The 5GENESIS project [20], which is focused on 5G trials, plans to assess KPIs concerning: services (e.g. data rates, reliability, and latency), applications (user-perceived Quality of Experience – QoE and security), network (coverage and density) and network life-cycle management. Performance metrics such as slice establishment time, VNF relocation and instantiating times, and the computational resource usage of the protocol stack will be analyzed. The ONE5G project studies RAN KPIs in the context of optimization of the end-to-end 5G performance [21] and follows the 3GPP recommendations [12], [13]. The 5G-MoNArch project is aimed at implementation architecture for 5G, including the issues of underlying network slicing, cross-domain management and run-time optimization. Hence, the project has provided a broad set of KPIs [22] grouped into:

- General KPIs: typical, 5G service-oriented requirements;
- Resilience and Security KPIs: end-to-end reliability, telco cloud reliability, service restoration time, security threats identification, security failure isolation;
- Resource Elasticity KPIs: availability, cost efficiency gain, elasticity orchestration overhead, minimum footprint, multiplexing gain, performance degradation function, rescuability, resource consumption, resource savings, response time, resource utilization efficiency, service creation time, time for reallocation of a device to another slice;
- Application-specific KPIs: frame rate judder, the maximum number of simultaneously active IoT devices, task success rate, time on task, use of search vs navigation, etc.

In the context of use case-oriented KPIs, the 5GCHAMPION, 5GCAR, and TRIANGLE projects are worth a mention. The 5GChampion has been focused on 5G use cases used dur-

ing the 2018 Winter Olympics in Korea [23]. The 5GCAR [24] has defined automotive (service) KPIs and include communication network (radio) KPIs also. The TRIANGLE project has defined and implemented a framework to test and benchmark 5G applications, devices, and services [25]. Additionally, the 5G PPP view on communication service KPIs can be found in [26]. All the mentioned projects and initiatives follow the ITU-R or 3GPP visions of requirements definitions of the end-to-end communication service.

So far, the majority of approaches to 5G are unaware of the performance and quality of network slicing. There are some, above-mentioned SDOs and research projects efforts related to network slicing, but the issue has not been solved yet. This has motivated us to formulate a small but representative set of KPIs that can be used in order to assess the impact of implementation of certain networking solution (e.g. EPC) as a network slice. It should be noted that on the one hand, the virtual implementation of the network solution brings some benefits, but on the other one, it creates some problems. Having in mind that the number of slices can be huge, the number of network slicing related parameters has to be keep to the minimum in order to minimize the overhead related to their collection, calculation, and interpretation. The proposed KPIs for network slicing are described in the next section.

III. Network slicing KPIs set proposal

In this section, we focus only on these KPIs that are related to network slicing. The KPIs related to the solution, which is implemented as a network slice, are out of the scope of our analysis because they should be exactly the same as defined for the non-sliced implementation of the solution. For example, if 4G EPC is implemented as a slice, KPIs that concern EPC implementation in a non-sliced (or virtualized) environment are applicable. So, we focus only on the parameters that are defined by 3GPP [16] as common for different network functions.

To define the network slicing KPIs, it is necessary to use a certain functional model of network slicing as well as its implementation. We follow the NGMN functional approach [2] with some extensions and ETSI MANO approach for slice orchestration. We will separately analyze KPIs for slice run-time and life-cycle operations. Moreover, we provide a separate analysis for RAN slicing as the technology virtualization differs from other domains. The proposed KPIs not only deal with a single slice but we also took into account the impact of the multiple slices on infrastructure and single slice operations. We also analyze KPIs for multi-domain slices. In the case of multiple domains, we follow the approach with local (per domain) orchestrators and stitching of domain slices for obtaining the end-to-end slice. Such an approach requires a master orchestrator on top of local orchestrators and well-known domain-slices. In comparison to the multi-domain orchestration that uses a single slice blueprint, single orchestrator and the capability of allocation of resources of all domains, such orchestration is faster. Moreover, local orchestrators can handle domain-specific issues that are of premium importance, e.g. for RAN.

In our concept, we partly use and update the KPIs defined by 3GPP [15] and performance measurements specified by ETSI [17]. In comparison to the existing approaches, we provide a set of KPIs which are calculated simpler (only thresholds exceeds are reported) but provide the same sufficient information about the behavior of the network slicing. We assume that in case of detected issues the management and orchestration system will trigger some action that will include more detailed monitoring and solving the issue. The proposed KPIs deal by definition with some performance-related indicators that typically change dynamically, therefore we did not address slice parameters that are static. The proposed KPIs can be split into slice run-time and slice life-cycle management related.

A. Slice run-time KPIs

Slice run-time KPIs concern performance of a network or a service that is implemented as a slice and typically are identical as in case of non-sliced implementation of the network or solution. The only new mechanisms that are slice-agnostic in the virtualized implementation are related to the usage of virtual resources by a slice and orchestration operations related. One of the key operations on resources is resource scaling according to their usage. We focus on three types of virtual resources, namely connectivity, computing, and memory. In ETSI NFV framework, memory (i.e. RAM and swap space) and disc measurements are performed separately [17] – we introduce a single, synthetic parameter related to usage of all kinds of memory.

We assume that the MANO orchestrator is used and it has the capability of the virtual resources dynamic allocation according to slice needs (resource scaling). In that context, we evaluate two cases: (i) underutilization of allocated resources and (ii) overutilization of resources. In fact, for both cases, the same issue of resource allocation is analyzed, and the allocation is done by the MANO orchestrator. Overutilization of resources may lead to the degraded performance of the sliced solution, whereas underutilization of resources leads to ineffective network slice implementation. In this paper, we propose to set two thresholds for too high (Th_{hi}) and too low (Th_{lo}) resources usage. We introduce also the observation period T_o, which is used for KPIs reporting. During the time the measured parameter is averaged. We propose to use $Th_{hi} = 80\%$ and $Th_{lo} = 20\%$ and observation interval $T_o = 30$ s. However, other values for thresholds as well as for observation time can be used. It has to be noted that reducing the observation time may lead to significant overhead related to KPIs calculations. Each of the measured KPIs has its timestamp. The proposed KPIs concerning resource usage by a slice are the following:

- KPI-R1: *Connectivity Resources Underutilization (ConRu)*. The KPI is calculated periodically as a number of virtual links of a slice with utilization under Th_{lo} of link capacity during the observation time T_o. In properly allocated connectivity resources such situation should

be rare. The abundant connectivity resources have no negative impact on slice behavior, but they affect the overall efficiency of the resources usage.

- KPI-R2: *Connectivity Resources Overutilization (ConRO)*. The KPI is calculated periodically as a number of virtual links of a slice which utilization is over Th_{hi} of link capacity during the observation time T_o. In properly allocated connectivity resources such situation should be rare. The lack of connectivity resources leads to increased transmission delay and packet loss. Therefore, it degrades slice behavior.
- KPI-R3: *Computing Resources Underutilization (ComRU)*. The KPI is calculated periodically as virtual CPU utilization of each VNF of a slice and represents the number of VNFs which computing utilization is under Th_{lo} during the observation time T_o. The abundant computing resources have no negative impact on slice behavior, but they affect the overall efficiency of the infrastructure resources usage.
- KPI-R4: *Computing Resources Overutilization (ComRO)*. The KPI is calculated periodically as virtual CPU utilization of each VNF of a slice and represents the number of VNFs which computing utilization is over Th_{hi} during the observation time T_o. The lack of computing resources leads to increased processing time. Therefore, it degrades slice behavior.
- KPI-R5: *Memory Resources Underutilization (MemRU)*. The KPI is calculated periodically as virtual memory utilization of each VNF of a slice and represents the number of VNFs with memory utilization of under Th_{lo} during the observation time T_o. The abundant memory resources have no negative impact on slice behavior, but they affect the overall efficiency of the infrastructure resources usage.
- KPI-R6: *Memory Resources Overutilization (MemRO)*. The KPI is calculated periodically as virtual memory utilization of each VNF of a slice and represents the number of VNFs with memory utilization of over Th_{hi} during the observation time T_o. Lack of memory resources leads to increased processing time. Therefore, it degrades slice behavior.

For all KPIs above we propose to calculate the absolute as well as normalized values, e.g. number of links in which the threshold has been crossed related to all links of the slice. In the case of memory KPIs we propose a synthetic approach. ETSI [17] enables monitoring of different types of memory (RAM, swap and disc space). We have simplified the approach and if at least one type of VPN's monitored memory resources crosses the threshold, the KPI is affected. The KPIs proposed above are oriented towards a single slice. Another measure of proper, dynamic resources allocation can be taken on the system level, i.e. by evaluation of under- or overutilization of all (aggregated) connectivity, computing and memory resources. In that context we define:

- KPI-R7: *Overall Connectivity Resources Underutilization (OConRU)*. This KPI is defined as the aggregation of *ConRU* KPIs of all slices.
- KPI-R8: *Overall Connectivity Resources Overutilization (OConRO)*. This KPI is defined as the aggregation of *ConRO* KPIs of all slices.
- KPI-R9: *Overall Computing Resources Underutilization (OComRU)*. This KPI is defined as the aggregation of *ComRU* KPIs of all slices.
- KPI-R10: *Overall Computing Resources Overutilization (OComRO)*. This KPI is defined as the aggregation of *ComRO* KPIs of all slices.
- KPI-R11: *Overall Memory Resources Underutilization (OMemRU)*. This KPI is defined as the aggregation of *MemRU* KPIs of all slices.
- KPI-R12: *Overall Memory Resources Overutilization (OMemRO)*. This KPI is defined as the aggregation of *MemRO* KPIs of all slices.

The increase of above-mentioned KPIs can be observed when the orchestrator part responsible for resources allocation (VNFM, VIM) is overloaded or driven by not optimal resource allocation algorithm.

There exists multiple RAN virtualization approaches. In most of them limited RAN virtualization is used, then other techniques of allocation of resources to slices are often used, e.g. a double-level scheduler responsible for Resource Blocks allocation. One level of the scheduler is used for splitting the radio resources (Resource Blocks) between slices; another one is responsible for the allocation of resources to slice users. The intra-slice scheduler KPIs can be linked with classical RAN KPIs. The above-proposed KPIs related to connectivity can be used, whereas the memory- and computing-usage KPIs make sense only in case of virtualized implementation of RAN nodes.

B. Slice life-cycle KPIs

The list of the proposed KPIs is presented in subsequent subsections.

- KPI-L1: *Slice Deployment Time (SDT)* is a parameter that describes the interval between the slice deployment request and the moment in which slice is ready for operation. The problem with this parameter is that this interval depends on "slice template (blueprint) complexity", the performance of orchestrator, and the allocation of resource time by the virtualized infrastructure. The slice complexity may deal with the footprint size of VNFs, their interconnection topology, amount of configuration parameters. Therefore, in a generic case, it is impossible to define the required value of *SDT*. It can be noted that SDT can be critical for some network slices, e.g. in case of on-demand or short-lived ones, but much less critical for long-lived slices.
- KPI-L2: *Slice Deployment Time Scalability (SDTS)* is a measure of scalability of slice deployment operations. To evaluate the scalability, we propose to send N slice

deployment requests of the same slice template and calculate *SDTS* in the following way:

$$SDTS = \frac{GSDT}{N \cdot SDT}$$

where *GSDT* is the overall time for the deployment of *N* identical slices and *SDT* is the deployment time of a single slice (as defined above). It is hard to define the *N* value *a priori*. If the *N* value is too big, then there can be a problem with the availability of the requested resources. If it is too small, then the obtained result may not express the scalability of the orchestration well. It can be recommended to calculate *SDTS* parameter for *N* = 10. *SDTS* is expected to be greater than 1.

- KPI-L3: *Reconfiguration Execution Time (RET)*. There are two reasons for run-time slice reconfiguration. The first is driven by NFVO, which decides to move an NFV from one data center to another. Such reconfiguration is not driven by orchestrator operator but is an autonomic decision of NFVO. The network slice template is unchanged. Such an operation is executed in the background. Therefore, no KPIs in such case are needed. Another kind of reconfiguration is a reconfiguration that is needed because of changes of the slice blueprint, for example by adding or removing VNFs. Such operation is driven by OSS, and its completion should be reported to OSS. Having in mind the operation, we propose to define a KPI called *Reconfiguration Execution Time (RET)*, which is defined as an interval between slice reconfiguration request and slice reconfiguration completion. It has to be noted that *RET* is not agnostic and depends on the slice template as well as a number of modifications (number of VNFs, number of links to be reconfigured, number of operations related to VNFs re-configuration). Hence, the expected value of this KPI cannot be defined, but it should be kept as low as possible.
  RET has a very specific usage in case of RAN. RAN can be virtualized partly and may be based on PNFs. The reconfiguration of PNFs that already are in use by a specific slice as a change of allocated resources has been already described. In case of RAN slice reconfiguration, the change in the area that is served by a slice, by adding and/or removing of some RAN nodes to a slice, should be considered. Such operation can still use the *RET* KPI. However, the value of *RET* can be different for the different number of RAN nodes involved in the reconfiguration process. It has to be noted, however, that the process of slice coverage modification has not been addressed by 3GPP, yet.
- KPI-L4: *Slice Termination Time (STT)* is a parameter that describes the interval between the slice termination request and the moment in which all slice allocated resources are released. If the time is long, it decreases the efficiency of the infrastructure resources usage.

C. KPIs for multi-domain slicing

The end-to-slice can be created not only in a single domain but also in multiple domains (administrative, technological or orchestration domains). The creation of the end-to-end slice in multiple administrative domains deals mostly with business-related issues as well as the definition of the operations that are allowed in the non-owned domains. The technological domain in the context of network slicing may imply specific orchestration or management functions, for example by the use of specialized hardware, legacy subsystems or special virtualization techniques. In the context of mobile networks, such specific operations are typical for RAN and may concern both, nodes of RAN or data transport (especially the front-haul). In general, there are two options for creating end-to-end slices in multiple domains.

The first one lies in the use of single orchestrator that is able to orchestrate the resources of all domains. Another one lies in the usage of per-domain orchestrator and higher level entity in order to coordinate the behavior of domain level orchestrators and to provide the end-to-end operations. Such case has the ability to handle domain-specific orchestration issues, but it requires well-defined types of local slices and their descriptors that can be selected for defining the end-to-end slice. From the KPI calculation point of view, the first case, i.e. single orchestrator-based, has no impact on KPI calculation – they are computed in the same way as in the single-domain case described above.

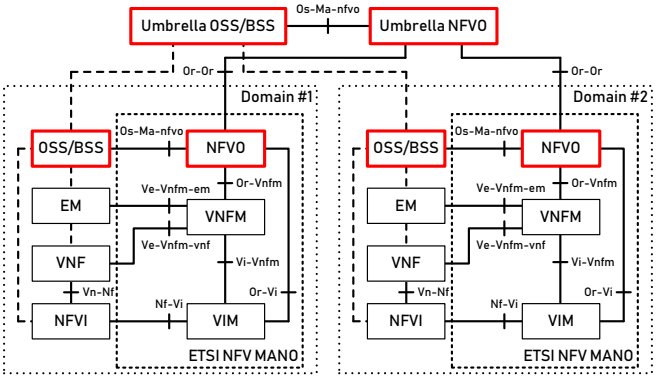

Fig. 1. Multi-domain management and orchestration architecture for end-to-end KPI calculation

The second case is much more complicated. Here, we propose computation of domain-level KPIs by local OSS/BSS and passing this information to the end-to-end orchestrator (*Umbrella NVFO*, which may also play a role of a local domain NFVO, cf. [27]). This orchestrator exposes overall KPIs as well as domain-level KPIs to the end-to-end slice operator and slice tenants. The overall KPIs are computed using the operations on domain-level KPIs of all domains. In some cases, the overall KPIs can be obtained by summing of domain-level KPIs in other cases other operations on-local KPIs are applicable. The end-to-end slice setup time is calculated as an interval between the first domain-level slice setup request and last domain level confirmation of

slice deployment. Please, note that in opposite to some multi-domain network slicing concepts, we see a significant role of domain-level OSS/BSS. The concept is illustrated in Fig. 1. The *Umbrella OSS/BSS* can collect and expose not only the network slices' KPIs but also KPIs of "sliced" solutions.

IV. Network slicing KPIs computation in the NFV MANO case

The KPIs described in Section III, in order to be calculated in the MANO environment, require some information about the resource allocation, usage, as well as the occurrence of certain operations and their completion time. In fact, we may split the required for KPIs data in the following way:

- information related to computing, memory, storage and connectivity resources allocated to VNFs and consumed by them;
- information about initiation and completion time of selected NFVO procedures that are driven by OSS/BSS;
- information about VNFM operations (initiation, completion).

In the presented approach, we propose to use the OSS/BSS of the MANO architecture to calculate and collect the network slicing KPIs. For the collection of information required for KPI calculation, the OSS/BSS has to interact with other components of the MANO architecture. VIM exposes the information about the underlying NFVI at the reference points *Vi-Vnfm* [28] and *Or-Vi* [29] to higher-level MANO entities, VFNM and NFVO. These entities are able to understand, correlate and further enrich this received information in the context of the installed network service description. The OSS/BSS can directly use the *Os-Ma-Nfvo* reference point of NFVO [30], for the purpose of: Network Service Life-cycle Management (instantiating, scaling, updating, healing, terminating, deleting, etc.), Performance Management (management of performance management jobs and thresholds), Fault Management (management of subscriptions to notifications, querying alarms lists, acknowledging alarms) and NFVI Capacity Information (querying and notifications about underlying infrastructure capacity and its shortage). Hence, the OSS/BSS is able either to determine the life-cycle operations performance based on request-response time interval or get directly the subscribed or requested run-time performance/fault/capacity information. While the information exchange between NFVO and OSS/BSS is at the level of Network Service Instance, the individual VNF's Element Manager (EM) is partially able to exchange similar information with the VNFM at the level of its VNF/VNFCs via the reference point *Ve-Vnfm-em* [31] and to share further this information with its own OSS. Here, Performance Management and Fault Management interfaces are available. In case of multi-domain operations [32], the reference point *Or-Or* between the producer-NFVO and the Umbrella NFVO copies the definition of Performance, Fault and Life-cycle Management interfaces exposed at *Os-Ma-Nfvo*. Additionally, the Network Service Instance Usage Notification Interface at *Or-Or* provides the awareness of the delivered Network Service utilization status to the producer-NFVO, which can accordingly adapt, e.g. the activities related to performance or the fault reporting for the specific Network Service.

The ETSI GS NFV-IFA 027 [17] concerns MANO performance issues and describes mechanisms of premium importance for KPIs calculation:

- VIM uses reference points *Vi-Vnfm* and *Or-Vi* to report NFVI-related performance indicators to VNFM and VNFO, respectively. The performance metrics include mean/peak usage of virtual CPU, memory, disk, and virtual storage, number of incoming/outgoing bytes/packets on the virtual computer (split per virtual interface) or virtual network (split per virtual port);
- VNFM maps the above-mentioned information from VIM to specific VNFs/VNFCs and exposes the performance measurements at reference points *Ve-Vnfm-em* (for VNFs/VNFCs) and *Or-Vnfm* (for VNFs only). These are VNF/VNFC-specific mean/peak usages of virtual CPU, memory, disk and virtual storage, numbers of incoming/outgoing bytes/packets at VNF internal/external connection points;
- The performance measurements produced by NFVO can be transferred to OSS/BSS via the reference point *Os-Ma-Nfvo*. They include numbers of incoming/outgoing bytes/packets at Network Service border interfaces.

According to ETSI GR NFV-EVE 008 [33], which deals with charging and billing, MANO enables charging of two categories: Usage Events and Management and Orchestration Events. Both types of events can be used in order to calculate KPIs. The Usage Events can be used for resource usage monitoring (computing, storage, networking). The VIM is responsible for such monitoring. The VNF instance monitoring (for example the VNF Instance scaling) is done by NFVO/VNFM, and the Network Service Instance is monitored by NFVO. The mentioned recommendation place the charging-related entities as a part of NFVI, VIM, and OSS/BSS. These functions are focused on averaged usage, but the continuous monitoring of resources is possible as well.

The presented capabilities of MANO enable OSS/BSS to collect data necessary for network slicing KPIs calculation and correlation. These data, processed mainly by VNFM, can be obtained via several paths by the direct interaction of OSS/BSS with NFVO or through EM. The EM of VNF can also be implemented in that way that it will calculate VNF-level KPIs directly. In some implementations, the OSS/BSS can interact with NFVI directly in order to obtain knowledge about resource allocation and consumption. The ways in which the required information is collected by OSS/BSS is partly implementation-dependent and therefore cannot be defined *a priori*. However, MANO provides enough information to calculate all the network slicing KPIs defined in Section III.

The performance management abilities of ETSI NFV MANO framework allow for the direct collection of all resource-related metrics defined in the paper. The mechanism called Performance Management Job (cf. [30], [31]) enables

TABLE I
CALCULATING KPIS USING MANO

KPI	MANO role in KPI calculation
ConRU, ConRO ComRU, ComRO MemRO, MemRU	The KPIs can be calculated using information obtained in several ways: (i) VNFM which produces the measurements on the basis of information from VIM, shares them with NFVO, which can further pass them to OSS/BSS, where KPIs will be calculated. (ii) EM can receive the reports for its VNF/VNFC from VNFM and pass them to OSS/BSS for KPIs calculation or calculate KPIs locally and pass the results to OSS/BSS.
OConRU, OConRO OComRU, OComRO OMemRO, OMemRU	These KPIs are calculated by OSS/BSS as aggregates of respective KPIs of single slices. The operation concerns all active slices.
SDT, SDTS, SRT, STT	The operations that are related to these KPIs are triggered by OSS/BSS and executed by NFVO. The NFVO reports their completion to OSS/BSS. Therefore computation of these KPIs is based on "request to response" time.
Multi-domain KPIs	The multi-domain OSS/BSS obtains all KPIs from domain-level OSS-es and calculates multi-domain KPIs.

creation of measurements of specified parameters upon the OSS/BSS or EM request. After creation of relevant jobs, the OSS/BSS requests MANO (directly or via EMs) to set thresholds on these measurements and then only the threshold-crossing notifications are sent by MANO entities to the requester. The measurements of computational and memory (all types) resources are produced and exposed as a percentage of maximum value. Hence, their thresholds settings are directly $Th_{hi/lo}$. The connectivity measurements are based on counts of packets/bytes at the measurement points. Therefore, the connectivity overutilization/underutilization thresholds settings should take into consideration also the observation time T_o and the maximum link speed.

The proposed life-cycle KPIs can be obtained using the interaction between the OSS/BSS and NFVO. The relevant procedures are based on a request-response handshake and OSS/BSS has to have the definition of message sequences implemented in API [34]. Hence, it is able to determine clearly both the beginning and the end of the procedure, also in case of disturbances of intra-MANO communication (e.g. OSS/BSS is notified about the delay of procedure execution due to the need of retrying). There are two possible ways of calculation of these KPIs: (i) based on events logging in the on-board OSS/BSS log – each event is logged with a timestamp and correlated search of beginning/finishing event for specific procedure is sufficient; (ii) the API for OSS/BSS-MANO communication will typically use the time-out mechanism and the time-out timer will be implemented – its value at the end of the procedure may be instantly passed to the Performance Management engine of the OSS/BSS. The OSS/BSS operations can be supported by EMs of VNFs in order to increase KPIs calculations scalability.

V. CONCLUSIONS

In this paper, we have described KPIs for network slices. Such set of KPIs is of premium importance for operators in order to compare solutions of different vendors for the verification of proper operations of the installed systems (also as a part of SLAs with business customers) and for trials of newly developed solutions. Our goal was to define a minimal but representative set of KPIs that will describe the network slicing impact on the behavior of the sliced solutions. However, we think that more work is needed for network slicing KPIs evaluation and estimation. For example, the network slice metrics that include the number and size of footprints of all VNFs that compose the slice, number of slice links, number of operations concerning slice configuration can be used for the estimation of slice deployment time.

As we noted in the paper, some KPIs are tightly coupled with MANO orchestration and should be defined as a part of MANO. So far this is not the case – in the paper we have outlined some additional operations and components of MANO that have to be implemented in order to support orchestration-related KPIs calculations. The work on the implementation of the presented concept for the OAI platform [35], in which the EPC is virtualized and sliced, is in progress.

REFERENCES

[1] "NGMN 5G White Paper", NGMN Alliance, Feb. 2015.
[2] "Description of Network Slicing Concept", NGMN Alliance, Sep. 2016.
[3] "Network Functions Virtualisation (NFV); Architectural Framework", ETSI GS NFV 002, V1.2.1, Dec. 2014.
[4] K. Kozłowski, S. Kukliński, L. Tomaszewski, "Open issues in network slicing", 2018 9th International Conference on the Network of the Future (NOF), Poznan, 2018, pp. 25-30.
[5] "Speech and multimedia Transmission Quality (STQ); QoS aspects for popular services in GSM and 3G networks", ETSI TS 102 250 series of standards.
[6] "Study on Key Quality Indicators (KQIs) for service experience", 3GPP TR 32.862, ver. 14.0.0, Mar. 2016.
[7] "End-to-end multimedia services performance metrics", 3GPP TR 26.944, ver. 15.0.0, Jun. 2018.
[8] "Definitions of terms related to quality of service", ITU-T E.800, Sep. 2008.
[9] "Definition of Quality of Service parameters and their computation", GSMA IR.42, ver. 9.0, Jun. 2018.
[10] "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond", ITU-R M.2083-0, Sep. 2015.
[11] "Minimum requirements related to technical performance for IMT-2020 radio interface(s)", ITU-R M.[IMT-2020.TECH PERF REQ] (draft), Feb. 2017.
[12] "Service requirements for next generation new services and markets", 3GPP TS 22.261, ver. 16.6.0, Dec. 2018.
[13] "Study on scenarios and requirements for next generation access technologies", 3GPP TR 38.913, ver. 15.0.0, Jul. 2018.
[14] "Recommendations for NGMN KPIs and Requirements for 5G", NGMN Alliance, Jun. 2016.
[15] "Management and orchestration; 5G end to end Key Performance Indicators (KPI)", 3GPP TS 28.554, ver. 15.1.0, Dec. 2018.
[16] "Management and orchestration; 5G performance measurements", 3GPP TS 28.552, ver. 16.0.0, Dec. 2018.
[17] "Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Performance Measurements Specification", ETSI GS NFV-IFA 027 V2.4.1, May 2018.
[18] "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Requirements and interfaces specification for management of NFV-MANO", ETSI GS NFV-IFA 031, V3.1.1, Sep. 2018.

[19] "NFV Infrastructure Metrics for Monitoring Virtualized Network", Alliance for Telecommunications Industry Solutions Deployments, ATISI0000062, Jan. 2018.
[20] H. Koumaras et al., "5GENESIS: The Genesis of a flexible 5G Facility", 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Barcelona, 2018, pp. 1-6.
[21] "Deliverable D2.1 Scenarios, KPIs, use cases and baseline system evaluation", ONE5G project, Nov. 2017.
[22] "Deliverable D6.1 Documentation of Requirements and KPIs and Definition of Suitable Evaluation Criteria", 5G-MoNArch project, Sep. 2017.
[23] "Deliverable D2.2: 5GCHAMPION Key Performance Indicator and use-cases defined and specification written", 5GCHAMPION project, Mar. 2017.
[24] "Deliverable D2.1 5GCAR Scenarios, Use Cases, Requirements and KPIs", 5GCAR project, Aug. 2017.
[25] "Deliverable D2.6 Final Test Scenario and Test Specifications", 5G Applications and Devices Benchmarking (TRIANGLE) project, Sep. 2018.
[26] "D2.6 Final report on programme progress and KPIs", Euro-5G project, Oct. 2017.
[27] "Network Functions Virtualisation (NFV); Management and Orchestration; Report on Architectural Options", ETSI GS NFV-IFA 009, V1.1.1, Jul. 2016.
[28] "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Vi-Vnfm reference point - Interface and Information Model Specification", ETSI GS NFV-IFA 006, V3.1.1, Aug. 2018.
[29] "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Or-Vi reference point - Interface and Information Model Specification", ETSI GS NFV-IFA 005, V3.1.1, Aug. 2018.
[30] "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Os-Ma-Nfvo reference point – Interface and Information Model Specification", ETSI GS NFV-IFA 013, V3.1.1, Aug. 2018.
[31] "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Ve-Vnfm reference point – Interface and Information Model Specification", ETSI GS NFV-IFA 008, V3.1.1, Aug. 2018.
[32] "Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Multiple Administrative Domain Aspect Interfaces Specification", ETSI GS NFV-IFA 030, V3.1.1, Sep. 2018.
[33] "Network Function Virtualisation (NFV) Release 3; Charging; Report on Usage Metering and Charging Use Cases and Architectural Study", ETSI GR NFV-EVE 008, V3.1.1, Dec. 2017.
[34] "Network Functions Virtualisation (NFV) Release 2; Protocols and Data Models; RESTful protocols specification for the Os-Ma-nfvo Reference Point", ETSI GR NFV-SOL NFV-SOL 005, V2.5.1, Sep. 2018.
[35] OpenAirInterface Software Alliance, https://www.openairinterface.org/.