Topology awareness for smart 5G eMBB network slicing VNF placement
Sánchez-Navarro, Ignacio; Alcaraz Calero, Jose M.; Wang, Qi

Published in:
2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)

DOI:
10.1109/WoWMoM49955.2020.00076

Published: 09/10/2020

Document Version
Peer reviewed version

Link to publication on the UWS Academic Portal

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the UWS Academic Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact pure@uws.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

“© © 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”
Topography Awareness for Smart 5G eMBB Network Slicing VNF Placement

1st Ignacio Sanchez-Navarro  
School of Computing Engineering and Physical Sciences  
University of the West of Scotland  
Paisley, United Kingdom  
Ignacio.Sanchez@uws.ac.uk

2nd Jose M. Alcaraz Calero  
School of Computing Engineering and Physical Sciences  
University of the West of Scotland  
Paisley, United Kingdom  
Jose.Alcaraz-Calero@uws.ac.uk

3rd Qi Wang  
School of Computing Engineering and Physical Sciences  
University of the West of Scotland  
Paisley, United Kingdom  
Qi.Wang@uws.ac.uk

Abstract—This paper presents an architecture to gather non-traditional metrics from 5G multi-tenant infrastructure using information about the network topology to take smart decisions on where to optimal placement VNFs that are used to provide services to network slices. The metrics considered are spatial metrics, where information about the shape and size of the network topology is taken into consideration. The architecture has been prototypical validated showing how optimal decisions are taken in an eMBB high-dense scenario, with topologies up to 65538 mobile users geographically concentrated on the same location. Our prototype is able to deal with the calculation of such spatial metrics over the 5G multi-tenant network with 65538 mobile users within 20 seconds, which make it viable at operational phase.

Index Terms—VNF Placement, Smart Computing, 5G, Network Topology, Network Slicing, Multy-Teancy

I. INTRODUCTION

The imminent Fifth-Generation (5G) mobile networks are intrinsically linked to virtualisation and containerisation. This introduces significant costs reduction making usage of softwarisation and key enablers such as Software Defined Networks (SDN) or Network Functions Virtualisation (NFV). SDN pursue the separation of network control functions from network forwarding functions whereas NFV seeks the execution of network functions in virtualised environments. A large-scale 5G multi-tenant network will require a significant amount of different Virtual Network Functions (VNF) to provide support for all the tenants of the infrastructure. The efficient management and orchestration of such large-scale amount of VNFs represents a significant challenge to be addressed.

VNFs are network functions that can deliver services to network slices by using Service Function Chain (SFC), which is a set of ordered network functions providing specific features to the network flows that compose such network slice. The placement of such VNFs along any possible location in the infrastructure will have a significant impact in the efficient management and orchestration of resources to optimise the performance of the delivery of such network services. A wrong choice for the placement of a VNF could incur in extra delays, packet losses, lack of scalability, possible resource shortages, etcetera, thus it is of critical importance to make an efficient decision on the allocation of such VNFs.

This is specially relevant in the context of 5G multi-tenant networks where the ambitious requirements imposed by the 5G use cases such as Enhanced Mobile Broadband (eMBB), Ultra-reliable Low Latency Communications (URLLC) and Massive Machine Type Communications (mMTC) will stress even more the importance of the optimal placement of VNFs.

There exists a number of research works focused on improving the placement of VNF. However, almost the vast majority of them only consider metrics and capabilities related to the state of the physical machines (PMs) and the requirements imposed by the VNFs to take decisions on their placement. Although these works represent a first step on taking informed decisions, they achieve results far to be optimal, mainly due to the fact that they completely ignores information about the shape and size of the network topology and the connectivity among the different physical and virtual VNFs that compose the 5G multi-tenant infrastructure.

The main aim of this contribution is the design and implementation of a novel architecture able to provide structural metrics that are used to achieve a smart VNF placement. To be specific, the main innovations achieved in this architecture are as follows:

1) New spatial metrics to allow metric-aware VNF placement optimisation
2) New topology awareness capabilities to allow end-to-end network slicing optimisation
3) Architecture compatible with multi-tenancy support for multi-operator operational environments

The remainder of the paper is structured as follows. Section II introduces and compares different solutions in the state of the art about VNFs placement strategies. Section III introduces the proposed architecture to achieve topology awareness metrics. Section IV defines the 5G-aware spatial metrics proposed to be used on VNF placement decisions. Then, section V describes the proposed VNF placement strategy using the metrics previously introduced tailored for specific use cases. Afterwards, Section VII presents the performance evaluation of the empirical results obtained. Finally, Section VIII concludes this paper.

II. RELATED WORK

Key multi-tenant virtualisation technologies such as OpenStack [1] and Kubernetes [2] still making use of a random VNF placement strategy where clearly is impossible to take informed decisions.

Askari et al [3] presents a comprehensive analysis of different VNF placement strategies for Service Chain (SC) provisioning. They conclude that an efficient placement strategy can reduce the operational cost in terms of delay, QoE and economy of service provisioning up to 16% or 23%. This clearly emphasises the importance of taking an optimal decision.

Oechsner and Ripke [4] propose a VNF placement strategy integrated with OpenStack where the infrastructure is modeled as a tree where each node represents a computer which the associated metrics of delay and availability and where such metrics are used to take optimum decisions based on these metrics. Similarly, Alahmad and Agarwal [5] make use of metrics to evaluate both availability and reliability to take optimised VNF placement decisions. Agarwal et al [6] propose a placement strategy based on vertical business requirements, to insert the notion of final-user in the loop. Also, Jemaa et al [7] present strategies that enable QoS-aware VNF placement based on real-time performance metrics and QoS models.

All the previous research works do not consider scenarios where there are multiple geographically separated potential locations to place VNFs. This is, in fact, addressed by Benkacem et al [8] extending the coverage of the network infrastructure to include both edge and core network segments as possible candidate options for VNF placement, which in turn, are based on quality of experience (QoE) metrics.

Laghrissi et al [9] present a benchmark to determine the best algorithms for VNF placement and provide their own predictive placement strategy.

Unfortunately, none of the research works and reference software implementations analysed have considered the use of spatial metrics to inform the decision making process to allocate VNFs. None of them have also ever consider tailored scenarios where not only multi-tenant infrastructures are at operations but also where there are deployments of multiple 5G networks running on top of such tenants where the mobility of 5G users need to be considered across multiple antennas.

Pei et al [10] has gone a step forward making use of the topological information of geo-distributed datatcenter to optimise two different use cases. The first case is to optimise the allocation of VNFs making use of the topological information and the service function chain associated to such VNFs. The second case is to enhance such placement by reducing the number of redundant VNF instances by allowing them to take combined workload. This is a solution closer to our approach since it take partially into consideration topological information. However, they only take a partial representation of the topology where there is not any 5G infrastructure elements, and not any user mobility across the infrastructure. Furthermore, their experiments have been tested only on a small scale (75 nodes) network topology, clearly need further optimisation. To the best of our knowledge, this is the first proposal that considers the topological information coming from different and heterogeneous data sources, including, physical infrastructure, virtual infrastructure, 5G mobile network infrastructure and 5G final-users in order to take VNF placement decisions based on the current shape and size of the network to make sure that such decisions will benefit to the majority of the users, which is a clear way to optimise resource allocation decisions.

III. TOPOLOGY-AWARENESS IN 5G MULTI-TENANCY NETWORKS

Fig 1 shows an overview of the proposed architecture to achieve a smart VNF placement strategy. The architecture consists on different software architectural components exchanging messages through a middle-ware communication message bus using a publication/subscription paradigm. All the subscribers of a given topic will receive the messages associated and the senders will publish those messages to react the interested parties.

The three bottom components in Fig 1 are responsible for gathering the topological information from the underline infrastructure, keeping such information up-to-date and reporting it every time there is a change in the topology (see Topology Exchange). These agents are described as follows:

Resource Inventory Agent (RIA) is a component deployed inside each physical and virtual machine of the 5G multi-tenant network. It is responsible for generating the topological information for all the devices, ports and connections between ports available inside of the machine. RIA is responsible for discovering the VMs and physical and virtual network interfaces as well as logical switches and interconnection between network interfaces.

Topology Inventory Manager (TIM) is a component deployed in the management plane that provides information not available inside the computers/devices, where RIA is deployed. This lack of availability could be for two different reasons. Firstly, RIA cannot be installed in closed hardware devices such as managed switches and routers and thus topological information needs to be collected from a centralised place.
Secondly, there are some cases where information is stored in the management software and not distributed to computers. These are architectural decisions of the management stack used. Therefore, TIM will extract the information by its integration with the management stack. As example, TIM will extract the tenant information about the networks through the management interfaces of OpenStack.

**5G Mobile Agent (MA)** is a component deployed in the control plane that provides information about the attachment of the 5G mobile users to the specific 5G distributed units and keeps tracking the user mobility across antennas. To achieve so, this component is directly integrated with the control plane of both 4G and 5G networks by retrieving such information from the 4G Mobility Management Element (MME) and 5G Session Management Function (SMF) / Access and Mobility Management Function (AMF). To be concrete, this integration has been performed by creating an ad-hoc API in the MME implementation of the Mosaic 5G project as described by [11]. Marco et al [12] provides a comprehensive explanation of this API.

The previously described architectural components report the topological information at periodic intervals to keep updated and in real-time the entire network topology discovered.

The key component of the proposed architecture is **VNF Spatial Metric Agent (vSMA)**. It is deployed in the management plane and computes in real-time 5G spatial metrics (explained in section IV), based on the topological information reported from the agents. The vSMA keeps the 5G topology updated in a graph structure and performs periodic calculation of the 5G spatial metrics by interrogating the graph in order to extract topology information about size and shape of the networks to make VNF placement optimal. The graph changes continuously according to the reports coming from the topology discovery agents. The framework computes continuously the metrics for each physical server capable of allocating VNFs.

These metrics are then received by the **VNF Scheduler** component of the framework to make advanced topology-aware decisions for 5G network VNF placement. The **VNF Scheduler** is in charge of deciding the VNF placements. In terms of prototypical validation, it is a new scheduling algorithm implemented over OpenStack. This component is also receiving the capabilities of the different options for placement. Thus, the **VNF Scheduler** aggregates information about the available resources (e.g. CPU availability) and combine this knowledge in order to provide the best decision to place the VNF requested.

Finally, the **Compute** components are responsible to enforce the VNF allocation in the associated physical machines. In terms of prototypical validation, this is the traditional OpenStack Nova service.

**IV. 5G Spatial Metrics**

Spatial networks metrics rely on the structure of the networks, dynamically gathered from the agents in charge of topology discovery to calculate relevant information about the size and shape of the network to make smart decisions on VNF allocation. However, current state of the art is not yet capable of providing spatial network metrics tailored for 5G multi-tenant infrastructures where complex topologies with hundreds of thousands composed the network, including physical and virtual infrastructures and mobile users connected, showing a

---

1 OpenStack Compute (nova) is available at https://docs.openstack.org/nova/
level of complexity and scalability that makes network graphs impracticable, specially when computing those metrics in real time and considering the whole 5G network topology.

VNF placement decisions require a complete overview of the network topologies to make valuable deployments, for instance, by deciding the allocation of a VNF in an appropriate MEC node depending on the number of connected users and their demands. Figure 2 represents graphically the abstraction layer of the graph used to represent a 5G network topology, where it can be identified several User Equipments (UEs) connected to two different 5G Distributed Units (DUs) which are used as radio access network to provide connectivity to the users towards the edge of the network. Two different EDGE computers allocated in two different geographical locations at the edge of the network. They represent physical servers. This two physical servers are connected to routers and switches towards the transport network to interconnect the CORE data center with the edges. The CORE network segment is represented by two physical servers. These 4 different physical servers are hosting a set of VNFs running different services. The naming convention is to make use of a "v" prefix to indicate that they are VNFs. This, vAUSF, vPCF, vSMF, vAMF, vUDM, are all the architectural components of the 5G network running in virtualised machines, the vCU nodes represent virtualised Centralised Units, the vVNF node represents a virtualised VNF and vCDN nodes represent a virtualised Content Delivery Network service.

$$SP_G(v, n)$$ is defined as the Shortest Path in the graph $$G$$ between the node $$v$$ whose metrics is being calculated and the rest $$N$$. The distance function used on the calculation of the shortest path algorithm is achieved using a weighted graphs. It allows the possibility to tailor the metric for specific use cases, for example, for bandwidth optimisation, delay optimisation, wider benefit optimisation, etcetera. The Shortest Path algorithm is based on \textit{Dijkstra’s algorithm}, which runs in time $$O(m + n\log n)$$. It should be noted that this algorithm has been tailored to 5G Multi-tenant Network Topology graphs and thus the metrics is only calculated for those nodes that can allocated VNFs.

$$\forall v \in V, C(v) = \frac{1}{\sum_{n \in N} SP_{GB}(v, n)}$$ (1)

Note that Fig 2 has the values calculated over the graph shown in the figure using Eq 1 to allow the reader to understand how the algorithm is applied over a given topology. The metric considered is based on wide benefit optimisation, i.e. the metric will indicate the place that will benefit more nodes of the network. Thus, the reader can see how the top score is on the physical machine on EDGE-A where the vast majority of UEs is attached to the radio access network. The calculation of this metric has been implemented inside of the \textit{VNF Spatial Metric Agent} previously described in Section III. This agent keeps a graph model of the whole topology and computes periodically the algorithm to provide the score for every placement option and shared with the \textit{VNF Scheduler}.

V. SMART VNF PLACEMENT STRATEGY

The \textit{VNF Scheduler} is receiving at periodic intervals the available resources on all the physical machines of the infrastructure. It is also receiving at periodic interval the metric values of the 5G spatial metric previously described for all the physical machines of the infrastructure. Then, when a VNF placement request is received to the \textit{VNF Scheduler}, the best location to perform the placement is done as follows. First, the subset of physical nodes is filtered according to the requirements for the VNF requested so that a subset of viable placement locations is considered. Then, such placement locations are ordered by the value of the 5G spatial metric and the most optimal location is then selected as a decision to enforce the placement. The most optimal is the one providing the maximum (or minimum) value of the metric depending if we are facing a maximisation or a minimisation problem. By making use of the 5G spatial metric, the decision is considering the information of the whole network topology and the best location is the location that is less used and closest to the majority of UEs will be selected. Our Smart VNF Placement Strategy consists on a solution that takes into account the status of the network topology in real-time. This status in a graph format provide a score for every placement option to be consider.

VI. eMBB MEDIA STADIUM USE CASE

This section provides a concrete example for a specific use case to allow the reader to understand the application of the...
VNF placement strategies in-situ. The use case represents a smart city scenario similar to a large stadium, that consists on a network topology where there is a large amount of UEs connected to the same edge and all of them require an enhanced mobile broadband (eMBB). This is a use case that is being studied by telecommunication operators as a demanding 5G use case where there is a need to adapt the existing infrastructures for the incoming 5G networks deployment.

Fig. 2 shown a concrete disposition of the UEs connected to the network. The disposition of the nodes in the graph allow to represent the difference between the amount of UEs connected to each DUs, which are located in different edges in order to represent a high dense scenario and its difference with a low dense one. Obviously, the number of nodes is just example to illustrate the unbalance, and in practice we will have hundreds of thousands of devices. Scalability is addressed in further sections, and the purpose of the graph is to illustrate the effectiveness of the metric so that this distribution of the nodes represent a proportional realistic situation where EDGE-A is serving coverage to the stadium.

The whole procedure to provide a smart VNF placement start gathering metrics from the Topology Exchange where the distributed and centralised topology discovery components are reporting. When achieved the topology state shown in Fig. 2 the vSMA will have calculated the $C(v)$ metric for every VNF placement candidate. Then, a VNF request to place a vCDN cache to deal with the density of the studio is received, as a high number of users are demand video delivery. Then, the VNF Scheduler receives the request and using the best value of metrics to make the decision. In our case, 99% associated to EDGE-A. As a result, the VNF Scheduler decides to place the vCDN tagged the most beneficial location for all the nodes of the network, i.e. EDGE-A.

VII. EMPIRICAL EVALUATION

The solution proposed has been implemented using OpenStack VNF placement and an integration of the new proposed scheduling algorithm to consider the metrics reported by our new architectural component VNF Spatial Metric Agent. This section describes scalability tests carried out to perform the empirical validation of the proposed architecture of or a large-scale 5G multi-tenant networks.

A. Testbed Description

All the empirical executions have been carried out in a Cyberserve XE5-308S v4 computer, with Dual E5-2660 v4 Intel Xeon, 14 Cores, 2.00GHz, 35M Cache, 105Watts with hyperthreading activated, 128GiB DIMM DDR4 Synchronous 2400 MHz, 1.6 TB Intel SSD PCIe, and Ubuntu 18.0.4.2 LTS. This computer runs the vSMA software component for the calculation of the spatial metrics. Our 5G infrastructure is composed by nine Cyberserve computers with the same specifications indicated. They run OpenStack Newton 2 where one computer is the cloud controller and the other eight are computes employed for a 5G mobile edge computing network. These computes have deployed inside the Mosaic 5G stack 3. The infrastructure is deployed with Ethus USRP X310 as DU. All the switches of the infrastructure are Netgear GS724T at 1 Gbp Ethernet. In order to stress the scalability of the topology, the three different topology discovery components (RIA, TIM and MA) have been extended with an extra functionality that allows to emulate the reporting of large scale infrastructure even if the physical components are not available. This allows us to scale up from only 8 UEs available in our premises to the 65k nodes reported in this contribution. But, it is important to say that the behaviour of the component will be similar when the real uses are present in the infrastructure since we are emulating them, and not any simulation is being taking place.

B. vSMA Scalability Results

An scalability test has been conducted with the intention of show how the VNF Spatial Metric Agent dealing with large-scale topologies to provide metrics to the VNF Scheduler. Fig. 3 shows the time required (y-axis) by vSMA after receiving the topological information to calculate and report all the values of the metrics for all the possible VNF locations involved in the scenario. The implementation matches the architecture described in Fig 1. For all the experiments, the number of physical machines available in the CORE network segment has been fixed to 16 with their associated virtualised components for simplicity. The EDGE physical machines are used to deal with scalability, ranging from 1 to 256, all of them connected to a DU. DUs allows UEs to be connected. A fixed ratio of 256 UEs per DU have been fixed as an example of high dense scenario. Each scenario has 8 different tenants, requiring VNFs to deliver services for the network slices associated to such tenants. These description leads to scenarios ranging exponentially from a minimum of 256 UEs to a maximum of 65536 UEs which is a significant number of UEs to be covered by the stadium use case presented. Notice that these numbers are completely different from the sizes of the topology graphs that are being managed to support such scenarios. To be concrete, the total number of nodes for these two cases are, 2558 nodes for the smallest scenario and 537803 for the biggest one.

As expected, results show an exponential ascendant trend. However, it is very relevant to mention that the metric calculation take less than 1 seconds for up to 8192 UEs and for the largest scenarios with 65536 nodes, it take less than 20 secs which is significant result that allows to take decision on a very decent level of freshness for the metric values.

Fig. 4 shows a different perspective for the scalability of the proposed VNF placement strategy. This time the number of UEs is fixed to 65536 UEs, which is the maximum being considered in our experiments due to lack of resources in the data center to perform higher emulations and the ranging is now focused on the number of edge physical machines. The number of UEs connected is always the same for all the

---

2OpenStack is available at https://www.openstack.org/

3Mosaic 5G is available at http://mosaic-5g.io/
experiments and what changes is the ratio of UEs connected to each EDGE, according to the number of EDGEs available in the scenario. Thus, the number of edge PMs has been ranged from 1 to 1024 which means that the number of UEs has been ranged from 65536 to 64 to keep the total. The rest of the topology is the same than the one used in the previous experiment.

Fig. 4 shows an almost constant trend and around 20 seconds (similar to the biggest scenario previous described) as the time required to calculate the 5G spatial metrics. It clearly shows how scalable the proposed architecture is despite the number of edge nodes. These results make the 5G spatial metric usable and feasible and demonstrate how the VNF placement decision are optimised to the place where they are most effective for the final user. This placement will allow to provide a smart service in the context of the network slices of each of the tenants available in the infrastructure.

VIII. CONCLUSION

This paper has proposed a solution based on the usage of topological information which is collected in real-time to provide a topology-aware description of the 5G multi-tenant network. Such information is then used to create new spatial metrics at periodic intervals that allows the VNF Scheduler to take optimal placement strategies by making use of the topological information available. The architecture has been empirically validated based on an OpenStack infrastructure and scalability tests carried out over a realistic 5G multi-tenant network show efficient metric calculation and VNF placement decision to make the system practical in operational phase. VNF placements are taking the same overhead as the original OpenStack scheduler while the time to provide the updated value of the metrics is around 20 seconds for a high dense eMBB scenarios with 65536 UEs, which clearly provides some useful insights of the practical aspects of the approach presented.

As future work, it is planed to improve the system to deal with other aspects that will enrich our existing abstraction of the network topologies, dealing with Lora IoT networks, Wi-Fi networks and other types of networks currently not being addressed in our research work.

REFERENCES