On using SDN in 5G: the controller placement problem

* Adlen Ksentini, ‡ Miloud Bagaa, ‡ Tarik Taleb
* adlen.ksentini@eurecom.fr, EURECOM, Sophia-Antipolis, France
‡ name.surname@aalto.fi, AALTO University, Finland

***Abstract*—To integrate Software Defined Networking (SDN) in the envisioned 5G system, a separation of the control and user data plane functions of the Evolved Packet Core (EPC) is required. This separation will impact mainly the functions available at the Serving GateWay (SGW) and Packet data GateWay (PGW) elements, and will result in two new entities; i.e. the S/PGW-C and S/PGW-U (PGW-C and PGW-U). The S/PGW-C integrates all the control plane functions (such as signaling and tunnel creation), while S/PGW-U contains only forwarding functions. The S/PGW-C will control the S/PGW-U in order to forward the UE traffic to the appropriate destinations by enforcing rules e.g., using the Openflow protocol. Usually, the S/PGW-C will run as a Virtual Network Function (VNF) running on a Virtual Machine or Container instantiated over a federated cloud. In this paper, we focus on the problem of the SGW-C placement, where a tradeoff is needed between reducing the SGW relocation frequency and balancing the traffic load among the underlying SGW-C VNFs. We formulate this problem using optimisation models, and a fair solution (i.e. Pareto optimal) is derived using Nash Bargaining game and the threat point.**

I. Introduction

The envisioned 5G system will be certainly built employing both Software Defined Networking (SDN) and Network Function Virtualization (NFV) technologies. Specifically, the 5G mobile core network will evolve from current deployments based on dedicated hardware to a fully virtualized environment, relying not only on SDN and NFV but also on cloud computing; giving birth to the mobile carrier cloud concept [3]. The 5G mobile core network will be an evolution of the current Evolved Packet Core (EPC) toward a fully virtualized system, known as virtual EPC (vEPC)[1][2]. The vEPC will be built on software versions of the 3GPP core network elements (HSS, MME, SGW and PGW), hosted on Virtual Machines (VMs) or containers in a cloud computing system. The way to virtualize the EPC elements through SDN and NFV is still an open issue, which is attracting high amount of propositions from academic and industry [1]. One proof of the relative importance of this topic toward 5G is reflected by the fact that major telecom vendors and operators have established their own way to implement the vEPC.

However, most existing proposals converge toward a common commitment in order to introduce SDN and NFV in 4G and beyond, which consists in separating control and user plane functions within the EPC elements (i.e., MME, and S/P-GW). Control plane functions represent all functions related to signaling and tunnel establishment, while user plane forwards UE data from and to different Packet Data Networks (PDN) (e.g., Internet, IMS). In this paper, we will focus particularly on the separation of the control and data plane functions within the SGW and PGW entities, while the MME element remains unchanged. This separation enables the creation of: (i) intelligent entities (S/P-GW controllers), which may be run in a centralized element (physical machine or virtual machine running in the cloud); (ii) simple entities (S/P-GW-user plane; S/P-GW-U) that forward user traffic according to rules defined by the controller. In this context, two new entities will emerge SGW-C/PGW-C and SGW-U/PGW-U; where U stands for User plane (Data Plane) and C stands for Control plane. The control gateway, GW-C, manages the GW-U elements using southbound API (e.g. Openflow) protocol, whereby rules are enforced at the GW-U to forward the UE traffic to the appropriate PDNs. Furthermore, GW-C may run as a VNF on a VM/container hosted in the cloud. Last and not least, the interest to integrate SDN in 5G through the separation of control and user planes, has led the 3GPP to create a new study item [4] dedicated to the separation of SGW and PGW functions. Whilst this separation represents a step forward to 5G, it introduces several challenges. First, there is a need for specifications that define and specify the interfaces between the GW-C and the GW-U. Second, GW-C has to be placed at optimal points in the underlying cloud. The latter defines the focus of this paper.

The remainder of this paper is structured as follows. Section II presents some research work related to the integration of SDN in mobile networks. Section III introduces the issue we are addressing and describes the envisioned solutions. The performance results of the introduced solutions are presented and discussed in Section IV. The paper concludes in Section V.

II. Related work

As stated earlier, a high number of research papers have been published discussing how the EPC would be virtualized, and proposing new architectures that integrate SDN and NFV. Particularly, high-level architectures have been introduced in [5],[6] and [7], which give a general way to introduce SDN and NFV in 4G and beyond. Other works, like those detailed in the following paragraph, give more details and mainly rely on separating data and control planes.

Authors in [8] envision different architectures to apply SDN and NFV in LTE. In addition to the idea of separating the control and data plane functions, the authors distinguish between

resource management and signaling functions at the control plane level. Based on the separation of functions (i.e. control and user planes), the authors proposed four architectures:

- Full cloud architecture, whereby all EPC entities (control and data planes) run as VNFs, hosted in the cloud. Although this architecture is entirely virtual, it does not use SDN.
- Control-plane migration proposes hosting all control plane functions and elements (i.e. MME and S/P-GW-C) in the cloud, while keeping data-plane functions running on dedicated physical machines.
- Signaling-plane architecture, wherein the MME and control signaling functions of the S/P-GW (i.e. tunnel establishment, etc.) are hosted in the cloud, while data plane and resource management functions remain outside the cloud, running on dedicated machines.
- Scenario-based architecture follows the same principle as the signaling-plane architecture. The difference concerns the location where the signaling control is executed. Indeed, this architecture distinguishes between delay-critical scenarios, wherein the signaling control should run on the data plane, and CPU-intensive scenarios, which run signaling control in the cloud.

In [9], the authors propose focusing only on separating the control and data planes of the SGW entity, while the P-GW remains executed in a COTS equipment. The authors propose merging MME and SGW-C functions on top of a SDN controller as an application. The SDN controller enforces the control plane decision on the SGW-U using Openflow protocol. To this end, the authors propose replacing S1-MME (i.e., between the eNB and the MME) and S11 (i.e., between the SGW and MME) interfaces by the Openflow protocol. One of the major differences with [8] concerns the fact that the eNB has to understand the Openflow protocol in order to forward the UE data to the appropriate SGW. The research Work in [10] focuses on the separation of the PGW control and user plane functions. The PGW-C may run on a dedicated machine or on top of a VNF hosted in the cloud, while the PGW-U is run on a dedicated machine. Obviously, the PGW-C manages the control plane signaling, while the PGW-U forwards traffic from UE to the PDN and vice versa. One difference with the precedent research work concerns the introduction of PGW-O (i.e. Orchestrator), which manages the GTP-U tunnel according to the PGW-C requests. In [11], the authors rely on the separation of S/P-GW functions in order to propose a novel organization of the mobile core network. They propose creating a new entity, called Mobile Controller (MC), which is the brain of the EPC. It includes MME as well as S/P-GW combined control plane functions. Moreover, the authors merge the S/P-GW-U into one entity, called GW-U, which is controlled by the MC through the OpenFlow protocol. Mainly, all these research work aim at introducing SDN in LTE by assuming that S/P-GW-U are common Openflow switches able to understand the GTP-U protocol; in order to match data packet coming from UE and apply forwarding rules on them. The authors have considered this approach to keep some compatibility with the current 3GPP specifications, and ease the integration of SDN without changing all the interfaces. The research Work in [12] proposes withdrawing GTP tunnels by proposing a complete IP flat architecture managed by a SDN controller hosted at the MME. S/P-GW are replaced by common Openflow switches. Whilst the proposed solution is novel, it contains many drawbacks and does not support user QoS and mobility.

Enabling SDN in 4G and beyond requires to separate the control plane and data plane functions of the SGW and PGW by creating new entities, i.e. S/P-GW-C and S/P-GW-U. Depending on the aim, it is possible to run S/P-GW-U on either dedicated or virtual machines. However, the S/P-GW will surely run as a VNF or combined with a SDN controller. Figure 1 illustrates the envisioned solution in this paper. We mainly focus on separating the data plane and control plane function at the SGW level, keeping the PGW as one entity hosted in dedicated or virtual machine. The SGW-C entities are run as VNFs hosted on the federated cloud infrastructure owned by the mobile operator. The SGW-U elements are deployed on the field, and could be seen as common openflow switches able to match GTP tunnels. Each SGW-U is connected to several eNBs using the S1-U interface. Both SGW-U and SGW-C are connected through S5/S8 interface to the PGW. Communication interfaces between the SGW-C and the SDN controller use any southbound API. Finally, we assume that the MME is connected to the SGW-C through S11 interface and to eNodeB through S1-MME interface (i.e., not shown on the figure for the sake of figure clarity). No connection between the MME and the SGW-U is needed. It should be noted that we assume that the SGW-C contains some SDN controller functions.

III. Problem formulation

Having described in details the integration of SDN in LTE and beyond, i.e. based on the separation of the data plane and the control plane functions, in this section we detail the problem we aim to solve in this work and how the envisioned solution. Assuming the architecture shown in Fig. 1, the target problem concerns the placement of the SGW-C in a federated cloud architecture. We assume that the mobile operator network has deployed one DC per location or region. A region covers one geographical area. Several SGW-U are deployed per region, where each SGW-U is connected to different eNBs. In the case of separating the SGW function, each SGW-C is covering one area. When a UE moves between two areas, SGW relocation is needed. This means, that the mobile operator should maintain the UE connectivity by transferring the UE context from one SGW-C to another SGW-C, while ensuring that flow rules are moved from one SGW-U to another SGW-U. The SGW relocation is costly for the operator due to the high number of involved signaling messages. Therefore, the SGW-C placement algorithm should limit the number of SGW relocation. To ensure this objective, the optimal solution will consist in creating only one SGW-

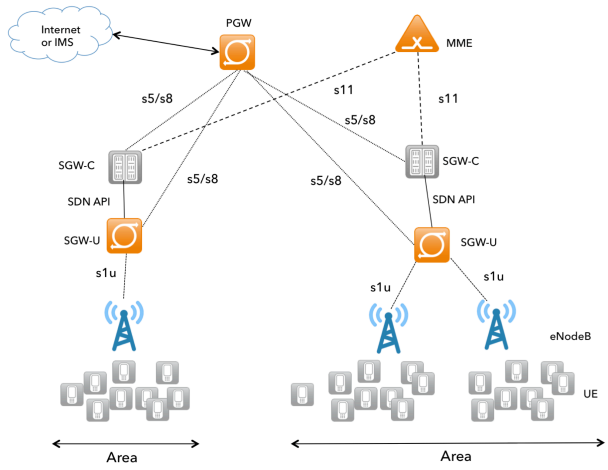


Fig. 1: Envisioned SDN-based mobile network architecture

C for the whole mobile network. Nevertheless, this solution would lead to overload the SGW-C; resulting in an increase of the flow installation delay in the SGW-Us. Consequently, the SGW-C placement algorithm should ensure that SGW-C entities are lightly loaded. To achieve this objective, the optimal solution will consist in instantiating only one SGW-C per DC, which in turn increases the number of SGW relocation.

Accordingly, the SGW-C placement algorithm should consider conflicting objectives, wherein a Pareto-optimal solution needs to be derived. To solve this problem and compute the Pareto-optimal solution we propose using Game Theory rather than classical multi-objective optimization approaches. Indeed, as indicated in [13], multi-objective solution needs to define weights for the different objectives, which are in different scale and hard to derive as we would like to satisfy both objectives. Before describing in details the envisioned model, based on Game Theory, we will begin by a formal description of the model and formulate two optimization problems for each objective.

We assume that the network operator owns small-scale data centers distributed over different $\mathcal{N}$ locations. Each location corresponds to an area (noted A_i). Each area includes a certain number of eNB (depending on the population density), which are connected to one SGW-U through the S1U interface. The number of SGW-U could be one or more, also depending on the population density and traffic to be carried out (noted by w_i). Each A_i is controlled by only one SGW-C. We denote by $h(i,j)$ the frequency of handovers between areas i and j. This information is easily obtained by monitoring the number of handovers seen on X2 and S1 interfaces.

A. Minimizing the SGW relocation

We recall that the first objective aims at reducing as much as possible the number of SGW relocations, while maintaining an acceptable load on each SGW-C. We consider two matrices: $\mathcal{S}$ and $\mathcal{P}$. If the areas i and j are managed by the same SGW-C, then $\mathcal{S}(i,j) = 1$. Otherwise, $\mathcal{S}(i,j) = 0$. Moreover, if $\mathcal{P}(k,m) = 1$, then A_k is controlled by G_m. Finally, $h(i,j)$ denotes the frequency of handovers between A_i and A_j, while w_i represents the traffic generated in A_i. Then to minimize the SGW relocations, we formulate the following Integer Linear Program (ILP) as follows:

$$\textbf{min}\, \mathcal{F}(\mathcal{S},\mathcal{P}) \tag{1}$$

$S.t,$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) + \mathcal{P}(j,t) - 1 \leq \mathcal{S}(i,j) \tag{2}$$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{S}(i,j) + \mathcal{P}(i,t) - 1 \leq \mathcal{P}(j,t) \tag{3}$$

$$\forall i \in \mathcal{N} : \sum_{t=1...|\mathcal{N}|} \mathcal{P}(i,t) = 1 \tag{4}$$

$$\forall t = 1...|\mathcal{N}| : \sum_{\forall i \in \mathcal{N}} w(i) \cdot \mathcal{P}(i,t) \leq Load_{max} \tag{5}$$

$$\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}, i \neq j} h(i,j)(1 - \mathcal{S}(i,j)) < \mathcal{F}(\mathcal{S}) \tag{6}$$

$$\forall i,j \in \mathcal{N} : \mathcal{S}(i,j) \in \{0,1\} \tag{7}$$

$$\forall i \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) \in \{0,1\} \tag{8}$$

where $Load_{max}$ represents the maximum load supported by a SGW-C. This load could represent the maximum number of UE flows to handle in order to maintain an acceptable rule installation latency on SGW-Us. The objective (1) minimizes as much as possible the SGW relocation, whereas the constraints are used to ensure the following conditions:

- Constraint (2) guarantees that two areas are not managed by the same SGW-C, if $\mathcal{S}((i,j) = 0$. Formally, $\mathcal{S}((i,j) = 0 \Rightarrow \forall t = 1...|\mathcal{N}| : \mathcal{P}((i,t) = 0 \vee \mathcal{P}((j,t) = 0$.
- Constraint (3) ensures that two areas should be managed by the same SGW-C if $\mathcal{S}((i,j) = 1$. Formally, $\mathcal{S}((i,j) = 1 \Rightarrow \forall t = 1...|\mathcal{N}| : \mathcal{P}((i,t) = \mathcal{P}((j,t)$.
- Constraint (4) guarantees that one area is managed only by one SGW-C.
- Constraint (5) permits to maintain the load on each SGW-C under certain threshold, which could be fixed by the mobile network operator.
- Constraint (6) allows minimizing the number of SGW relocations.
- Constraints (7) and (8) ensure that the matrices $\mathcal{S}$ and $\mathcal{P}$ are binary.

B. Minimizing the load on the SGW-C

In the second optimisation model, we aim at minimizing the load on SGW-C, while maintaining the SGW relocation frequency under a fixed threshold. By using the same notation as in the previous model, we formulate this problem using an ILP as follows:

$$\mathbf{min}\, \mathcal{G}(\mathcal{S},\mathcal{P}) \tag{9}$$

$S.t,$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) + \mathcal{P}(j,t) - 1 \leq \mathcal{S}(i,j) \tag{10}$$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{S}(i,j) + \mathcal{P}(i,t) - 1 \leq \mathcal{P}(j,t) \tag{11}$$

$$\forall i \in \mathcal{N} : \sum_{t=1...|\mathcal{N}|} \mathcal{P}(i,t) = 1 \tag{12}$$

$$\forall i,j \in \mathcal{N} : \mathcal{S}(i,j) \in \{0,1\} \tag{13}$$

$$\forall i \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) \in \{0,1\} \tag{14}$$

$$\sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}, i\neq j} h(i,j)(1-\mathcal{S}(i,j)) < Reloc_{max} \tag{15}$$

$$\forall t = 1...|\mathcal{N}| : \sum_{\forall i\in\mathcal{N}} w(i)\cdot\mathcal{P}(i,t) \leq \mathcal{G}(\mathcal{S},\mathcal{P}) \tag{16}$$

where $Reloc_{max}$ indicates the maximum number of tolerated SGW relocations. This information could be derived by the network operator according to its own objective. The objective in this optimization problem minimizes as much as possible the load on different SGW-C. In addition, the first five constraints are identical to those defined for the first model. The main differences are:

- Constraint (15) permits to maintain the SGW relocation under a certain defined threshold (i.e. by the network operator).
- Constraint (16) allows to minimize the number of SGW relocations.

C. Pareto-optimal solution

As indicated earlier, we will use Game Theory, specifically Nash bargaining game and threat value model, to derive the Pareto optimal solution – named Fair and Optimal SGW-C placement in $5G$ Network (FOSNet) – that satisfies both objectives in the same time; i.e. minimizing the SGW relocation and traffic load on each SGW-C.

Nash bargaining game is a cooperative game with non-transferable utility. In the envisioned model, SGW-C relocation and traffic load are two players in the game. To find the Pareto optimal solution, a fair and reasonable point need to be found, which satisfies both players. The optimal point that finds the optimal trade-off between minimizing the SGW relocation and the traffic load, is obtained by solving the following non-convex optimization problem:

$$\mathbf{max}(Reloc_{worst} - \mathcal{F}^*(\mathcal{S},\mathcal{P}))\cdot(Load_{worst} - \mathcal{G}^*(\mathcal{S},\mathcal{P})) \tag{17}$$

$S.t,$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) + \mathcal{P}(j,t) - 1 \leq \mathcal{S}(i,j) \tag{18}$$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{S}(i,j) + \mathcal{P}(i,t) - 1 \leq \mathcal{P}(j,t) \tag{19}$$

$$\forall i \in \mathcal{N} : \sum_{t=1...|\mathcal{N}|} \mathcal{P}(i,t) = 1 \tag{20}$$

$$\sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}, i\neq j} h(i,j)(1-\mathcal{S}(i,j)) < \mathcal{F}^*(\mathcal{S},\mathcal{P}) \tag{21}$$

$$\forall t = 1...|\mathcal{N}| : \sum_{\forall i\in\mathcal{N}} w(i)\cdot\mathcal{P}(i,t) \leq \mathcal{G}^*(\mathcal{S},\mathcal{P}) \tag{22}$$

$$\forall i,j \in \mathcal{N} : \mathcal{S}(i,j) \in \{0,1\} \tag{23}$$

$$\forall i \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) \in \{0,1\} \tag{24}$$

$$\mathcal{F}^*(\mathcal{S},\mathcal{P}) \leq Reloc_{worst} \tag{25}$$

$$\mathcal{G}^*(\mathcal{S},\mathcal{P}) \leq Load_{worst} \tag{26}$$

The Nash bargaining game requires the threat point, which represent the utility of different players that fail to achieve an agreement. For more theoretical background on Bargaining games and how to find the optimal point, interested readers may refer [14] [15]. The threat point in FOSNet is defined as the point $(Reloc_{worst}, Load_{worst})$ that represents the worst utility values of both players. $Rloc_{worst}$ represents the worst value of the player SGW-C relocation, while $Load_{worst}$ represents the worst value of the SGW-C traffic load player. To find the values of $Reloc_{worst}$ and $Load_{worst}$, we propose to use the two first ILP. Indeed, the first ILP has as objective to minimize the SGW relocation frequency, while maintaining the traffic load under a certain value. Therefore, solving this problem will derive best value for the SGW relocation, while achieving the worst value for traffic load. Hence, we may use the obtained traffic load value as the threat point for the second player. The same reasoning is done to fix the value of $Reloc_{worst}$.

Let $\mathcal{P}_{Load}$ and $\mathcal{S}_{Reloc}$ be the obtained matrices by resolving the optimization problems (1)...(8) and (9)...(16), respectively. Formally, $Load_{worst}$ can be defined as $Load_{worst} = \mathbf{max}_{\forall t=1...|\mathcal{N}|} \sum_{\forall i\in\mathcal{N}} w(i)\cdot\mathcal{P}_{Load}(i,t)$, while $Reloc_{worst}$ can be defined as follow: $Reloc_{worst} = \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}, i\neq j} h(i,j)(1-\mathcal{S}_{Reloc}(i,j))$.

Following the same approach as in [13], the optimization problem in (3) could be transformed to a convex-optimization problem without changing the solution. The main idea consists in introducing the Log function, which is an increasing function. Accordingly, we obtain:

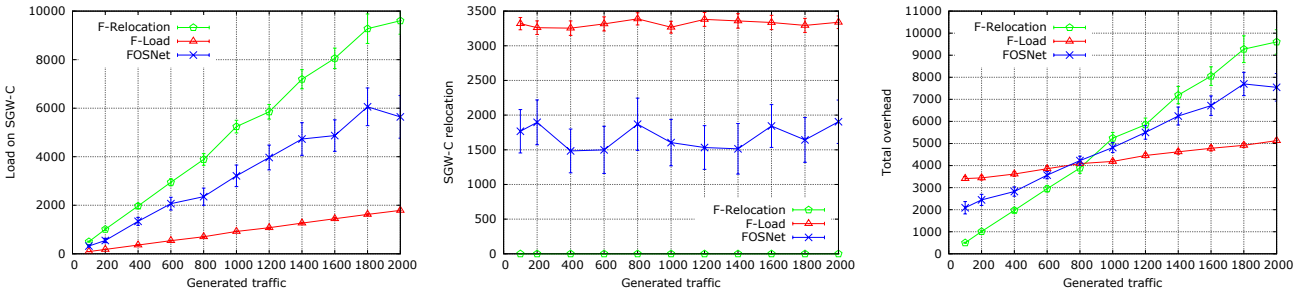
$$\mathbf{max}\log(Reloc_{worst} - \mathcal{F}^*(\mathcal{S},\mathcal{P})) + \log(Load_{worst} - \mathcal{G}^*(\mathcal{S},\mathcal{P})) \tag{27}$$

$S.t,$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) + \mathcal{P}(j,t) - 1 \leq \mathcal{S}(i,j) \tag{28}$$

$$\forall i,j \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{S}(i,j) + \mathcal{P}(i,t) - 1 \leq \mathcal{P}(j,t) \tag{29}$$

$$\forall i \in \mathcal{N} : \sum_{t=1...|\mathcal{N}|} \mathcal{P}(i,t) = 1 \tag{30}$$



(a) (b) (c)

Fig. 2: Performance of the three solutions versus traffic load.

$$\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N},i\neq j} h(i,j)(1-\mathcal{S}(i,j)) < \mathcal{F}^*(\mathcal{S},\mathcal{P}) \tag{31}$$

$$\forall t = 1...|\mathcal{N}| : \sum_{\forall i\in\mathcal{N}} w(i)\cdot\mathcal{P}(i,t) \leq \mathcal{G}^*(\mathcal{S},\mathcal{P}) \tag{32}$$

$$\forall i,j \in \mathcal{N} : \mathcal{S}(i,j) \in \{0,1\} \tag{33}$$

$$\forall i \in \mathcal{N}, t = 1...|\mathcal{N}| : \mathcal{P}(i,t) \in \{0,1\} \tag{34}$$

$$\mathcal{F}^*(\mathcal{S},\mathcal{P}) \leq Reloc_{worst} \tag{35}$$

$$\mathcal{G}^*(\mathcal{S},\mathcal{P}) \leq Load_{worst} \tag{36}$$

To derive the optimal value satisfying both players (i.e. Pareto optimal), we should solve the optimization problem defined in (27...36).

IV. Performances evaluation

A. Scenario

In this section, we compare the performance of *FOSNet* to both solutions *F-Relocation* and *F-Load*. The solution *F-Relocation*, modeled through the Integer Linear Program (1)...(8), favors the relocation of SGW-C overhead over the load overhead. The solution *F-Load*, modeled through the Integer Linear Program (9)...(16), promotes the load overhead of SGW-C over the relocation overhead. We evaluate the proposed solutions in terms of the following metrics:

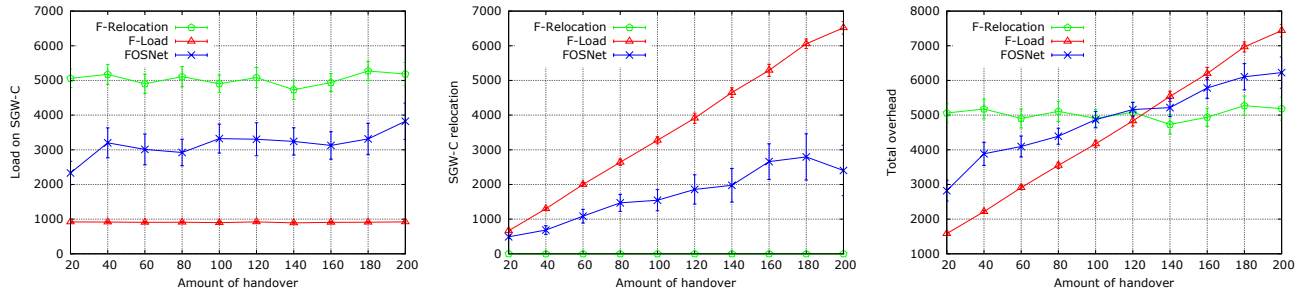
1) *Load on SGW-C:* the overhead of the load on SGW-C of each solution.
2) *SGW-C relocation:* the overhead of SGW-C relocation of each solution.
3) *Total overhead:* the generated overhead due to both load and relocation of SGW-C. The aim of this metric is to show the achieved Pareto-efficiency of the proposed solution.

The three solutions are evaluated by deploying the different areas using uniform distribution. In the simulations, we fixed the number of deployed areas $\mathcal{N}$ to 200. We model the mobility of different UEs through Random Waypoint Mobility Model, which will be reflected to the number of handover between each couple of areas. Both solutions *F-Relocation* and *F-Load* are implemented through Python and Gurobi, a tool for solving Integer Linear Program, while *FOSNet* is implemented through a Matlab and CVX (a package for disciplined convex optimization and geometric programming) [16]. The three solutions are evaluated by conducting two sets of experiments. Firstly, we vary the rate of generated traffic while fixing the rate of handover to 100 handovers/min between each area. Secondly, we vary the rate of handover while fixing the rate of generated traffic per area to 1000 Packets/min.

B. Results

Fig. 2 illustrates the performance of the three solutions versus the generated traffic (i.e. generated by UE). The considered metrics are: load on SGW-C (Fig. 2a), SGW relocation (Fig. 2b) and total overhead (Fig. 2c). Regarding the load on SGW-C, we clearly observe that F-Load solution outperforms the other solutions as its aim is to minimize the traffic load on each SGW-C. For SGW-C relocation (Fig. 2b) the worst case is achieved by F-Relocation, while the solution based on Game Theory obtains acceptable performances. Regarding the SGW relocation, we remark that the best solution is F-Relocation, which guarantees the best results in comparison to the other solutions. This is expected as its objective is to reduce the number of SGW relocations in the network. In this solution, the derived number of SGW-C is one, which leads to not observe any SGW relocation. We also observe from Fig. 2b that the SGW-C relocation remains constant for the three solutions, as the SGW-C relocation is triggered by the number of UE handovers, which is constant in this scenario. Finally, the total overhead figure (Fig. 2c) clearly indicates the ability of FoSNET to find a fair trade-off between relocation and traffic load on SGW-C. Indeed, we remark that from certain point (i.e., 600) the best performance is achieved by the FoSNET. Here, we argue that F-load ensures best results before 600 due to the fact that the total overhead depends only on traffic load as the handover is constant.

Fig. 3 depicts the performance of the three solutions for different numbers of handovers using the same metrics as in Fig. 2. We observe the same trend as in Fig. 2; F-Load achieves the best results for traffic load, while F-Relocation obtains the best results in terms of SGW relocation. Moreover, we observe in Fig.3c that FoSNET ensures the best results after that the



(a) (b) (c)

Fig. 3: Performance of the three solutions for different numbers of handovers.

number of handover exceeds 120 handover/min. This is mainly due to its ability to obtain a fair trade-off between minimizing SGW Relocation and minimizing the load on SGW-C.

V. Conclusion

In this paper, we devised a new algorithm for the gateway controller placement in a SDN-based virtual mobile network. The proposed algorithm finds a fair trade-off between reducing the SGW relocation, which is costly for mobile operators, and reducing the load on the SGW-C that permits to reduce the flow installation latency. To find this trade-off (i.e. Pareto optimal), we relied on Game Theory, and particularly Bargaining game, which derives the threat points and solves the problem. The simulation results showed the ability of the Game Theory based approach to derive a solution that enforces the above-mentioned trade-off.

Aknowledgement

This work was supported in part by the European Unions Horizon 2020 research and innovation programme under the 5G!Pagoda project with grant agreement No. 723172

References

[1] T. Taleb, M. Corici, C. Parada, A. Jamakovic, S. Ruffino, G. Karagiannis, and T. Magedanz, "EASE: EPC as a Service to Ease Mobile Core Network," in IEEE Network Magazine, Vol. 29, No. 2, Mar. 2015. pp.78 - 88.

[2] F.Z. Yousaf, P. Loreiro, F. Zdarsky, T. Taleb, and M. Leibsch, Cost Analysis of initial deployment strategies of a Virtual Network Infrastructure in a Datacenter, in IEEE Communications Magazine, Vol. 53, No. 12, Dec. 2015, pp. 60 - 66.

[3] T. Taleb, "Towards Carrier Cloud: Potential, Challenges, & Solutions,"in IEEE Wireless Communications Magazine, Vol. 21, No. 3, Jun. 2014. pp. 80-91.

[4] 3GPPP TR 23.714, "Study on control and user plane separation of EPC nodes", http://www.3gpp.org/DynamReport/23714.htm.

[5] X. jiny et al., "SoftCell: scalable and flexible cellular core network architecture", Proceedings of the ninth ACM conference on Emerging networking experiments and technologies, Conext'13, New york, USA, 2013.

[6] A. Braddai et al., "Cellular software defined networking: a framework", IEEE Communications Magazine, Vol.: 53, Issue: 6, pp: 36-43, June 2015.

[7] C. J. Bernardos et al., "An architecture for software defined wireless networking", IEEE Wireless Communications, Vol.: 21, Issue: 3, pp: 52-61, June 2014.

[8] A. Basta et al., "A Virtual SDN-Enabled LTE EPC Arch12itecture: A Case Study for S-/P-Gateways Functions", in Proc. of IEEE SDN for Future Networks and Services (SDN4FNS), Trento, Italia, 2013.

[9] S. Ben Hadj Said et al., "New Control Plane in 3GPP LTE/EPC Architecture for On-Demand Connectivity Service", in Proc. of IEEE 2nd International Conference on Cloud Networking (CloudNet), San Francisco, USA, 2013.

[10] X. An, "Virtualization of Cellular Network EPC Gateways based on a scalable SDN Architecture", in Proc. of IEEE Global Communications Conference (Globecom), Austin, USA, 2014.

[11] V. Nguyen and Y. Kim, "Proposal and evaluation of SDN-based mobile packet core networks", EURASIP Journal on Wireless Communications and Networking, December 2015.

[12] S. Shanmugalingam, P. Bertin "Programmable Mobile Core Network", in Proc. of IEEE Symposium on Computers and Communications (ISCC), Funchal, Portugal, 2014.

[13] Y. Zhao et al., "Load Balance vs energy efficiency in traffic engineering: A game theroretical perspective", in Proc. ofThe 32nd IEEE International Conference on Computer Communications (Infocom), Turin, Italy, 2013.

[14] M. Bagaa et al. "Efficient Tracking Area Management Framework fo 5G networks", to appear in IEEE Transactions on Wireless Communications (TWC).

[15] M. Bagaa et al. "User Mobility-aware Wirtual Network function Placement for Virtual 5G Network Infrastructure", in Proc. of IEEE International Conference on Communications (ICC), London, UK 2015.

[16] Michael Grant and Stephen Boyd "Graph implementations for non smooth convex programs", Recent Advances in Learning and Control, Recent Advances in Learning and Control, 95–110, 2008.