Network Slicing Cost Allocation Model

Asma Chiha, Marlies Van der Wee, Didier Colle & Sofie Verbrugge

Journal of Network and Systems Management

ISSN 1064-7570

J Netw Syst Manage
DOI 10.1007/s10922-020-09522-3

Springer

Journal of Network and Systems Management
https://doi.org/10.1007/s10922-020-09522-3

Network Slicing Cost Allocation Model

Asma Chiha[1] · Marlies Van der Wee[1] · Didier Colle[1] · Sofie Verbrugge[1]

Received: 15 August 2019 / Revised: 20 December 2019 / Accepted: 28 February 2020

Abstract
Within the upcoming fifth generation (5G) mobile networks, a lot of emerging technologies, such as Software Defined Network (SDN), Network Function Virtualization (NFV) and network slicing are proposed in order to leverage more flexibility, agility and cost-efficient deployment. These new networking paradigms are shaping not only the network architectures but will also affect the market structure and business case of the stakeholders involved. Due to its capability of splitting the physical network infrastructure into several isolated logical sub-networks, network slicing opens the network resources to vertical segments aiming at providing customized and more efficient end-to-end (E2E) services. While many standardization efforts within the 3GPP body have been made regarding the system architectural and functional features for the implementation of network slicing in 5G networks, techno-economic analysis of this concept is still at a very incipient stage. This paper initiates this techno-economic work by proposing a model that allocates the network cost to the different deployed slices, which can then later be used to price the different E2E services. This allocation is made from a network infrastructure provider perspective. To feed the proposed model with the required inputs, a resource allocation algorithm together with a 5G network function (NF) dimensioning model are also proposed. Results of the different models as well as the cost saving on the core network part resulting from the use of NFV are discussed as well.

Keywords NFV · SDN · Resource allocation · Cost model · 5G · Satellite · Network function resource modelling

✉ Asma Chiha
asma.chihaepharbi@ugent.be

Marlies Van der Wee
marlies.vanderwee@ugent.be

Didier Colle
didier.colle@UGent.be

Sofie Verbrugge
sofie.verbrugge@ugent.be

[1] IDLab, Ghent University-imec, Ghent, Belgium

Published online: 10 March 2020

1 Introduction

The exponential increase of the Internet speed together with emerging services and applications are considered obvious from a customer point of view. However, telecom operators face serious investment challenges to meet these increasing expectations of the end user while keeping a reasonable Average Revenue Per User (ARPU) to not experience a customer churn.

Currently, traditional network architectures are static, decentralized and complex. In addition, network infrastructures are designed to fulfil the requirements of dedicated, specific services using specialized network hardware (known as hardware appliances or middleboxes). These services, however, do not make full use of the available resources all the time. This inefficiency leads to high capital costs for building, maintaining and operating the networking infrastructure, leading to long return on investment cycles. Furthermore, the optimization of the network infrastructure for a specific service prevents the addition of new features and services, slowing down or even preventing innovations. Technical improvements and new technological trends, such as SDN, NFV and network slicing, have to be investigated in order to make the network architecture more flexible and agile.

For the sake of flexibility and easy troubleshooting of the network, a centralized network intelligence component is needed. Nevertheless, we cannot achieve a centralized control of the network without disassociating the routing process (control plane) from the forwarding process of network packets (data plane). The SDN paradigm is an approach that enables dynamic and programmatically efficient network configuration by splitting the control plane and the data plane and centralizes the control of the network by using an SDN controller [1].

On the other hand, Network Function Virtualization (NFV) moves the network functionalities from hardware into software such that they can run on a range of standard-based hardware which may easily be moved to any location within the network if needed. The focus of the NFV concept is on reducing the need to add new equipment. This enables more flexibility to scale up or scale down, and more opportunity to innovate, experiment and deploy new services with lower risk [2].

These new emerging technologies being SDN and NFV are jointly beneficial and complementary to each other since they both share the same feature of encouraging innovation, creativity and competitiveness [1, 3]. Therefore, NFV and SDN together are considered as the key enablers of the network slicing concept. According to [4]: "A network slice is viewed as a logical end-to-end network that can be dynamically created. A given User Equipment (UE) may access to multiple slices over the same Access Network (e.g. over the same radio interface). Each slice may serve a particular service type with agreed upon Service-level Agreement (SLA)". Network slicing allows different service providers with disparate traffic requirements to share the same infrastructure resources. This does not mean deploying Virtual Local Area Networks (VLANs) to isolate traffic flows, but it means splitting physical network resources and network functions to deploy an exclusive end-to-end (E2E) implementation for each application, e.g. enhanced Mobile Broadband (eMBB), ultra-Low Latency Communication (uRLLC).

The deployment of multiple network slices requires the network to reserve enough resources for each specific slice instance according to its requirements. For the radio access network (RAN) part, the network needs to allocate physical radio resources to each specific network slice based on the required bandwidth. On the core network part, an isolated set of logical network functions may be used which are provisioned for this specific network slice combined with logical and physical network functions which are shared among multiple network slice instances. Network slicing gives Mobile Network Operators (MNOs) the opportunity to open their virtualized networks to vertical segments such as the healthcare sector, car manufacturers, intelligent transport providers etc., as well as to service providers that lack physical network infrastructure. MNOs need resource allocation models to distribute the resources among themselves and their tenants in a way that first adheres to the SLA agreed beforehand and, second, makes effective use of the network. Resource allocation problem is seen as one of the challenges that face slicing-enabled networks according to many surveys on network slicing [5, 6].

However, not only the allocation of network resources to the different slices is considered as a crucial element towards the efficient use of the network slicing—and hence the efficient use of the network—also techno-economic challenges pop-up with the introduction of this new technique [5]. From a techno-economic perspective, two main research questions are raised due to the use of the network slicing concept, being, first, how much cost can be saved by using network virtualization and network slicing and second, how to allocate the cost of the network to the running network slices/services?

In this paper, we propose a cost allocation model for network slicing that aims to allocate the cost of the network to the different network slices deployed on this network, as well as a cost model to derive how much can be saved by using virtualization on the core network. To this end, we present, in the next section, a literature review of the allocation models for network slicing and cost allocation models in general. In Sect. 3, we describe in a detailed way the proposed model, which requires a slice dimensioning model in order to distribute the cost of the network to the running slices based on their consumption in terms of network resources. The resource allocation model or slice-dimensioning model demands the hardware requirements of each 5G network functions NFs as inputs. Thus, Sect. 4 explains how we modelled the 5G network functions requirement for both the control and the data plane. Section 5 presents the results of the application of the developed models to a specific use case, which was selected based on availability of cost model and slice assumptions. Finally, Sect. 6 concludes the paper and presents our planned next steps.

2 Related Work

In order to allocate the cost of the network to the different slices, we have to follow an optimized resource allocation model, optimized in the sense that this algorithm assigns the needed network resources efficiently to each slice taking into consideration all Key Performance Indicators (KPIs) and dimensioning.

However, in literature, the resource allocation models are often designed in order to optimize one of the network metrics e.g. round-trip time (RTT), Signal-to-noise ratio (SNR) etc. and several of them focus only on the radio resources and spectrum sharing, not on the core network resources. For example, authors of [7] proposed a subchannel allocation algorithm that allocates radio frequencies in spectrum-sharing two-tier systems to the different tiers while considering the co-tier interference and cross-tier interference. In addition, in [8] as well, a radio resource framework is designed in order to allow tenants to dynamically and flexibly distribute base station resources among their users taking into account both the admission control and user dropping mechanisms. Authors in [9] consider that by using cloud RAN (C-RAN), network slicing at the spectrum level can be easily implemented. The C-RAN technology is based on moving the Baseband Units (BBUs) from distributed base station locations into a centralized BBU place. This centralization of BBUs aims to achieve a multiplexing gain and increase network capacity by using load balancing and cooperative processing of the traffic sent by different base stations.

On the core network side, a network slice is seen as a set of Virtual Network Functions (VNFs) which are deployed on virtual machines (VMs) or containers. The resource allocation for the core network resources is a matter of mapping VNFs to VMs based on their hardware requirements such as processing power, memory and storage. Researchers often see the allocation problem as (1) a forecasting technique that requires three steps to be implemented: (a) a forecasting module to predict the traffic of a network slice based on previous traffic and user mobility, (b) a network slicing admission control algorithm, (c) a network slicing scheduler algorithm to fulfil SLAs and report back to the forecasting module [10] or as (2) an VNF-RA (VNF resource allocation) problem as presented in a wide survey elaborated by [6]. Such an VNF-RA problem consists of 3 stages: VNFs—chain composition, VNF-forwarding graph embedding and VNFs—scheduling. However, few researches consider the three steps of the resource allocation problem together and no research considers the VNF itself. e.g. the consumed time of the VNF installation and removal processes.

The proposed resource allocation models for the core network part somehow assume the infinite availability of network resources and consider a cloud-based network model where the allocation of resources is a matter of optimization. However, for network infrastructure providers and data centres who own the network infrastructure, allocation models are not only needed to efficiently use the resources, but also as a forecasting tool of the required investment in the future to cope with the growth of demand and as an input to the cost allocation model.

To conclude, the state of the art of the resource allocation models shows that a complete model that allocates both radio and core network resources is still missing in literature.

On the economic side, SDN, NFV and network slicing technologies add more flexibility and easy control and to the network management in addition to the efficient use of the network resources. Certainly, all these advantages have an impact in term of cost. For example, the use of NFV may reduce the Capital Expenditures (CAPEX) due to the possibility to deploy VNFs on a vendor-independent hardware. SDN can lead to a possible reduction of the Operational Expenditures (OPEX)

because of the easy troubleshooting and maintenance of the network. However, this latter may also add more complexity due to the centralization of the network control, hence increasing the signalling traffic on the control plane. Therefore, the influence of these new technologies on the cost of the network must be investigated carefully. Many researches in literature investigated the impact of SDN, NFV and network slicing from a techno-economic perspective and are summarized in the following Table 1.

These researchers focused on the quantification of the cost saving resulting from the use of SDN and NFV [11–14], but none of them tackled the cost model for network slicing. For example, [15] presented a revenue model for network slicing based on MOOP and [16] proposed a new business and services model for network slicing as a service. However, how much exactly does an eMBB slice that requires a high network capacity costs, consume compared to an uRLLC slice that requires low latency? How can a network infrastructure provider dimension slices effectively considering the hardware requirement of each slice, in order to reduce costs and maximize revenues? Though they are very interesting research questions, to the best knowledge of the authors, there is no paper available so far that models the allocation of the network costs to the different slices.

Cost allocation models are built using two main components, being (1) the cost model of the sliced network and (2) the hardware requirements of each of the network slices. Both parts are indispensable for any operator to ensure a fast Return on Investment (ROI). To this end, we propose, in the next section, our proposal for a network slicing cost allocation model.

3 Network Slicing Cost Allocation Model

A slice is a logical subnetwork defined on top of the physical network to deliver a specific service such as video streaming, Internet of Things (IoT) services, etc. This logical network is customized in a way to fulfil a set of KPIs: availability, reliability, capacity, efficiency, and latency etc. To satisfy those KPIs, different types of network resources have to be reserved for this specific slice.

The network consists of three parts: core network, transport network and radio access network. An E2E slice has to be built taking into account these three different network components. On the core network side, a slice is a chain of VNFs and physical network functions. Yet, on the transport network part, the slice can be seen as a pipe or tunnel with a specific bandwidth reserved for this slice. However, on the RAN side, the slicing is made by means of a frequency subchannel reservation. Therefore, the question that raises here, as discussed previously in the introduction, is how to derive the cost of the slice given its exigence in term of these different resource types?

In addition, from a network infrastructure provider point of view, a model that allocates cost to the different slices is a useful input to their pricing models. From an MNO's point of view, such a model is also needed in order to identify what are the most consuming slices in term of hardware and radio resources.

Table 1 SOTA of cost–benefit model for network slicing

Paper title	Studied technology	Proposed model	Methodology and studied use case	Main results
Techno-economic analysis of software defined networking as architecture for the virtualization of a mobile network [11]	SDN & virtualization	Adaptive cost model for SDN concept	Cost model for SDN networks but this model considers only the CAPEX German reference network scenario: - Classical scenario: a distributed network architecture with distributed network control - SDN scenario: a centralized network architecture with decoupled network control from data plane using Open-Flow as communication interface - Sharing scenario: network virtualization and network sharing between several network operators with a FlowVisor controller	- SDN and virtualization of the first and second aggregation stage of the network infrastructure leads to considerable CAPEX cost reductions for the mobile network operator - A 13.81% CAPEX reduction can be achieved for the SDN scenario in comparison with the classical scenario - A 58.04% CAPEX reduction can be achieved for the SDN based sharing scenario in comparison with the classical scenario
Cost Modeling for SDN/NFV Based Mobile 5G Networks [12]	SDN & NFV	Cost model	Comparison between Traditional network (with Virtual EPC) and cloud RAN network also with vEPC for Sweden network However, authors consider only for the OPEX the cost of the power consumption (no maintenance cost nor a failure cost etc.)	- A 63% OPEX reduction can be achieved in comparison with the traditional scenario - The considered CAPEX could be reduced by 68% in comparison with the traditional scenario - The considered TCO could be reduced by 69% in Comparison with the traditional scenario

Table 1 (continued)

Paper title	Studied technology	Proposed model	Methodology and studied use case	Main results
Cost Efficiency of SDN in LTE-based Mobile Networks: Case Finland [13]	SDN	Cost model	Cost model taking into account both CAPEX and OPEX applied to a Finnish reference network	- SDN decreases the network related annual CAPEX by 7.72% and OPEX by 0.31% compared to non-SDN LTE - The cost reduction is a small fraction of the total expenses of a Finnish MNO, but may has a important influence on the profit levels
Life-cycle cost modelling for NFV/SDN based mobile networks [14]	NFV/SDN	Life-cycle cost (LCC) models	Cost model including both CAPEX and OPEX used to compare non-virtualized network costs with virtualized one with many options The modelling considers the hardware requirement for each virtualized network function	- The non-virtualized network has the highest TCO - Among the virtualized flavours, the setup with 6WINDGate speed-up technology is cheapest
Modeling Profit of Sliced 5G Networks for Advanced Network Resource Management and Slice Implementation [15]	Network slicing	Revenue model	Multi-Objective Optimization Problem (MOOP) to increase operator Profit: mathematical optimization	- Novel methodology of modeling profit generated by 5G network slices - The expenditures and revenues of a network can be estimated according to its slice properties, such as KPI requirements and service prices - Based on this, the slice implementations can be optimized to maximize the profit

Table 1 (continued)

Paper title	Studied technology	Proposed model	Methodology and studied use case	Main results
Network Slicing as a Service: Enabling Enterprises' Own Software-Defined Cellular Networks [16]	Network slicing	Business Model and Service Model for NSaaS	Used a detailed process of NSaaS concept by typical examples, together with the configuration process, product management of NSaaS and management APIs for customers	Detailed business and service model for the NSaaS concept that help operators to offer tailored end-to-end cellular network as a service

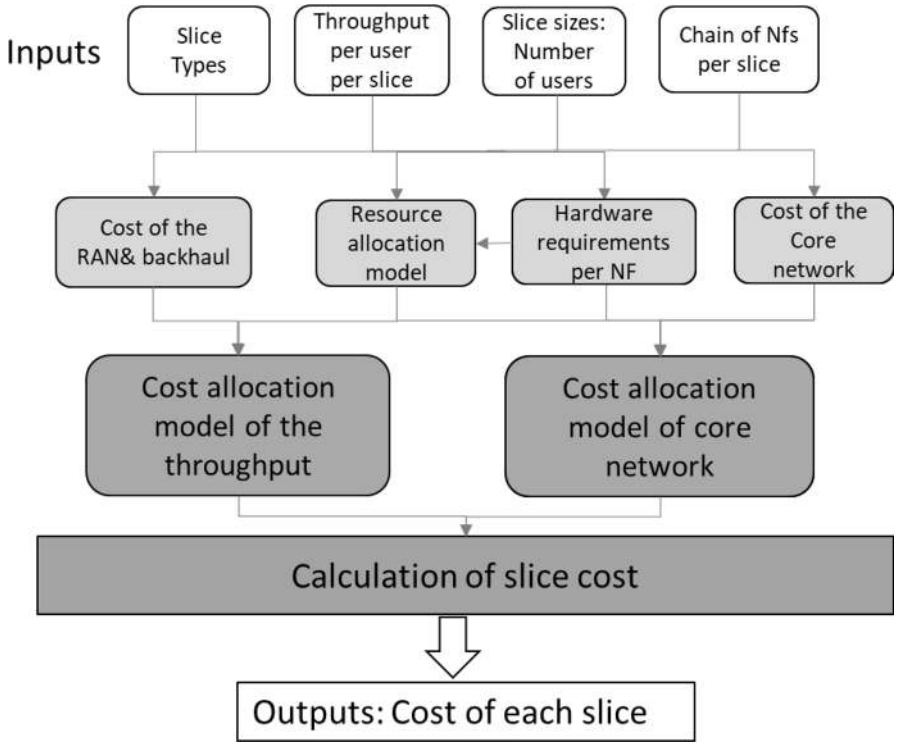

Fig. 1 Cost allocation model diagram

Therefore, we propose a cost allocation model that aims, first, to fairly allocate network costs to deployed slices (services) and, second, to show that network slicing makes more efficient use of the network. The proposed model is built on the following assumptions[1]:

- A network slice will support one specific service.
- A predefined throughput is reserved per service on the RAN and the transport network in a static way.
- On the core network side, a service/slice consists of a chain of (virtual) network functions, such as 5G functions besides physical network functions.
- Virtual network functions (VNFs) are running on virtual machines (VMs), which are in turn executed on servers, and one VM can only host one VNF.

The different building blocks of the suggested model are presented in Fig. 1. The inputs of the model consist of two different type of inputs: (i) available inputs which are related to the use case (e.g. slice type and the number of users) or can be derived

[1] The authors are aware that these assumptions simplify the modelling but aim to extend the model to more complex cases in a later stage (as will also be described in Sect. 6).

easily from literature e.g. the chain of network functions per slice, those inputs are represented in the diagram with white boxes; (ii) derived inputs from other models that we developed in order to feed this model with the required data, those are represented with the light grey boxes.

Both the cost model of the RAN and the backhaul network and the cost model of the core network used as cost inputs to this model are described in [17]. Since a complete model that allocates radio, transport and core network resources in the context of a sliced network is still missing in literature, and since existing models do not dive deep into the VNF characteristics-level of granularity (such as how much memory, processing power and storage each VNF needs to mitigate the overall QoS imposed by the slice), building our cost allocation model in the light of an existing resource allocation model is not possible. Consequently, we propose an allocation model for network slicing based on the hardware requirements and the throughput required for each slice. This model aims at allocating the right amount of resources to each slice. In order to accomplish this, we have to distinguish between the three type of network resources discussed previously: RAN, transport network and core network resources. The RAN and transport network resources can be allocated by reserving a predefined throughput per slice on the base station and the backhaul network. Yet, for the core network resources this allocation is done in two steps: (i) identify the hardware requirements for each 5G VNF and (ii) map the 5G NFs onto VMs and then onto servers. It is noteworthy that the step of the hardware requirements identification per VNF needs a separate model that we developed and will be described in Sect. 4.

Afterwards, a cost allocation model to assign the cost of each resource type to each slice is needed. The cost of the RAN and transport network can be modelled as the cost of the throughput reserved on these two network parts and is presented in Sect. 3.1. For the core network part, Sect. 3.2 summarizes the cost model of the core network resources. Hence the cost of a slice can be derived using Eq. 1.

$$C_{slice} = C_{Thp} + \sum_{i=1}^{K} C_{VNF(i)} + \sum_{j=1}^{l} C_{phNF(j)} \tag{1}$$

with C_{slice} is the cost of the slice; C_{Thp} is the cost of the throughput of the slice; C_{VNF} is the cost of the VNF; K is the number of VNFs in the slice; C_{phNF} is the cost of physical NF; l is the number of physical NFs in the slice.

3.1 Cost Allocation Model of the Throughput

For the throughput metric, a predefined throughput is reserved per slice in a static way.[2] The cost of this metric can be concluded based on consumed throughput from the overall throughput provided by the base station. The throughput required by

[2] In the current version of the model, we consider that the traffic is static. Yet, in a later stage of the model we will include the dynamicity of the traffic, hence the allocation of the network resources will be dynamic as well (as specified in Sect. 6).

each slice has to be reserved over the transport link as well. Hence, the cost of the throughput should incorporate a part of the cost of the transport link capacity. Moreover, each slice has its own exigence in terms of quality of service (QoS), such as priority level, packet error loss rate, packet delay etc. Thus, for each slice, we reflect the cost of the QoS requirement by the mean of a weighting coefficient. This latter takes into consideration the required throughput of the slice as well as the desired priority level, thus the latency requirement, as shown in Eq. 3. Therefore, the cost of the throughput for a specific slice s is calculated using Eq. 2.

$$C_{Thps} = \left(C_{BS} + C_{TrCap}\right) \times \frac{Qco_s}{\sum_{s=1}^{N} Qco_s} \tag{2}$$

$$Qco_{Sj} = Cost_{sharC} \times \frac{Thp_{Sj}}{\sum_{i=1}^{N} Thp_{S(i)}} + \left(1 - Cost_{sharC}\right) \times \left(1 - \frac{L_{Sj}}{\sum_{i=1}^{N} L_{S(i)}}\right) \tag{3}$$

with N is the number of slices running; C_{BS} is the cost of the base station; L_S is the Latency of the slice S; Qco is the weighting coefficient according to throughput and the QoS level of the slice; C_{Thps} is the cost of throughput of slice s; C_{TrCap} is the cost of the transport capacity link; $Cost_{sharC}$ is the Cost sharing coefficient.

To be able to quantify the QoS weighting coefficient for each slice, we need to know the latency for the video and voice services as well as the throughput. For the former, we adopt the agreed 3GPP traffic types described in detail in [18], the latter will be provided in the scenario description used to validate the model.

3.2 Cost Allocation Model of Core Network Resources

In order to fairly allocate the core network resource costs to the different slices based on the hardware requirements of each slice, we should first find a way to map those requirements to the resources used. This can be elaborated following two different approaches: an infrastructure provider or a telecom operator approach. An infrastructure provider, responsible for the operations of a data centre, tries to find a way to allocate the total cost of the data centre to the running services/slices to maximize the ROI. A telecom operator, on the other hand, leases the slices from the data centre and considers the business model of pay-per-use. A telecom operator hence only needs to know the cost of the service that he wants to offer (hence the cost of the slice that he wants to lease) beforehand to study the viability of providing such a service. Giving this, the cost allocation model differs according to whom it is designed for. In this section, we will detail the cost allocation model from a telecom infrastructure provider perspective.

In order to achieve this goal, the infrastructure provider needs to allocate resources to the different slices according to the latter's requirements in term of processing power, memory and storage capacity. This allocation consists of two steps mapping process: (1) mapping the VNFs to VMs, (2) mapping VMs to the data centre resources namely servers, as shown in Fig. 2, and (3) the allocation of the

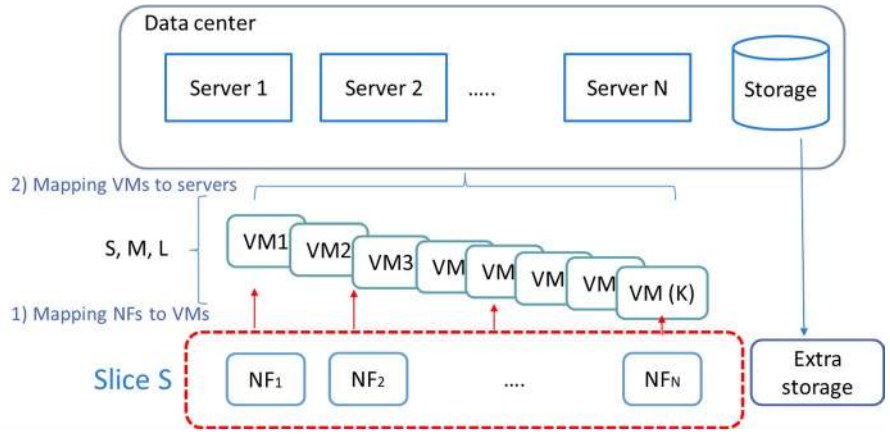

Fig. 2 Mapping data center resources to slices

network costs to the different slices can be achieved based on the resource consumption of each slice, which is the final step of the cost allocation model.

3.2.1 Step 1: Mapping VNF to VM

In order to accomplish the mapping of the VNFs onto VMs, we should identify first, based on the slice characteristics, the technical requirements for each VNF in term of computing, storage and networking. Giving those requirements, we can pick up the suitable VM for each specific VNF. To the best of authors knowledge, there are no references describing these requirements for the 5G networks available in the literature. Therefore, we propose, in Sect. 4, our model to identify the hardware requirements of each 5G NF based on the number of users and the offered traffic of the slice.

After identifying the requirement of each NF, we should, then, be able to select the suitable VM for this NF, which is explained in the next section.

3.2.2 Step 2: Resource Mapping: VM to Data Centre Resources

There are three options to map the VMs to servers and storage resources within a data centre: (i) VM with fixed configuration, (ii) VM with predefined categories and (iii) Personalized VM according to the need of the NF. The first option was adopted in the model used in [14]; a simple mapping between the blade server resources and the resource offered to each VM is elaborated by dividing the server capacities in an equal way among VMs. However, this method is not efficient in term of resource usage; because VMs can be over- or under-dimensioned. For this, the second option with different VM classes or types was suggested. In this second option, the VMs can be classified into classes (for example: small, medium and large) such that the mapping between the VNF and VM can be more accurate. Finally, the third option allows to create a personalized VM according

Table 2 VM classification

	Small	Medium	Large
Core	1	4	8
RAM	8 GB	16 GB	40 GB
Temporary storage	50 GB	100 GB	1500 GB

the NF requirements. Though, this option assigns the required resources to each NF the most accurately, it considers only a static traffic without a margin of safety if the NF would receive more traffic than expected. This option also increases the unused resources, hence raises the question about how to allocate the unused resources cost to the different slices? Hence, for our allocation model we chose the second option of mapping VMs onto the data centre resources, being the predefined categories of VMs. Those categories are presented in Table 2.

After selecting the suitable mapping option, a server consolidation algorithm that considers the co-location of the highly communicating VMs such as [19, 20], can be applied. It aims at placing the different VMs at the suitable server to reduce data centre network traffic and thus decrease the latency for the deployed slices.

The algorithm of mapping VNFs to VMs is described in the following table:

Algorithm: NF to VM mapping algorithm
Input: Set of slices S= {S_1, S_2, …, S_N}; set of network functions per slice NF= {NF_1; NF_2; NF_M};each NF:{CPU(NF), RAM(NF), HDD(NF)} **set of VMs with different hardware characteristics VM={S, M, L};**
Output: allocated VMs for each slice
1: **FOREACH** slice in S **do**
2: **FOR** i=1 to M **do**
3: **FOREACH** j in **VM do**
4: vm_{ij}(CPU)=Ceil (CPU(NF_i)/ CPU_j);* nb of VM type j to satisfy the CPU required by NF_i *\
5: vm_{ij}(RAM)=Ceil (RAM(NF_i)/ RAM_j);* nb of VM type j to satisfy the RAM required by NF_i *\
6: vm_{ij}(HDD)=Ceil (HDD(NF_i)/ HDD_j); * nb of VM type j to satisfy the HDD required by NF_i *\
7: vm_j(NF_i)=MAX (vm_{ij}(CPU), vm_{ij}(RAM), vm_{ij}(HDD));
8: **ENDFOREACH**
9: **vm(NF_i)=MIN (vm(NF_i)(S), vm(NF_i)(M), vm(NF_i)(L));**
10: **ENDFOR**
11: **ENDFOREACH**

As inputs of the algorithm, we consider a set of slices S. Each slice S_i consists of chain of network functions NFs. In addition, each VNF has its requirements in term of CPU, RAM and HDD. Similarly, each VM from categories S, M and L has its own characteristics in terms of the same metrics i.e. CPU, RAM and HDD as well.

For each slice in the set S, we calculate the number of VMs of type "j" that satisfy the requirement of each metric (line 4 in the algorithm for CPU metric, line 5 for RAM and line 6 for the HDD). Then, the mapping of VNFs to VMS of type "j" is determined by the maximum number of VMs required based on all the metrics being CPU, RAM and HDD, and this is done for all the three VM categories (line 7).

Afterwards, we pick the category (i.e. small, medium or large VM) that gives the minimum required number of VMs (line 9).

After mapping the right amount of resources to each VNF and hence to each slice, the questions that raises is how to allocate the core network cost to the different slices based on this mapping process?

3.2.3 Step 3: Allocate The Network Costs to Slices

There are two options to distribute the core network costs among the slices. The first option uses the hardware requirements of each slice as a cost driver for splitting the total cost of the core network between the running slices. Here, unused resource costs will be covered by assigning them proportionally to the different services/slices based on the cost driver. The second option relies on allocating only the cost of the used resources to its slice and does not consider the cost of unused resources. Within this option, mapping VMs onto servers (step 2 of the resource allocation algorithm) is used as input to the cost allocation model, wherein only the used servers cost will be allocated to the considered slice.

For our model, we will adopt the first option. Our slices are deployed on the 5G mobile network. Hence, on the core network side, the slice is a chain of 5G NFs. For each 5G NF, we will identify a cost driver based on its key characteristic, as defined by the 3GPP spec 123.502 [21]. For example: for the User Plane Function (UPF), the cost driver is the number of packets per second, yet for the Access and Mobility Management Function (AMF), it is the number of handover requests. Secondly, we prioritize the following technical requirements (CPU, Memory, Storage, Bandwidth) according to the cost driver identified for this NF. For example, for the UPF, the cost driver of number of packets per second leads to the main technical requirements of CPU and memory.

3.2.4 Resulting Allocated Cost

Bringing the three steps together, allows to allocate the cost of the core network resources. As mentioned above, every NF is translated into hardware requirements in terms of CPU, RAM and HDD. Different virtual machines are in parallel defined in terms of the same hardware requirements, which allows mapping both. By using the hardware requirements of each NF as a cost driver, an allocated cost for the NF can be determined from the overall core network cost (as described in more detail in Sect. 5.3).

4 Modelling the Hardware Requirements of the 5G NFs

In order to be able to calculate the cost of the network slices, we need to derive the cost of each 5G network function (NF) of which the network slice is composed. The cost of this latter depends on the hardware requirements of this NF (in terms of CPU, RAM and HDD). In this section, we present our model used to quantify the needed hardware requirement for both control and data plane NFs.

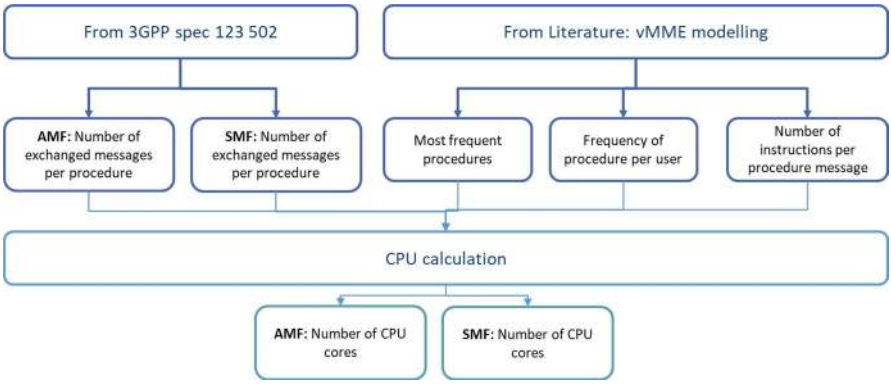

Fig. 3 Traffic quantification model for AMF and SMF

4.1 5G Control Plane Modelling: Focus on AMF and SMF

For the control plane we start by modelling the traffic handled by the AMF and Session Management Function (SMF), then derive the required processing power. The AMF and SMF form together the Mobility Management Entity (MME) in the former evolved Packet Core (EPC) of the 4G network. Therefore, in the approach presented in this section, we base ourselves on the literature that models the control traffic for virtualized MME (vMME).

4.1.1 Model Description

Our proposed model to quantify the needed hardware requirements for the AMF and SMF is presented in Fig. 3. It consists of two main steps. The first step is to understand the vMME modelling and use it as input to model the AMF and SMF. This step consists in identifying: (i) the most frequent procedures; (ii) the frequency of procedure per user and (iii) the number of instructions per procedure.

The second step identifies the number of exchanged messages per procedure for both AMF and SMF and uses it as a driver to split the overall traffic of the vMME between those two functions.

It should be noted though, that when we compare the 4G and 5G service-based architecture, we conclude that there is a difference in terms of the number of messages exchanged between the MME and the rest of the 4G NFs compared to those exchanged between AMF/SMF and the rest of 5G NF. Since AMF and SMF need to communicate with much more NFs than the vMME does (as indicated with the red circles in Fig. 4), we assume the total number of the messages for the AMF and SMF to be the same as the one of the vMME, but we include a correction factor. In [22], authors compared five different Distributed mobility management (Dmm) models of the 45/5G architecture based on the number of exchanged signalling messages for two procedures being the initial attachment and handover procedures.

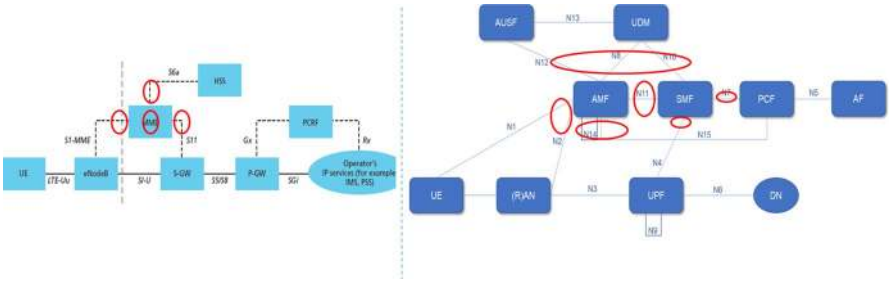

Fig. 4 4G VS 5G service-based architecture

Results of this comparison showed that the number of signalling messages increases with an increase of the number of functional nodes in the core network. More specifically, for an SDN/NFV-based core network architecture, like the 5G core network, the number of signalling messages for the handover procedure is twice the number as for the 4G core network (12 versus 6). Based on these results, we assume a correction factor of two in our model.[3]

The AMF performs most of the functions that the MME performs in a 4G network such as terminating the RAN CP interface (N2), NAS signalling, NAS ciphering and integrity protection, Mobility Management (MM) layer NAS termination, Session Management (SM) layer NAS forwarding, authentication of UE, etc. [23]. Giving the nature of the tasks performed within these functionalities we can assume that the AMF is CPU-intensive. Many researchers in literature such as [24–28] made this same assumption. Moreover, authors in [29] proved by calculating the CPU utilization in function of the number of users that AMF's CPU utilization is the highest among all the network functions (such as AUSF, SMF).

On the other hand, the SMF performs the session management functions that are handled by the 4G MME, SGW-C, and PGW-C: allocating IP addresses to UEs, NAS signalling for session management (SM), sending QoS and policy information to RAN via the AMF, downlink data notification, and selecting and controlling UPF for traffic routing. The UPF selection function enables Mobile Edge Computing (MEC) by selecting a UPF close to the edge of the network [23]. Based on characteristics of these functionalities and following the same reasons argued in [24–28], we assume that the SMF is also a CPU-intensive function and it needs a good networking interface as well.

Therefore, we will investigate for both AMF and SMF how much CPU cores are needed in order to serve a given number of users. Following the same reasons argued in [24–28], we assume that the most frequent procedures are the service request (SR), the service release request (SRR) and the X2-based handover (HO). Following the reasoning presented above, both AMF and SMF are CPU-intensive, which allows us to consider the CPU power as the main cost driver. Therefore, in the

[3] Sensitivity analysis on this factor will be performed in future work.

Table 3 Frequency of the considered control procedures

Procedures	Service request	Service release	X2-based HO	Source
Frequency 1: nb of events per user/s	0.00126	0.00126	0.00112	[28]
Frequency 2: nb of events per user/s	0.0045	0.0045	0.0012	[26]

Table 4 Number of instructions per user per procedure

Procedures	Service request	Service release	X2-based HO
Number of instructions per procedure	3,580,000	3,200,000	2,140,000

proposed model, only the CPU power will be calculated, from which we will derive the RAM and storage by proportionality.

As explained previously, the signalling traffic modelling for AMF and SMF requires the following inputs: the procedures' frequency, number of instructions per procedure message and the number of exchanged messages per procedure for both AMF and SMF. In the next subsections we detail each input and explain from where it was derived.

1. Procedures' frequency:

There are two different assumptions regarding the frequency of the considered procedures. On the one hand, assuming a user inactivity of 10 s, authors in [26] calculated by the mean of a mathematical framework the signalling rates per user equipment for each signalling procedures for the vMME as presented in Table 3 labelled with frequency 2. On the other hand, authors in [28] derived the SR, SRR and HO frequencies from an operational network measurement as presented in Table 3 with frequency 1.

2. Number of instructions per procedure message:

We assume that the number of run instructions for the different control messages are the same as in [24–28], as presented in Table 4.

3. Number of instructions per procedure for AMF and SMF:

In order to model the contribution of both AMF and SMF in the signalling traffic, we use the 3GPP specification document [21] describing the sequence diagrams for each control procedure, and from there we count the number of messages handled by AMF versus those handled by SMF for each procedure. We derive a ratio of AMF/SMF contribution in the exchanged messages for each procedure in order to be used to split the number of instructions per procedures of vMME between them. For each procedure, several assumptions were made in order to count these numbers:

Table 5 Number of messages for each procedure for both AMF and SMF

Procedures	Service request	Service release	X2-based HO
The ratio of AMF contribution in exchanged messages	0.533333333	0.6	0.5
The ratio of SMF contribution in exchanged messages	0.4	0.8	0.666666667

Service request:

- UE Triggered Service Request is considered.
- For this procedure, the impacted SMF and UPF are all under control of the PLMN serving the UE, e.g. in the home-routed roaming case, the SMF and UPF in HPLMN are not involved.
- We assume that the Service Request was sent integrity protected.
- The UE identifies list of PDU sessions to be activated in the Service Request message.
- We assume that PDU Session ID corresponds to a Local Area Data Network (LADN) and the SMF determines that the UE is within the area of availability of the LADN.
- We assume that the SMF accepts the activation of uplink connection and continue using the current UPF(s);
- No N4 Session Modification is established.
- No dynamic Policy and Charging control (PCC) is deployed.

Service release request:

- Service release procedure corresponds to AN Release in 5G CP procedures.
- We assume that 3 PDU sessions were active at the time of initiating this procedure. This assumption is based on the fact that each network slice will be served with a separate PDU session. Thus, assuming two different slices with a PDU session for each of them and a third session for the normal calls and chat. Since we do not have a good reference that strengthens this assumption, a sensitivity analysis on this is elaborated. Results demonstrates the number of active PDU sessions does not significantly affect the required CPU cores for AMF and SMF (see Appendix for more information).
- The procedure is triggered by the User Inactivity.

X2-based handover:

- It corresponds to Xn based inter NG-RAN handover in 5G CP procedures.

Table 6 Variation of the number of CPU cores required for AMF and SMF in function of the number of users for both frequency 1 and 2

Frequency/number of users	Network function	100×10^3	1000×10^3	$10{,}000\times10^3$
Frequency 1	AMF	0.4	4	40
	SMF	0.5	5	50
Frequency 2	AMF	1.23	12.3	123
	SMF	1.51	15.1	151
Average of frequencies	AMF	0.82	8.2	82
	SMF	1	10	100

The ratio of contribution of both AMF and SMF per procedure are presented in Table 5.

4.1.2 Mathematical Formulation

The number of instructions per user per second for AMF and SMF for each procedure is calculated using Eq. 4.

$$\text{Nb\_Inst(NF)}_{\text{prcd/user}} = \text{Nb\_Inst}_{\text{Prcd}} \times \text{Freq(prcd)}_{\text{user/s}} \times \left(\text{ratio\_Nb}_{\text{Msg}}\text{(NF)}_{\text{prcd}} \times \text{Corct}_{\text{factor}}\right) \tag{4}$$

The number of CPU core required to run the total instructions per procedure per network function (AMF or SMF) is calculated via applying Eq. 5.

$$\text{Nb\_CPU}_{\text{NF/prcd}} = {}^{\text{Nb\_Inst(NF)}_{\text{prcd/user}}}\!/_{\text{CPU}_{\text{power}}}. \tag{5}$$

The total CPU cores required for all the procedures per NF is derived using Eq. 6.

$$\text{Nb\_CPU}_{\text{NF}} = \sum_{\text{i=1}}^{3} \text{Nb\_CPU}_{\text{NF/prcd(i)}} \tag{6}$$

with $\text{Nb\_Inst(NF)}_{\text{prcd/user}}$ is the number of instructions per procedures per user per second for NF; $\text{Nb\_Inst}_{\text{prcd}}$ is the number of instructions per procedures; $\text{Corct}_{\text{factor}}$ is the the correction factor of the exchanged messages between the AMF/SMF and the rest of 5G NFs; $\text{Nb\_CPU}_{\text{NF/prcd}}$ is the number of CPU cores required for the procedure (prcd) within the NF; $\text{Freq(prcd)}_{\text{user/s}}$ is the the frequency of the procedure per user per second; $\text{ratio\_Nb}_{\text{Msg(NF)/pr}}$ is the ratio of NF contribution in the exchanged messages within procedure; $\text{CPU}_{\text{power}}$ is the the power of one CPU core (number of supported instructions per second); $\text{Nb\_CPU}_{\text{NF}}$ is the number of CPU cores required for NF.

4.1.3 Results of the Simulation

We simulated the proposed model to calculate the required CPU cores for both AMF and SMF and for the two frequencies while varying the number of users. We

consider the same CPU power as in [26], one CPU core has the power of 2.845×10^9 float operations per second. Results of the simulation are presented in Table 6.

Results show that, for AMF, for frequency 1, half of a CPU core is needed in order to serve 100 k users, yet for frequency 2, almost 1.3 CPU cores are required to satisfy the signalling traffic of the same number of users and this is. For SMF, similar results are found regarding the difference between results for the two frequencies. The difference between the two frequencies for a small number of users is not that significant, however if we increase the number of users drastically, the gap between the results of these frequencies becomes significant.

Therefore, and since only those two frequencies are found in literature, and since we could not find a reason to prefer one of them, we decided to take the average of the two frequencies (see third row in the table above). Hence, only one core is needed to serve 100 k users for both AMF and SMF and 5 cores are needed to serve 1 million users.

We can conclude from Table 6 that both the number of CPU cores needed for AMF and SMF scales linearly with the number of users:

$$Nb\_CPU_{AMF} = 8.2 \times 10^{-4} \times N_{users}$$

$$Nb\_CPU_{SMF} = 10 \times 10^{-4} \times N_{users}$$

with Nb_CPU is the number CPU cores required for the NF; N_{users} is the number of users.

4.1.4 The Rest of the Control Plane Functions

The remaining control plane functions are the Policy Control Function (PCF), NF Repository function (NRF), Network Exposure function (NEF), Unified Data Management (UDM), Authentication Server Function (AUSF), Network Slice Selection Function (NSSF) and Application Function (AF) [23]. Since there is no starting point in literature to dimension those network functions, we analyse their functionalities and dimension them accordingly.

- The PCF consists of unified policy framework delivering policy rules to control plane functions and has access to subscriber information for policy decisions. Thus, it does not require significant CPU power.
- The AUSF acts as an authentication server, thus it requires a good processing power to elaborate the hashing tasks and the integrity checking.
- The UDM needs first a good memory and storage since it handles subscription management, user identification and it requires enough processing power to support the generation of Authentication and Key Agreement (AKA) credentials and access authorization.

NEF offers the exposure of capabilities and events and assures the security provisioning of information from non-3GPP application to 3GPP network, we have to think about how much external application needs to communicate with the 3GPP

Service chains considered for each service and bandwidth requirements.

Service	Chained VNFs *	Bandwidth req.
VoIP	NAT-FW-TM-FW-NAT	250 Kbps
Video Conference	NAT-FW-TM-VOC-IDPS	2 Mbps
Web Service	NAT-FW-TM-WOC-IDPS	4 Mbps

* IDPS: *Intrusion Detection Prevention*, FW: *Firewall*, NAT: *Network Address Translation*, TM: *Traffic Monitor*, VOC: *Video Optimization Controller*, WOC: *WAN Optimization Controller*

Fig. 5 Service chains and bandwidth requirements [32]

network but as a first order of estimate we can assume that it does not happen that often thus a modest CPU and storage resources are sufficient.

- The same holds for the NSSF, which supports the selection of the network slice instances to serve the UE and the correspondent AMF, which only happens within the service request procedure.
- For the NRF, which maintains NF profile and instances, we can assume that it requires a good processing power and memory as well.

4.2 5G Data Plane Modelling

Many researchers model the data plane function (i.e. UPF) for network slicing within the LTE-A and 5G networks using virtualized middlebox network functions [30–36]. Middlebox network functions here refer to Network Address Translator (NAT), Firewall (FW), WAN Optimization Controller (WOC), Intrusion Detection Prevention System (IDPS), Video Optimization Controller (VOC) and Traffic Monitor (TM). Authors of these papers have identified a chain of VNFs (which are virtualized middleboxes) per service with specific requirements in terms of bandwidth, latency and hardware per VNF. The chain of network functions per service is presented in Fig. 5. The main goals of these papers are, first, to find the best placement of the VNFs in order to optimize the network performance (e.g. minimize the latency) and, second, to investigate the number of active VNF nodes in specific scenarios in order to optimize the resource dimensioning.

Three main approaches are adopted in literature to dimension the UPF using virtualized middlebox functions. The first method translates the total traffic handled by VNFs to the number of concurrent operations for each VNF and deduce from that the needed hardware requirements such as in [32]. Yet, the second approach derives from the middlebox datasheets the required processing per user and given the number of users they deduce the needed CPU and others hardware requirements for each VNF [33]. However, the third approach defines a CPU-core-to-throughput relationship for each VNF and uses it to calculate the required number of CPU cores per VNF, like the method established in [35].

Table 7 Chain of NFs the total generated traffic per service for the considered scenario

Slice	Chain of VNFS	Throughput Per user	% traffic	Total traffic (for 1050 users)
VoIP	NAT-FW-TM-FW-NAT	64 kbps	20	13.44 Mbps
Video	NAT-FW-TM-VOC-IDPS	4 Mbps	80	1.68 Gbps

Since those dimensioning methods use different approaches and different assumptions, we will apply them to a specific use case (see Sect. 5.2), compare their outcomes and adopt the option that gives the most realistic results that align with other findings in literature.

5 Application of the Cost Allocation Model to SaT5G Use Case

As discussed previously, the main two goals of the developed cost allocation model are, first, to fairly allocate the network costs to the deployed slices and, second, to investigate the cost savings introduced by network slicing if any. In order to achieve these goals, two auxiliary models are developed to feed the main model with the necessary inputs. These auxiliary models are the identification of the hardware requirements of 5G NFs and the slice resources allocation. Therefore, in this section, results of these models as well as of the main model are discussed sequentially starting with the description of the specific scenario to which these models are applied. This description is elaborated in Sect. 5.1. Section 5.2 summarizes the main outcomes of the 5G network function modelling algorithm. Section 5.3 presents the resource allocation algorithm's results for the considered slices. Those results feed the cost allocation model, which results are discussed in Sect. 5.4. Afterwards, the cost savings on the core network side resulting from the use of NFV concept and network slicing are presented in Sect. 5.5.

5.1 Description of the Scenario

The scenario used in this paper was defined within the SaT5G project [37]. It considers providing broadband connectivity and 5G services to rural areas where currently no terrestrial network infrastructure is deployed. In order to reach those remote areas, satellite communication is used as a backhaul network that links the 5G core network to the radio access network. This latter consists of many base stations that will be installed in the considered area, which covers two villages about 5 km apart connected via a rural main road. The villages are home to 350 families, with an average of 3 users per home, resulting in 1050 users. The full description of the scenario as well as the network architecture installed are detailed in [17]. The 5G services considered here are eMBB voice and eMBB video. The quality of service of these two 5G services is described in [18].

Table 8 Hardware requirements per VNFs per service for the three approaches

	Number of CPU cores approach 1	Number of CPU cores approach 2	Number of CPU cores from approach 3
NAT	0.35	0.966 ~ = 1	1
FW	0.7	0.945 ~ = 1	1
TM	0.35	13.965 = 14	2
VOC	0.7	5.67 = 6	1
WOC	0.35	5.67 = 6	1
IDPS	0.7	11.235 = 12	2

Table 9 Hardware requirements of the data plane network functions for eMBB voice and video slices

VNF/service	Voice, throughput = 13.44 Mbps			Video, throughput = 1.68 Gbps		
	CPU	RAM: GB	HDD: GB	CPU	RAM: GB	HDD: GB
NAT	2	2	4	1	1	2
FW	2	4	6	2	3	5
TM	1	3	2	2	6	4
VOC	–	–	–	1	1	10
WOC	–	–	–	–	–	–
IDPS	–	–	–	2	2	7

5.2 5G NF Modelling for the Specific Scenario

The chain of network functions for the data plane and the required throughput per user for each service are derived based on input from literature [33, 34], and they are presented in Table 7. According to the Cisco VNI Forecast Highlights Tool, Internet video traffic will be 80% of all consumer Internet traffic by 2022 compared to 73% in 2017 [38], hence we assume that the video traffic represents 80% of the overall traffic and 20% of the traffic is voice. The total generated traffic per service is calculated based on the number of served users and is presented in Table 7.

For the data plane network functions, we applied the three approaches described in Sect. 4.2 and we compare their results to the findings in literature. Results of the model are presented in Table 8. The first column presents the required CPU cores for the different virtualized middleboxes based on the number of concurrent operations. The number of concurrent operations that each function needs to handle is derived by linking the original findings in [34] with the carried traffic in terms of Mbps. Results generated based on the second approach use the processing requirement per user considering the carried traffic as well. Finally, results of the third approach are driven using the CPU-core-to-throughput relationship. The comparison of our findings to those in literature [30–36] allows to conclude that the number of required CPU based on the throughput-to-CPU relationship (approach 3) seems to be the more realistic one.

Adopting the CPU-core-to-throughput approach, the hardware requirements for both eMBB voice and video slices are summarized in Table 9.

For the control plane network functions, we applied the proposed model described in Sect. 4.1 assuming a user inactivity of 10 s, the mobility model adopted in [26] and 1050 users. Based on the 3GPP view on network slicing, several NFs of the control plane are shared between slices. These common network functions are the AMF, NRF, NEF, UDM, AUSF and NSSF [39, 40]. Those functions are represented with grey columns in Table 10. Moreover, since the video slice is carried over the satellite link, we need to deploy a satellite multicast function and a Nano-CDN node next to the edge for popular video's caching purpose.

Results of the of the model for the two slices being eMBB voice and video are recapitulated in Table 10.

It is clear from results presented in Table 10 that, similar to the RAN, the video slice requires more core network resources than the voice slice, for example, the UPF of the video slice requires 8 CPU cores against only 5 CPU cores for the voice slice.

5.3 Resource Allocation for Video and Voice Slices

Given the hardware requirements of both voice and video slices presented in the previous section, we applied the proposed allocation model. The model is based on the preconfigured VMs (presented in Table 2 and described in Sect. 3.2). The results of the required VM types for each slice as well as for the shared network functions are presented in Table 11.

Similar to [14, 41], we consider a Blade server due to its capability of providing more processing power in less space which allow to simplify cabling and storage. The Blade server consists of 8CPU, 64 GB RAM, 1000 GB HDD and 4 Ethernet cards of 10 Gbps each. Moreover, for the reliability of the network, we consider a redundant VM for each NF. Hence, within the preconfigured VM option, we need 10 servers to satisfy the required hardware requirements of the two slices. Orthogonally, within the configurable VM option discussed in Sect. 3.2, we need only 8 servers from the same server type.

These findings are used as an input to the cost model of the virtualized core network presented in Sect. 5.5 in order to deduce the cost saving resulting from the use of network slicing.

On the other hand, we argued previously in Sect. 4.1.1, that the SMF needs a good networking interface. Thus, to investigate if we have a limitation in term of networking interface for the Blade servers reserved based on the CPU metric, we calculate the generated traffic per user for SMF and AMF as well. We assume that the average packet size is 250 Bytes for the control messages generated by SMF and AMF (similar to the assumption for the vMME in [26]). The calculation results in 18.418 and 14.626 bits per second per user for the SMF and AMF respectively. Hence, for our scenario with 1050 users, the AMF requires 15.4 kbps to be reserved on the network interface, while the SMF requires 38.8 kbps (for both video and voice slices). Therefore, 1Gbps is allocated on the network interface card (NIC) for

Table 10 Hardware requirements for the eMBB voice and video slices

	Technical requirement	UPF	AF	AMF	SMF	PCF	NRF	NEF	UDM	AUSF	NSSF	Satellite multi-cast function	AF (CDN)
eMBB voice	CPU	5	1	1	1	0.5	1	1	2	2	1		
	Memory	9	2	1	1	1	1	1	4	4	1		
	HDD	12	2	2	2	2	2	2	10	10	2		
eMBB video	CPU	8	2		1	0.5						1	1
	Memory	13	4		1	1						2	2
	HDD	28	4		2	2						100	100

Table 11 Results of mapping VNFs onto VMs

VM type	S-VM	M-VM	L-VM
Shared Nfs	4	2	0
Video slice	4	1	1
Voice slice	3	0	1

Table 12 Cost inputs for the cost allocation model

Item	Value	References
RAN cost (in euro)	50,067	[17]
Cost of Satellite backhaul (in euro)	160,149	[17]
Cost of 5G core network (in euro)	76,230	[17]
latency eMBB video in ms	300	latency: Packet Delay Budget from "3GPP standard QoS class identifiers" for Non-Conversational Video (Buffered Streaming) [18]
latency eMBB voice in ms	100	latency: Packet Delay Budget from "3GPP standard QoS class identifiers" for Conversational Voice [18]
Throughput eMBB video (in Mbps)	4	Assumption
Throughput eMBB voice (in Mbps)	0.64	Assumption
Cost sharing coefficient	0.7	Assumption
Qco Video	0.718	Calculated based on Eq. 3
Qco Voice	0.282	Calculated based on Eq. 3
CPU cost driver: video	0.59	Calculated using the model
CPU cost driver: voice	0.41	Calculated using the model

each VNF, it is more than enough. Hence, we do not dispose of a limitation on the networking interface resources.

5.4 Cost Allocation for Video and Voice Slices

The cost of the network infrastructure installed to serve the remote two villages described previously with eMBB voice and video services are taken from [17]. These cost figures and the required inputs for the cost allocation model are detailed in Table 12.

As described in Sect. 3, the CPU metric will be used as a cost driver to split the total network costs between the running services. This cost driver for to the two slices being eMBB video and eMBB voice is calculated using the outcomes of the hardware requirements identification model (discussed in Sect. 5.2).

The calculation of the CPU cost driver for both eMBB video and voice is presented in Table 12. It shows that 59% of the core network cost will be allocated to the eMBB video slice versus 41% for the voice slice. For the RAN side, the throughput

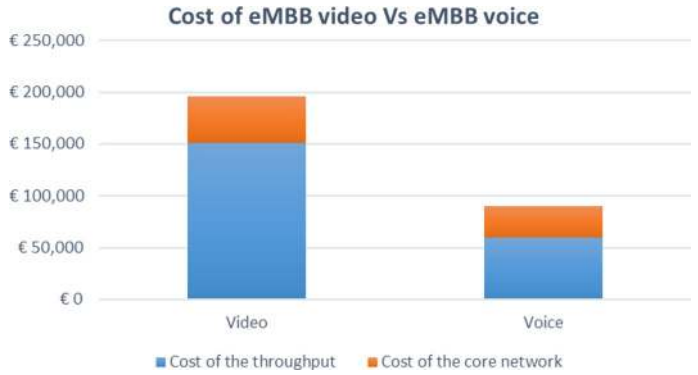

Fig. 6 allocation of the network cost to eMBB video and voice slices

and the quality of service (QoS) coefficient used to allocate the RAN cost to the two slices are also presented in Table 12. The QoS coefficient of the eMBB video is higher than the one of the eMBB voice (0.7 versus 0.3) since the capacity required by the video service is much more important than the voice' capacity and since also the cost sharing coefficient (0.7) is in favour of the throughput at the expense of the latency.

Considering input values presented in Table 12, we run our model to allocate the cost of the network. Results of the model are represented in Fig. 6.

Several interpretations can be extracted from Fig. 6. First, the eMBB video slice bears the significant amount of the network costs (68% compared to only 32% for the voice slice). This is reasonable, because it requires more throughput on the RAN and backhaul and more computing resources on the core network part as well. However, the cost of the throughput of the voice slice is still important, which can be justified with the fact that not only the required throughput is counted in the cost allocation algorithm but also the latency (Eq. 3), for which the voice service presents a more stringent requirement.

5.5 Cost Savings Resulting from the Use of NFV in the Core Network

NFV and network slicing paradigms promise to reduce overall network costs and allow for more cost-efficient network deployment. In this section, we investigate these promises. Given the dimensioning model for network slicing presented in Sect. 5.3, we know the exact amount of resources that each slice (from the two slices eMBB voice and video) requires. Furthermore, we count redundant resources for reliability purposes. Following these inputs, the 5G core network can be dimensioned accordingly. Afterwards, calculating the cost of the core network resources and comparing it to the cost of the traditional core network (without NFV deployment) allows to calculate the cost saving of the use of virtualization technology. Figure 7 presents the flow between the different sub-models that feeds the cost saving calculation.

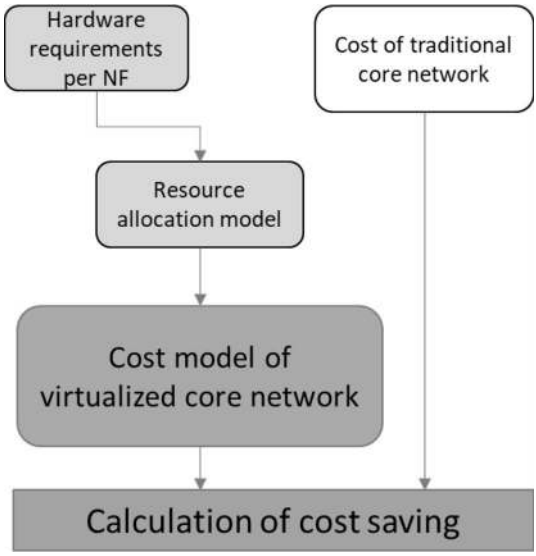

Fig. 7 Calculation of the virtualization cost saving for the core network

Table 13 Cost reduction resulting from the use of NFV in the core network

	Cost virtualized 5G core	Cost of the traditional core network	Cost reduction (%)
Cost 5G core per Mbps	21€	38 euro per Mbps (5 to 6 euro per Mbps for the hardware and 30 to 35 euro per Mbps for software) [42]	44.73
Cost 5G core per user	53€	60 euro per user [17]	11.67

In Appendix, we detail the cost inputs used to derive the cost of the virtualized 5G core network. We have two figures for the traditional core network cost, the first one is quoted in terms of the number of users, the second in terms of traffic, i.e. Mbps. Therefore, we calculate the cost of the core network and we quote it also per user and per Mbps.

The cost saving due to the use of virtualization on the core network is about 45% if the cost is quoted in Mbps compared to the cost used in [42], yet, it is only 12% if the cost is calculated per user compared to the cost found in [17]. The result of the comparison between the core network cost with NFV deployment versus without, is summarized in Table 13. Despite this difference, the savings can be identified as significant, and prove that the use of NFV and network slicing reduces the cost of the core network deployment.

6 Conclusions and Future Work

This paper proposed a novel cost allocation model for sliced networks. The model aims, first, to allocate the network cost to the deployed slices and, second, to investigate the cost reduction promises of using the NFV and network slicing technologies. Furthermore, to feed the proposed model with the required inputs and since we did not find these inputs in literature, we developed two other models. The first one is a resource allocation model that assigns the right amount of network resources to each slice aiming at fulfilling its KPI requirements, for which we needed to derive the hardware requirements of each VNF that is included in the chain of VNFs forming the slice. Therefore, we designed the second model that identified the hardware requirements of the 5G network functions, with a focus on those of the control plane. This latter is also the first of its kind in literature because, first, the control plane traffic was before often considered as a percentage of the data plane traffic and, second, because it is the first model for 5G networks. We applied the proposed model to a specific scenario considering the offer of two slices (eMBB video and voice) to a remote area. Results of the model shows that the eMBB video bears the significant part of the network cost. On the other hand, results demonstrate a cost saving of 12 to 45% (depending on the approach adopted) of the core network costs resulting from the use of NFV and network slicing. As next steps, we plan to validate our models and especially the hardware requirements of 5G NFs with, first, more scenarios, such as dense-urban scenarios, wherein higher traffic is required per slice and, second, with the testbed results within the SaT5G project. It would be a good exercise also to apply the model on two completely different types of slices (e.g. IoT versus eMBB) in order to investigate the impact of more diverse KPI requirements on the cost. Up till now, this was unfortunately not possible because of a lack of traffic model and scenario details for an IoT scenario. Finally, the traffic considered here is static, future work should extend the model to the case where the dynamicity of the traffic is considered. The same applies for the resource allocation model, as currently the assignment of the network resources is done statically based on the maximum throughput and traffic carried by the different network functions.

Acknowledgements The authors would like to acknowledge all partners in the SaT5G project, funded by the European Union's Horizon 2020 research and innovation program under grant Agreement No 761413.

Appendix

1. Inputs of the cost model of the NFV-based core network:

   See Table 14.

2. Variation of the number of PDU sessions:

Table 14 Cost inputs and calculation for the cost saving model of network virtualization

Item	Value	References
Blade server price	1300	Average price in the internet
Number of servers	8	We only need 4 but 8 in total for redundancy purposes
Capex Data Centre	12,753	
Air conditioning	500	[17]
Installation	1912.95	
Maintenance*5 year	6376.5	
Power consumption	240	Power consumption per server per year in kWh: from the datasheet
Electricity price kwh	0.014	[17]
cost power consumption	134.4	
VNF license cost (Dollar/vCPU): 100	25,200	[43]
Nokia router 7750	1853	[42]: 20,383 euro for 11 small sites (like our scenario site)
Total Capex	14,666	
Total Opex	31,711	
Overhead costs	9275.37	
TCO	55,652.22	
Total traffic (Mbps)	2700	

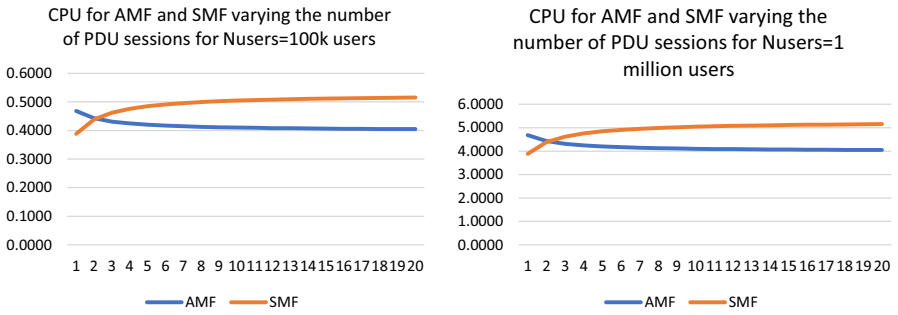

Fig. 8 Number of CPU cores for AMF and SMF in function of number of users for different number of PDU sessions

We assumed before in modelling the SRR procedure that 3 PDU sessions were active when this procedure is launched, but we do not have a strong assumption on this value. Hence in Fig. 8, we vary the number of PDU sessions to visualize its effect on the required CPU cores for both AMF and SMF using the average frequency without the correction factor. From these results we can deduce, up to until 1 million users, that the number of active PDU sessions does not significantly affect the required CPU cores for AMF and SMF.

References

1. Wood, T., Ramakrishnan, K.K., Hwang, J., Liu, G., Zhang, W.: Toward a software-based network: integrating software defined networking and network function virtualization. IEEE Netw. **29**(3), 36–41 (2015)
2. TSI NFV.: Network functions virtualization: An introduction, benefits, enablers, challenges & call for action. Darmstadt, Germany, SDN & OpenFlow World Congress, White Paper, Oct. 2012
3. Sun, S., Kadoch, M., Gong, L., Rong, B.: Integrating network function virtualization with SDR and SDN for 4G/5G networks. IEEE Netw. **29**(3), 54–59 (2015)
4. Network Slicing and 3GPP Service and Systems Aspects (SA) Standard.: https://sdn.ieee.org/newsletter/december-2017/network-slicing-and-3gpp-service-and-systems-aspects-sa-standard. Accessed 15 May 2019
5. Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A., Flinck, H.: Network slicing and softwarization: a survey on principles, enabling technologies, and solutions. IEEE Commun. Surv. Tutor. **20**(3), 2429–2453 (2018)
6. Herrera, J.G., Botero, J.F.: Resource allocation in NFV: a comprehensive survey. IEEE Trans. Netw. Serv. Manage. **13**(3), 518–532 (2016)
7. Zhang, H., Liu, N., Chu, X., Long, K., Aghvami, A.H., Leung, V.C.: Network slicing based 5G and future mobile networks: mobility, resource management, and challenges. IEEE Commun. Mag. **55**(8), 138–145 (2017)
8. Caballero, P., Banchs, A., De Veciana, G., Costa-Pérez, X., Azcorra, A.: Network slicing for guaranteed rate services: admission control and resource allocation games. IEEE Trans. Wireless Commun. **17**(10), 6419–6432 (2018)
9. Checko, A., Christiansen, H.L., Yan, Y., Scolari, L., Kardaras, G., Berger, M.S., Dittmann, L.: Cloud RAN for mobile networks—a technology overview. IEEE Commun. Surv. Tutor. **17**(1), 405–426 (2014)
10. Sciancalepore, V., Samdanis, K., Costa-Perez, X., Bega, D., Gramaglia, M., Banchs, A.: Mobile traffic forecasting for maximizing 5G network slicing resource utilization. In: IEEE INFOCOM 2017-IEEE conference on computer communications, pp. 1–9. IEEE, New York (2017)
11. Naudts, B., Kind, M., Westphal, F. J., Verbrugge, S., Colle, D., Pickavet, M.: Techno-economic analysis of software defined networking as architecture for the virtualiazation of a mobile network. In: European Workshop on Software Defined Networking (EWSDN-2012), pp. 1–6 (2012)
12. Bouras, C., Ntarzanos, P., Papazois, A.: Cost modeling for SDN/NFV based mobile 5G networks. In: 2016 8th international congress on ultra modern telecommunications and control systems and workshops (ICUMT), pp. 56–61. IEEE, New York (2016)
13. Zhang, N., Hämmäinen, H.: Cost efficiency of SDN in LTE-based mobile networks: Case Finland. In: 2015 international conference and workshops on networked systems (NetSys), pp. 1–5. IEEE, New York (2015)
14. Knoll, T. M.: Life-cycle cost modelling for NFV/SDN based mobile networks. In: 2015 conference of telecommunication, media and internet techno-economics (CTTE), pp. 1–8. IEEE, New York (2015)
15. Han, B., Tayade, S., Schotten, H. D.: Modeling profit of sliced 5G networks for advanced network resource management and slice implementation. In: 2017 IEEE symposium on computers and communications (ISCC), pp. 576–581. IEEE, New York (2017)
16. Zhou, X., Li, R., Chen, T., Zhang, H.: Network slicing as a service: enabling enterprises' own software-defined cellular networks. IEEE Commun. Mag. **54**(7), 146–153 (2016)
17. Chiha, A., Van der Wee, M., Colle, D., Verbrugge, S.: Techno-economic viability of integrating satellite communication in 4G networks to bridge the broadband digital divide. Telecommun. Policy (2019). https://doi.org/10.1016/j.telpol.2019.101874
18. Policy and charging control architecture, 3GPP specification 3.203.: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=810. Accessed 10 June 2018
19. Wu, G., Tang, M., Tian, Y. C., Li, W.: Energy-efficient virtual machine placement in data centers by genetic algorithm. In: International conference on neural information processing, pp. 315–323. Springer, Berlin (2012)
20. Ahmad, R.W., Gani, A., Hamid, S.H.A., Shiraz, M., Yousafzai, A., Xia, F.: A survey on virtual machine migration and server consolidation frameworks for cloud data centers. J. Netw. Comput. Appl. **52**, 11–25 (2015)

21. G; Procedures for the 5G System (3GPP TS 23.502 version 15.2.0 Release 15).: https://www.etsi.org/deliver/etsi_ts/123500_123599/123502/15.02.00_60/ts_123502v150200p.pdf. Accessed 10 Jan 2019
22. Sun, K., Kim, Y.: Gap analysis for adapting the distributed mobility management model in 4G/5G mobile networks. In: 2017 IEEE conference on network softwarization (NetSoft), pp. 1–5. IEEE, New York (2017)
23. Grandmetric.: 5G Core Network Functions. https://www.grandmetric.com/2018/03/02/5g-core-network-functions. Accessed 10 Jan 2019
24. Prados, J., Laghrissi, A., Bagaa, M., Taleb, T., Lopez-Soler, J.M.: A complete LTE mathematical framework for the network slice planning of the EPC. IEEE Trans. Mobile Comput. **19**, 1–4 (2019)
25. Prados-Garzon, J., Ramos-Munoz, J. J., Ameigeiras, P., Andres-Maldonado, P., Lopez-Soler, J. M.: Latency evaluation of a virtualized MME. In: 2016 Wireless Days (WD), pp. 1–3. IEEE, New York (2016)
26. Prados-Garzon, J., Ameigeiras, P., Ramos-Munoz, J. J., Andres-Maldonado, P., Lopez-Soler, J. M.: Analytical modeling for virtualized network functions. In: 2017 IEEE international conference on communications workshops (ICC Workshops), pp. 979–985. IEEE, New York (2017)
27. Prados-Garzon, J., Ramos-Munoz, J.J., Ameigeiras, P., Andres-Maldonado, P., Lopez-Soler, J.M.: Modeling and dimensioning of a virtualized MME for 5G mobile networks. IEEE Trans. Veh. Technol. **66**(5), 4383–4395 (2017)
28. Hirschman, B., Mehta, P., Ramia, K.B., Rajan, A.S., Dylag, E., Singh, A., McDonald, M.: High-performance evolved packet core signaling and bearer processing on general-purpose processors. IEEE Netw. **29**(3), 6–14 (2015)
29. Buyakar, T. V. K., Agarwal, H., Tamma, B. R.: Prototyping and load balancing the service based architecture of 5G core using NFV. In: 2019 IEEE Conference on Network Softwarization (NetSoft), pp. 228–232. IEEE, New York (2019)
30. Savi, M., Hmaity, A., Verticale, G., Höst, S., Tornatore, M.: To distribute or not to distribute? Impact of latency on virtual network function distribution at the edge of FMC networks. In: 2016 18th international conference on transparent optical networks (ICTON), pp. 1–4. IEEE, New Yrok (2016)
31. Ruiz, L., Durán, R. J., Miguel, I. D., Khodashenas, P. S., Pedreno-Manresa, J.-J., Merayo, N., Aguado, J. C., Pavon-Marino, P., Siddiqui, S., Mata, J., Fernández, P., Lorenzo, R. M., Abril, E. J.: Genetic algorithm for effective service mapping in the optical backhaul of 5G networks. In: 20th international conference on transparent optical networks (ICTON) (2018)
32. Pedreno-Manresa, J. J., Khodashenas, P. S., Siddiqui, M. S., Pavon-Marino, P. Dynamic QoS/QoE assurance in realistic NFV-enabled 5G access networks. In: 2017 19th international conference on transparent optical networks (ICTON), pp. 1–4. IEEE, New York (2017)
33. Savi, M., Tornatore, M., Verticale, G.: Impact of processing-resource sharing on the placement of chained virtual network functions. IEEE Trans. Cloud Comput. 2019. https://doi.org/10.1109/TCC.2019.2914387
34. Savi, M., Tornatore, M., Verticale, G.: Impact of processing costs on service chain placement in network functions virtualization. In: 2015 IEEE conference on network function virtualization and software defined network (NFV-SDN), pp. 191–197. IEEE, New York (2015)
35. Gupta, A., Habib, M.F., Mandal, U., Chowdhury, P., Tornatore, M., Mukherjee, B.: On service-chaining strategies using virtual network functions in operator networks. Comput. Netw. **133**, 1–16 (2018)
36. Gupta, A., Jaumard, B., Tornatore, M., Mukherjee, B.: A scalable approach for service Chain mapping with multiple SC instances in a wide-area network. IEEE J. Sel. Areas Commun. **36**(3), 529–541 (2018)
37. Liolis, K., Geurtz, A., Sperber, R., Schulz, D., Watts, S., Poziopoulou, G., et al.: Use cases and scenarios of 5G integrated satellite-terrestrial networks for enhanced mobile broadband: the SaT5G approach. Int. J. Satell. Commun. Netw. **37**(2), 91–112 (2019)
38. CISCO VNI: VNI Forecast Highlights Tool. https://www.cisco.com/c/m/en_us/solutions/service-provider/vni-forecast-highlights.html#. Accessed 8 Mar 2019
39. Kiran, M., Jun, Q., Ravi, R., Liang, G., Li, Q., Shuping, P., et al.: IETF Report: Network Slicing Use Cases: Network Customization and Differentiated Services. https://tools.ietf.org/id/draft-netslices-usecases-02.html
40. IETF: Network Slicing–3GPP Use Case. IETF. https://tools.ietf.org/id/draft-defoy-netslices-3gpp-network-slicing-02.html. Accessed 5 Jan 2019

41. Economou, D., Rivoire, S., Kozyrakis, C., Ranganathan, P.: Full-system power analysis and modeling for server environments. In: International symposium on computer architecture-IEEE (2006)
42. SAT5G: Internal document (2017)
43. Dieye, M., Ahvar, S., Sahoo, J., Ahvar, E., Glitho, R., Elbiaze, H., Crespi, N.: CPVNF: cost-efficient proactive VNF placement and chaining for value-added services in content delivery networks. IEEE Trans. Netw. Serv. Manage. **15**(2), 774–786 (2018)

Asma Chiha received an M.Sc. degree in Telecommunication Engineering from The Higher School of Communication of Tunis SUP'Com in June 2010. In October 2017, she joined the Techno-Economic research unit at the Internet and Data Lab research group at Ghent university-imec. She is currently working towards a Ph.D. in Engineering in the scope of techno-economic impacts of network virtualization. Her research interests include 5G networks, network virtualization, network slicing, satellite communication and business modelling.

Marlies Van der Wee received an M.Sc. degree in Engineering, option Industrial Engineering and Operations Research from Ghent University in July 2010. She joined the Techno-Economic research unit at IBCN in September 2010, at the same university. She finished a Ph.D. on social cost–benefit analysis of broadband networks in a multi-actor setting. Marlies is author or co-author of more than 35 publications, both in international journals and presented at conferences world-wide. Marlies is involved in several national and European research projects, where she performs techno-economic analysis and business modelling in different application domains (such as broadband networks, media, energy and eHealth).

Didier Colle received a M.Sc. degree in electrical engineering (option: communications) from the Ghent University in 1997. Since then, he has been working at the same university as researcher in the department of Information Technology (INTEC). He is part of the research group INTEC Broadband Communication Networks (IBCN). His research lead to a Ph.D. degree in February 2002. He was granted a postdoctoral scholarship from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) in the period 2003-2004. Currently, he is co-responsible for the research cluster on network modelling, design and evaluation and is coordinating the research on fixed internet architectures and optical networks. His research deals with design and planning of communication networks. This work is focusing on optical transport networks, to support the next-generation Internet. Up till now, he has actively been involved in several national and international research projects. His work has been published in more than 250 scientific publications in international conferences and journals.

Sofie Verbrugge received an M.Sc. degree in computer science engineering from Ghent University (Ghent, Belgium) in 2001. She obtained the Ph.D. degree from the same university in 2007 for her thesis entitled "Strategic planning of optical telecommunication networks in a dynamic and uncertain environment". Since 2008, she has been working as a researcher affiliated to imec (previously iMinds), where she is the research coordinator for the techno-economic research group within the Internet and Data Lab (IDLab). Since October 2014 she is appointed as an associate professor in the field of techno-economics at Ghent University. She teaches courses on Information Technology and Data Processing as well as on Engineering Economy. Sofie's main research interests include technology selection, infrastructure as well as operational cost modeling, advanced evaluation techniques as well as business modeling and value network analysis.