To Distribute or Not to Distribute? Impact of Latency on Virtual Network Function Distribution at the Edge of FMC Networks

Marco Savi*, Ali Hmaity*, Giacomo Verticale*, Stefan Höst§, Massimo Tornatore*

*****Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy*

§Department of Electrical and Information Technology, Lund University, Lund, Sweden

E-mails: {marco.savi; ali.hmaity; giacomo.verticale; massimo.tornatore}@polimi.it, stefan.host@eit.lth.se

ABSTRACT

The Network Function Virtualization paradigm enables the possibility to dynamically instantiate Virtual Network Functions (VNFs) in Commercial-Off-The-Shelf (COTS) hardware. Such VNFs are then concatenated together in Service Chains (SCs) to provide specific Internet services to the users. Depending on latency requirements for such services and considering the aim of maximally consolidating the VNFs (i.e., of minimizing the COTS hardware), the VNFs can be centralized in few datacenters in the core network or they can be distributed closer to the edge of the network. In this paper we evaluate the impact of latency requirements of SCs on VNF distribution towards the edge of the network, by also showing the benefits of a Fixed and Mobile Convergent (FMC) metro/access network, with respect to a non-convergent network, in terms of consolidation.

Keywords: Network Function Virtualization, Service Chaining, Fixed and Mobile Convergence, Metro network

1. INTRODUCTION

Network Function Virtualization (NFV) [1] is a novel network-architecture paradigm that helps fixed and mobile network operators to reduce both capital and operational costs. NFV is based on the concept of *network function*: a network function is an abstract building block performing a specific task. Examples of network functions are Firewalls, Traffic Monitors, etc. So far, network functions have been implemented using dedicated hardware, usually referred as *middleboxes*, that are able to handle very high traffic load, but are expensive and inflexible. NFV allows a move towards a *softwarization* of network functions in a virtualized environment. Multiple *Virtual Network Functions* (VNFs) can thus be instantiated and consolidated in the same Commercial-Off-The-Shelf (COTS) hardware, that can potentially be placed in any powered location of the network. NFV also eases service deployment by exploiting the concept of *Service Chaining* [2]: a Service Chain (SC) is a sequential concatenation of VNFs to provide a specific Internet service (e.g., VoIP, Web Service, etc.) to users.

One of the main problems faced by a network operator adopting NFV is deciding where to locate the VNFs in the network (in which nodes) to minimize network cost, while still satisfying the latency requirement of the supported Service Chains. To achieve such objectives, the instantiation (i.e., placement) of VNFs must be carefully planned. On one side, the optimal solution in terms of costs for a network operator would be placing the VNFs in a datacentre (DC) to provide all Internet services from a centralized and cheap location. However, this solution may degrade the performance of latency-sensitive SCs, due to the excessive distance of the DCs that, in some cases, can be even thousands of kilometers far from the users. Hence, to avoid such degradation, it is necessary to locate the VNFs closer to the users at the edge of the network [3]. The main candidate nodes to host VNFs at the network edge (i.e., in the metro/access segment) are the Central Offices (COs) at different hierarchical levels of the fixed and mobile aggregation networks (i.e., COs, Main COs and Core COs in Figure 1). However, the effectiveness of VNF distribution across the edge of the network is hindered by the fact that fixed and mobile access and aggregation networks have evolved and deployed independently and VNFs placed in fixed network COs cannot be easily accessed by mobile users (and vice versa). Recent studies [4] have targeted the definition of novel architectures for Fixed and Mobile Convergent (FMC) networks, where the fixed and mobile networks are jointly designed and optimized both from a functional (i.e., by unifying network functionalities) and structural (i.e., by sharing equipment and infrastructures) perspective. In this study we argue that the adoption of a novel FMC access and aggregation network can help network operators in consolidating the VNFs in shared locations (i.e., over the same COTS hardware) for fixed and mobile users. To explore the benefits of FMC on VNF consolidation, we evaluate the impact of latency on distribution and centralization of VNFs when different SCs must be *embedded* in the network, by comparing a *FMC* and a *No FMC* architecture. We show that for some low-latency Internet services the involved VNFs must be distributed across the edge to avoid a significant degradation of the service quality.

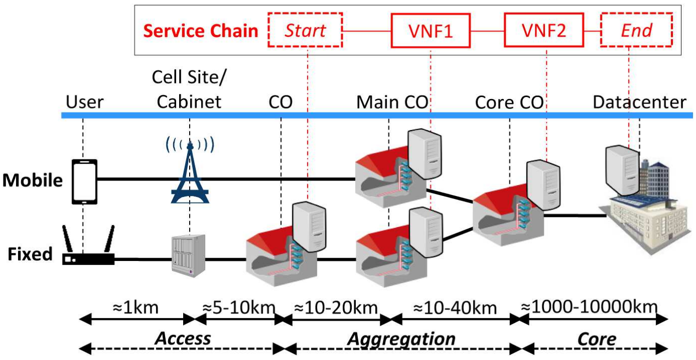

Figure 1: NFV-enabled fixed and mobile aggregation network

2. PROBLEM STATEMENT

The optimization problem associated to the VNF placement and SC embedding can be seen as an extension of some Virtual Network Embedding (VNE) problems [5], where the SCs are *virtual networks*, chaining together a start point, an end point and some VNFs that must be crossed in an appropriate order, as shown in Figure 1. The VNFs are associated to some *processing requirements*, expressed in terms of fraction of required dedicated CPU cores. The SCs, that are associated to a maximum end-to-end *latency requirement*, must be embedded in a *physical network*, which comprises physical nodes and physical links. Some of the physical nodes are *NFV nodes* (i.e., nodes able to host the VNFs in COTS hardware), and are associated to some *processing capacity*, expressed in terms of number of adopted CPU cores. We consider that both physical links and NFV nodes introduce some latency in crossing the SC from the start to the end point. The latency introduced by the links is due to propagation delay over the physical links and to transmission time of the network devices (i.e., switches), while the latency introduced by the NFV nodes is due to processing resource sharing among multiple VNFs hosted by the NFV nodes. As in [6], we consider the *context switching*, i.e., the operation of saving/loading the state for parallel execution of the multiple VNFs sharing the NFV node, as the primary source of latency in the traversal of NFV nodes. The objective of the optimization problem is *maximally consolidating* the VNFs (i.e., minimizing the number of nodes to be upgraded to host VNFs). This means placing the VNFs in the minimum number of NFV nodes while meeting the end-to-end latency requirement for the SCs and satisfying processing capacity constraints for the NFV nodes. Note that a VNF can be shared among multiple SCs by properly scaling up the processing requirements for that VNF. For a more detailed representation of the system model and of the problem statement the reader is referred to our work [7].

2.1 Heuristic algorithm for latency-aware SC embedding

We developed a *heuristic algorithm* [7] for the embedding of SCs in the physical network that takes into account latency requirements. The algorithm works in two distinct phases. The main idea is building an embedding solution for each SC by greedily trying to scale up already-placed VNFs or to place new instances of VNFs on already-active NFV nodes (*phase 1*). Note that phase 1 has as main objective the VNF consolidation, since it tries to exploit already-used resources first. At the end of phase 1 the end-to-end latency for the selected SC is evaluated. If latency requirements are not met due to an inadequate embedding on the physical topology, a *phase 2* is performed to improve the solution. Phase 2 consists in releasing the resources allocated in phase 1 and placing a new instance for each chained VNF on an inactive (i.e., turned off) NFV node on the *latency shortest path* between the start and the end point of the SC. This way, the algorithm tries to adjust the solution by minimizing the latency introduced by the links and by the NFV nodes.

3. SIMULATIVE SCENARIO AND ASSUMPTIONS

3.1 Physical network

We consider the metro/access network topology shown in Figure 2. Such network topology is based on the urban geotype proposed in [8]. In particular, it consists of 1 Core CO, 6 fixed Main COs and 6 mobile Main COs, each covering an area of 15 km^2. Each fixed Main CO aggregates the traffic of 3 fixed COs, all connected in a ring. Each fixed CO aggregates the traffic of 95 cabinets, and each cabinet aggregates the traffic of up to 160 fixed users, i.e., homes. For the mobile network, each mobile Main COs aggregates the traffic of 23 cell sites, each one aggregating the traffic of 3000 mobile users. The total coverage area of this network is in the order of the size of a large European metropolitan city. The Core CO and fixed/mobile Main COs are connected by a ring network. We assume that this network is connected to a DC placed in the core network. We model the core network as a single link whose latency reproduces the total latency experienced to reach the DC location. The Core CO, Main COs and COs can host COTS hardware with limited processing capacity, while the DC has no limitations on processing capacity. The Core CO is assumed to support 20 CPU cores, i.e., twice the processing capacity of the fixed/mobile Main COs, and we consider the COs processing capacity as 30% of the total capacity of a Main CO. In this work we investigate two network architectures: *FMC* and *No FMC*, as shown in Fig. 2. In the *No FMC* architecture, fixed (mobile) network users can access only VNFs that are placed into the fixed (mobile) network infrastructure, i.e., in fixed Main COs, fixed COs (i.e., in mobile Main COs). We will generically refer to such nodes as *fixed (mobile) NFV nodes*. The Core CO and the DC can be accessed by both fixed and mobile users. On the other hand, in the *FMC* scenario fixed and mobile users can share the network infrastructure and, thus, also the NFV nodes. We considered 3 different DC location configurations: *Short-range DC, Mid-range DC* and *Long-range DC* with latencies equal to 15, 75

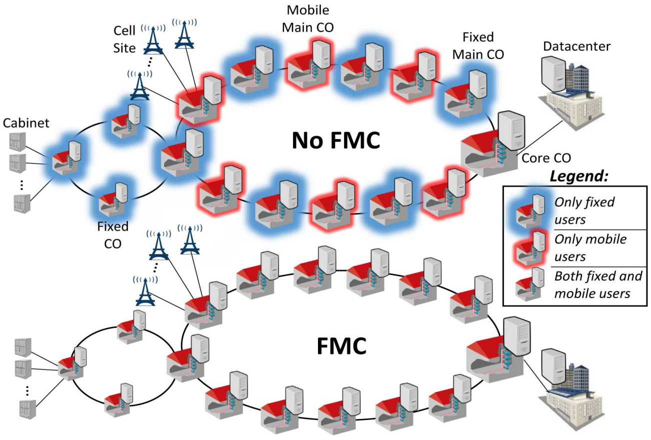

Figure 2: NFV node accessibility in *FMC* and *No FMC* architectures

and 150 *ms*, corresponding to a national, continental and intercontinental DC location. These latencies are based on measurements performed using [9], a tool to evaluate the ping distance from *Amazon Web Services* around the globe, and express the one-way latency to the DC location from the Core CO. Note that, given the geographical coverage of the metro/access network, the latency introduced by propagation and transmission in the metro/access links is negligible with respect to the latency introduced by the core network. In addition, we consider the *context switching* latency introduced by an NFV node to increase linearly with respect to the number of VNFs sharing such NFV node and to be equal to 100 μs per VNF.

3.2 Service Chains and Virtual Network Functions

We consider a set of 5 different SCs, as shown in Table 1, each one chaining different VNFs in sequential order. Each SC type is associated to a different end-to-end *latency requirement*. The Web Service (WS) is recognized to have a loose latency requirement, while novel 5G Services (5GS, e.g., Augmented Reality Service) require a very strict end-to-end latency [10]. It is worth mentioning that some cloud gaming categories might have a latency requirement comparable with 5G services, while other categories might tolerate higher latencies. In this work we consider that Cloud Gaming SCs have a latency requirement of 60 ms. We consider also other three SC types, i.e., VoIP, Video Conferencing (VC) and Cloud Gaming (CG). Note also that each VNF is associated to a *processing requirement* per user, obtained by middleboxes datasheet (e.g. [11]).

Service	**Chained VNFs**	**Latency req.**
Web Service (WS)	NAT-FW-TM-WOC-IDPS	500 ms
VoIP	NAT-FW-TM-FW-NAT	100 ms
Video Conferencing (VC)	NAT-FW-TM-VOC-IDPS	80 ms
Cloud Gaming (CG)	NAT-FW-VOC-WOC-IDPS	60 ms
5G Service (5GS)	NAT-FW-TM-WOC-VOC	20 ms

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

Table 1: Details of the considered Service Chains and latency requirements

Each SC aggregates the traffic of the users connected to the fixed/mobile Main CO or fixed CO it starts from. In particular, the start points for fixed (mobile) SCs are all the fixed Main CO and CO (mobile Main CO), resulting in an overall number of 30 SCs (see Figure 2). Concerning the end points, we compare three different cases with a different percentage of local traffic terminating in the metro network: *0%*, *50%*, *100%*. The first setting (*0%*) represents the case where all the SCs have as destination point at the DC location in the core network. In the second setting, (*50%*) half of the SCs has as destination the Core CO in the metro network and the remaining half terminate at the DC location. Finally, in the last setting (*100%*) all the SCs terminate at the Core CO (i.e., at the edge of the metro network). Note that considering a fraction of SCs that terminate in the metro network follows the current trend of telecom operators, which tend to push the content towards the users. For example, a Video Content Provider, Cloud Game Provider etc. may place Video Servers, Game Servers etc. in the metro/access network (i.e., in our case, in a *micro DC* located in the Core CO).

4. RESULTS

We implemented the heuristic algorithm described in Section 2.1 in Matlab. We compare results obtained by considering five different *homogeneous scenarios*: in each homogeneous scenario, only one specific type of SC among the types defined in Table 1 is embedded in the network. This way, we can independently evaluate the impact of different Internet services on VNF consolidation. We also focus on both a *FMC* and a *No FMC* architecture as described in Section 3.1, and we consider the three different settings related to the percentage of local traffic as shown in Section 3.2. Our investigation takes into account the three DC-location configurations introduced in Section 3.1.

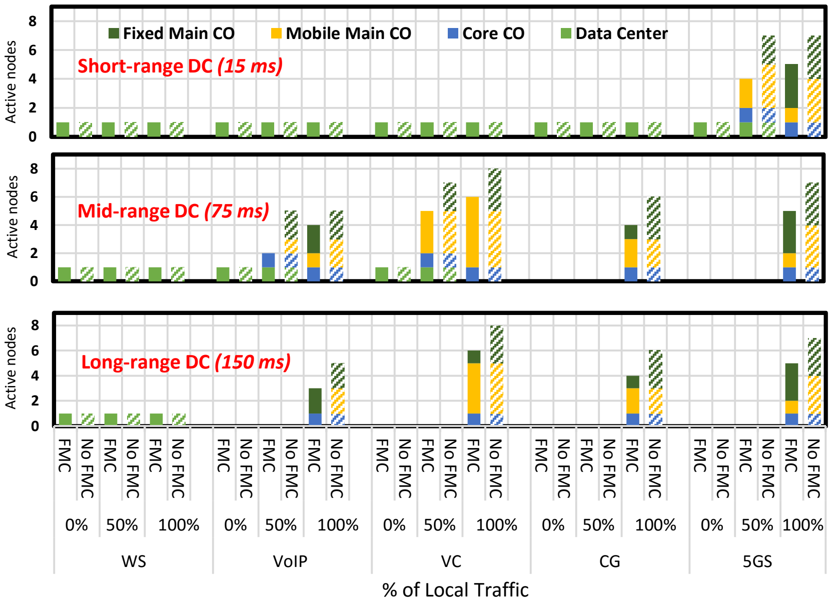

Figure 3: Number of active nodes in the *FMC* and *No FMC* architectures for three different DC-location configurations

Figure 3 shows the number of NFV active nodes for the various traffic configurations and network architectures discussed so far. For the *Short-range DC* configuration, we observe that the most convenient solution in terms of VNF consolidation is to host all the VNFs in the DC for every homogeneous scenario, every architecture and every percentage of local traffic, except for the 5GS SCs for *50%* and *100%* of local traffic. In these cases, it is required the activation of some fixed/mobile Main COs and of the Core CO. In fact, for all the SC types but the 5GS, consolidating the VNFs in the DC, even though part or all the SCs terminate in the Core CO, is a feasible solution because the Round Trip Time to the DC (30 ms) does not affect the latency requirement of the SCs. This is not true for the 5GS homogeneous scenario due to the very strict latency requirement of its SCs (20 ms). In this case and in conditions of local traffic, placing all the VNFs in the DC would degrade the performance. For this reason, it is necessary to distribute the VNFs in the metro/access network to meet latency requirements for the SCs terminating in the Core CO. For the *Mid-range DC* configuration, only the VNFs for the WS scenario can be all consolidated into the DC. Finally, for the *Long-range DC* configuration, only the WS homogeneous scenario can still be guaranteed by placing the VNFs in the DC for all the traffic conditions. For the other scenarios, the only feasible solution to meet latency requirements is to have all the VNFs placed in the metro/access network and to keep all the traffic local (*100%*).In general, from Figure 3 we can see how the impact of latency on VNF consolidation is similar for the *FMC* and *No FMC* architectures. However, when the VNFs are distributed in the metro/access network, the *FMC* architecture requires from 30% to 60% less NFV active nodes than the *No FMC* one. This happens because in the *FMC* architecture the NFV nodes as well as the VNFs placed on those nodes are shared between fixed and mobile users. This means that the adoption of a FMC metro/access network can consistently improve the consolidation of VNFs. We now focus on some *processing* aspects. By looking at Figure 3, we can notice that the VC homogeneous scenario requires the activation of more NFV nodes than the other scenarios. This happens because, in average, the VNFs chained by VC have a higher processing requirement than the other types of SCs. Moreover, the placement of VNFs is slightly different when NFV nodes dispose of more processing capacity. In Figure 4 we compare the results for the most processing-hungry service (i.e., the VC), obtained with the previous simulation settings (i.e., normal bars), with the case where the NFV nodes processing capacity is increased by 50% (i.e., diagonal-shaped bars). The increase of processing capacity allows placing the VNFs in less NFV active nodes (from 30% to 40%) for both the *FMC* and *No FMC* architectures. This means that increasing the processing capacity of NFV nodes is beneficial for VNF consolidation.

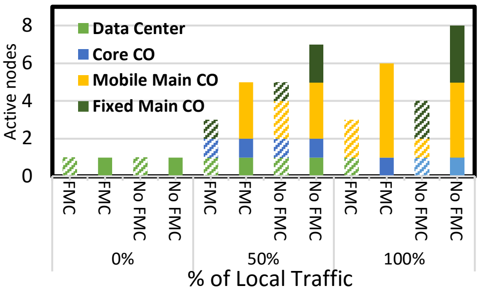

Figure 4: Number of active nodes for VC with increased processing capacity for the NFV nodes by 50% (standard bars) with respect to the settings of Fig. 3 (diagonal-shaped bars) in the *Mid-range DC* configuration

5. CONCLUSION

In this work, we evaluated the impact of latency on the distribution of VNFs in the metro/access network. We considered different Internet services with different latency requirements and we compared the results for different DC location configurations. Results show that Internet services with a very strict latency requirement (e.g., 5G Services) can be only satisfied by placing the VNFs in the metro/access network, while Internet services with looser latency requirement (e.g., a Web Service) can still be guaranteed by consolidating the VNFs in a DC far from the users. Moreover, an *FMC* architecture allows a higher VNF consolidation than a *No FMC* architecture, since the NFV nodes can be shared between fixed and mobile users.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community Seventh Framework Programme FP7/2013-2015 under grant agreement no. 317762 COMBO Project.

REFERENCES

[1] R. Mijumbi, *et al.*, Network Function Virtualization: State-of-the-Art and Research Challenges, in *IEEE Communications Surveys & Tutorials*, vol.18, no.1, pp.236-262, First Quarter 2016

[2] W. John, *et al.*, Research Directions in Network Service Chaining, in *Conference on SDN for Future Network and Services*, *IEEE*, Nov. 2013

[3] R. V. Rosa, *et al.*, M2D-NFV: The Case of Multi-domain Distributed Network Functions Virtualization, in *International Conference and Workshops on Networked Systems (NetSys) 2015*, Mar. 2015

[4] S. Gosselin, *et al.*, Fixed and Mobile Convergence: Needs and Solutions, in *European Wireless 2014; 20th European Wireless Conference*, May 2014

[5] A. Fischer, *et al.*, Virtual Network Embedding: A Survey, in *IEEE Communications Surveys & Tutorials*, vol.15, no.4, pp.1888-1906, Fourth Quarter 2013

[6] I. Cerrato, *et al.*, An Efficient Data Exchange Algorithm for Chained Network Functions, in *15th International Conference on High Performance Switching and Routing (HPSR) 2014*, Jul. 2014

[7] M. Savi, *et al.*, Impact of Processing-Resource Sharing on the Placement of Virtual Network Functions, submitted to a journal, 2016

[8] FP7 COMBO Project, Analysis of Transport Network Architectures for Structural Convergence, *Deliverable D3.3*, Jun. 2015

[9] www.cloudping.info

[10] ETSI, Mobile-Edge Computing, Introductory Technical White Paper, Sept. 2014

[11] "Dell SonicWALL SuperMassive 9000 Series Firewall." [Online]. Available: www.sonicwall.com/documents/sonicwall-supermassive-next-generation-firewall-series-datasheet-68226.pdf