Virtualized Local Core Network Functions Placement in Mobile Networks

Jad Oueis, Razvan Stanica, Fabrice Valois

Virtualized Local Core Network Functions Placement in Mobile Networks

Jad Oueis*, Razvan Stanica*, Fabrice Valois*
*Univ Lyon, INSA Lyon, Inria, CITI, F-69621 Villeurbanne, France

***Abstract*—A novel trend in mobile networks is to co-locate the base stations with virtualized core network functions, such as session management and routing. The goal is to lose the long-standing physical dependency between the radio access and the core network, and improve network resiliency. In this work, we focus on the placement of virtualized core functions within a network of multiple base stations interconnected via a potentially limited backhaul. Since all data and signaling traffic are exchanged on the links interconnecting the base stations, the placement of these functions deeply impacts the backhaul load. We compare centralized and distributed placement strategies, with respect to the overall backhaul bandwidth consumption. Results show that distributing instances of the core functions (e.g., routing) in the network is significantly less costly from a backhaul point of view, and can economize backhaul consumption by 86%.**

I. Introduction

Recent technology advances in microelectronics allowed reducing the size and weight of wireless network equipment [1]. On the other hand, the advances in virtualization techniques allowed easier virtualization of different network functions, including the mobile core network (CN) functions [2]. The combination of those two concepts instigated the development of lightweight, easy to move base stations (BSs), that can be co-located with virtualized CN functions, and application servers. The result is a stand-alone self-contained BS, with both radio access (e.g., radio signal processing, radio resource management) and CN capabilities (e.g., routing, authentication, session management) [3]. Such a BS is capable of autonomously providing network coverage and local services to users in its vicinity, depending on the hosted application servers, even without connectivity to external networks. Moreover, the BS can establish a backhaul connection to an external packet data network (PDN). The set of virtualized network functions (VNFs), providing the functionalities of a traditional CN and co-located with the BS, are referred to as Local CN [4]. To cover larger areas, several BSs must interconnect and form a network. With the Local CN functions co-located with the BSs, the network interconnecting these BSs constitutes the backhaul, as shown in Fig. 1.

In this work, we focus on the Local CN functions placement in such networks. The goal is to determine which BS(s) hosts which functions of the Local CN, and how many instances of each function are needed. In other words, we answer the following question: do we locate the Local CN functions with each BS, a subset of them, or only one? The placement problem faces a number of constraints, mostly related to the backhaul network dimensioning, and its potentially limited bandwidth [4]. Indeed, all traffic, whether data or signaling, exchanged between a BS and the Local CN functions, is routed locally on the inter-BS backhaul links. When their bandwidth is limited, these links can represent a bottleneck by limiting the amount of traffic that can be exchanged between the BS and the Local CN. Hence, the Local CN functions must be carefully placed to avoid backhaul saturation.

The contributions of this paper are the following. We compare both centralized and distributed Local CN placement approaches. For each approach, the functions are optimally placed, with the objective of minimizing the backhaul bandwidth consumption caused by the user data and signaling traffic exchanged among the BSs and the different Local CN entities. For the distributed placement, we further study the number of needed instances for each function, in addition to their optimal placement. We formulate and solve the underlying optimization problems, then use the overall backhaul consumption as evaluation metric for each of the approaches. Moreover, we assess the impact of the signaling traffic on the backhaul bandwidth consumption. Following extensive simulations, we conclude that distributing the Local CN functions in the network, with each of the BSs co-located with at least one function instance, largely outperforms a centralized Local CN. For example, results show that co-locating each BS with its own routing function would cause a significant economy in the backhaul consumption, reducing the latter by around 86% with respect to having only one default routing function in the network.

The paper is organized as follows. We discuss related works in Sec. II, and describe the network model in Sec. III. The placement problem relevance is discussed in Sec. IV, before formulating it in Sec. V. Results are discussed in Sec. VI, followed by concluding remarks in Sec. VII.

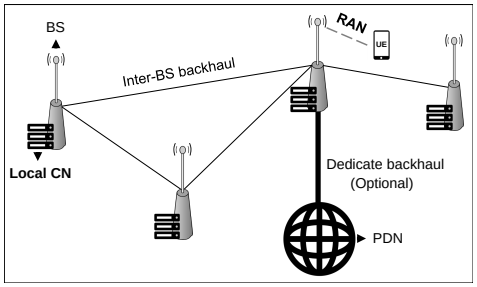

Fig. 1. Network architecture with BSs co-located with Local CN functions.

II. Related Work

Mobile operators are increasingly interested in exploring the benefits of CN virtualization through VNFs, to decrease the network operation and deployment costs, while increasing its scalability and flexibility [5]. In this context, the optimal placement of the VNFs has been studied, whether across federated clouds (i.e., geographically distributed and interconnected data centers) [6] or within the same data center [2]. However, the objectives of the placement problem in such networks are different from our case, where VNFs are actually co-located with the BSs and not in distant data centers. First, the two problems scale differently: an operator virtualized CN is usually designed at the scale of a country to serve a large number of customers, while BSs co-located with a Local CN serve a limited area and few users [4]. Second, the constraints in both problems are different. The main limiting factor in our case is the limited backhaul bandwidth interconnecting the BSs. Nevertheless, this problem is not necessarily relevant in a classical operator network with a dedicated over-provisioned backhaul, where the focus is mostly on runtime management [6], and minimizing computing and networking resources consumption within the data center(s) [2].

In a previous work, the limited backhaul bandwidth served us as a key criterion for placing the Local CN within a network [7]. However, that work focused on a centralized placement, by considering that all the Local CN functions are co-located with the same BS. Furthermore, user requests were ignored, and the centralized placement was determined under an assumption of demand equity among the BSs [7]. In this paper, we further study a distributed Local CN placement, by allowing functions to be placed on multiple BSs in the network, and compare it to the centralized approach. Moreover, we tackle the problem form a different point of view, by explicitly taking into consideration the number of users in the network, their distribution among the BSs, and their requests.

III. System model

We consider a mobile network based on the architecture in Fig. 1, with Local CN functions co-located with the BSs and an inter-BS backhaul. Let $\mathcal{J}$ be the set of BSs, $\mathcal{L}$ the set of directional inter-BS links, and $\mathcal{U}$ the set of user equipments (UEs). We consider that, regardless of the used technology, there is no contention between the backhaul links for resource utilization. We assume that potentially interfering links are operating on distinct channels, allowing interference-free parallel transmissions on the backhaul [8].

A. Local core network

We focus on two types of functions of the Local CN: those classically implemented by the mobility management entity (MME) and those of the serving gateway (S-GW). The MME handles network management such as paging, authentication, session management and gateway selection. Since these functions are difficult to distribute and the network size is limited, we consider that network management is ensured by a single MME entity, co-located with one of the BSs. The S-GW mainly handles local data routing. We consider that each BS of the network has the possibility to be co-located with a S-GW, and we focus on the distribution of this routing function in the following. All signaling traffic passes through the MME, and all data traffic passes through the S-GW. The two entities also exchange signaling traffic. In addition to associating to a BS, each UE in the network also attaches to the Local CN, more specifically to an MME and a S-GW. One of the objectives of the attachment is to assign the flows of each user to a specific S-GW, responsible of locally routing all of the UE flows [9].

B. Traffic model

1) Data traffic: We adopt a data traffic model consisting of bidirectional flows between two parties, that is two directional flows, one in each direction. Flows can be intra-network (i.e., between two parties belonging to the same network) or inter-network (i.e., one of the parties belong to another network). For brevity, the given examples, notations, and the subsequent numerical applications are limited to intra-network flows between two UEs. However, we note that the problem formulation is general enough to include both intra and inter-network flows, as well as flows between a UE and any application server. Let $\mathcal{F}$ be the set of directional flows. For each directional flow from UE u to UE v in $\mathcal{F}$, denoted $f = \{u, v\}$, there exists a flow $f' = \{v, u\} \in \mathcal{F}$ in the opposite direction. We denote by d_f the requested data rate of a flow f, in bits/second. One UE can have several simultaneous flows.

2) Signaling traffic: We consider two main signaling types: between the MME and the BS to which the UE is associated, and between the MME and the S-GW to which the UE is attached. Quantifying the amount of signaling traffic in a network depends on the particular scenario in question, the users' activity, and the number, size and frequency of the exchanged signaling messages.While some signaling procedures have a minimum imposed frequency, such as tracking area update, others are timely, such as paging and session management [9] Due to the lack of a thorough signaling traffic quantification model, and to avoid limiting the study to a particular use case, we consider that each data flow f requires a signaling traffic of bit rate Si_f. For our numerical results, we consider Si_f represents a percentage σ of the flow data rate d_f: $Si_f = \sigma \cdot d_f$.

C. Routing

Depending on the network topology, data and signaling traffic are routed on the inter-BS links either directly, if the two end-BSs are at one hop from each other, or through the interconnected BSs, otherwise. In the latter case, a routing policy is needed. We define $Z^l_{j,j'}$ as an indicator function, such that $Z^l_{j,j'} = 1$ if link $l \in \mathcal{L}$ belongs to the routing path between BSs j and j'. We assume that, for all bidirectional flows, the route is the same in both directions.

D. Radio access network

We consider an OFDMA-based system, with orthogonal channels allocated to the set of BSs. Distinct channels are reserved for each BS on the downlink (DL) and the uplink

(UL), with their respective numbers denoted $\mathcal{K}_j^{DL}$ and $\mathcal{K}_j^{UL}$. We denote by $R_{u,j}^{DL}$ and $R_{u,j}^{UL}$ the per channel rates seen by UE u from BS j, on the DL and the UL, respectively. A physical RAN model similar to the one described in [10] is adopted. We assume that, for each of its flows, a UE is granted a fraction of the channels available on the BS, depending on the rate it gets from that BS, and on the requested throughput.

IV. Placement Problem Overview

In a network where BSs are co-located with Local CN functions, the placement of the S-GW to which a user attaches determines the data traffic routing path on the backhaul. Likewise, the placement of the MME handling all signaling procedures affects the signaling traffic routing path. Fig. 2 illustrates in a simple example how the placement of the Local CN functions can impact the traffic on the backhaul, and the consequent bandwidth consumption.

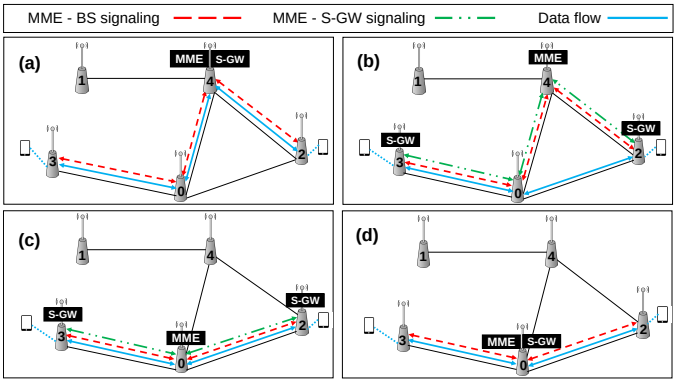

Fig. 2. An example on the different data and signaling traffic paths for a flow between two UEs, in scenarios with different Local CN functions placement.

The four cases in Fig. 2 correspond to a flow between two UEs associated to BSs 2 and 3, respectively, accompanied by the two types of signaling traffic: between the BSs and the MME, and between the MME and the S-GW(s). When the MME and the S-GW are co-located, signaling traffic between them is not routed on the backhaul (cases (a) and (d)), further economizing bandwidth. When the S-GWs are co-located with the BSs to which the UEs are associated (case (b) and (c)), or belong to the shortest path between the BSs (case (d)), the flow takes a shortest path between the two BSs. Hence, it consumes less backhaul bandwidth than the one having to go through a S-GW that does not belong to the shortest path between the two BSs (case (a)). The same can be said about the MME and the resulting signaling traffic path: the shortest the path between the BSs and the MME, the less is the MME-BS signaling backhaul consumption (cases (c) and (d)). While this example shows the trade-offs for a particular flow, the placement must actually take into account the overall traffic in the network, i.e., all the flows between the different BSs and the total consumption incurred by signaling and data.

In this work, our goal is to compare the backhaul bandwidth consumption for different optimal placement schemes. To that end, we compare the three following strategies:

i) $\mathcal{P}_\mathbf{1}$: there is one and only one S-GW co-located with one BS. The raised question here is whether this S-GW should be co-located with the MME (i.e., a centralized Local CN), or if the optimal placements of those two entities are different;

ii) $\mathcal{P}_\mathcal{J}$: there are $|\mathcal{J}|$ S-GWs, as many as the BSs in the network, such that each BS is co-located with an S-GW;

iii) $\mathcal{P}_\mathbf{o}$: there are multiple S-GWs distributed in the network. The number of the S-GWs is optimized to determine whether a routing function should be placed on all the BSs of the network, or only on a subset.

The placement problem is closely related to user attachment. With only one S-GW placed in the network ($\mathcal{P}_\mathbf{1}$), UEs have no choice but attaching to that S-GW. However, when multiple S-GWs are present, an attachment policy must be defined. In $\mathcal{P}_\mathcal{J}$, where all BSs are forcibly co-located with a S-GW, we consider that each UE attaches to the S-GW co-located with the BS it is associated to. In $\mathcal{P}_\mathbf{o}$, we optimize the attachment such that each UE can attach to any S-GW (not necessarily the one on the BS it is associated to) in a way that minimizes the backhaul bandwidth consumption. This allows us to deduce the number of needed S-GWs and their distribution.

V. Problem Formulation

To compare the three aforementioned placement schemes, we formulate three mixed integer linear programming (MILP) problems. These problems return, for the three placement strategies discussed in Sec.IV, the optimal MME placement and, if needed, the optimal S-GW(s) placement, with the objective of minimizing the backhaul bandwidth consumption.

Each BS of the network must be served by an MME co-located with one of the BSs. We define vector W, such that $W_j = 1$ if the MME is co-located with BS j and $W_j = 0$ otherwise. Vector W is an output in all the problems. Moreover, each UE must associate to one and only one BS. Hence, we define vector X, such that $X_{u,j} = 1$ if and only if UE u is associated to BS j. Vector X is an input, with $X_{u,j}$ known $\forall u \in \mathcal{U},\ j \in \mathcal{J}$.

A. $\mathcal{P}_1$: *One default S-GW in the network*

In this scenario, only one S-GW exists, and we want to optimally place it in the network in a way that minimizes backhaul bandwidth consumption. We define vector G as an output of our problem, such that $G_j = 1$ if and only if the S-GW is co-located with BS j. No attachment decision is needed in this case, since all UEs attach to the only available S-GW. To better illustrate the problem in this scenario, Fig. 3 shows the data and signaling traffic paths for a flow between UEs u and v, respectively associated to BSs j_1 and j_2, with the MME at j_0 and the S-GW at j_5. We denote by d_f the data flow rate between UEs u and v, and by $S1_f$ and $S2_f$ the signaling traffic rates between the MME, on the one side, and BSs j_1 and j_2, on the other side. $S5_f$ is the signaling traffic rate between the MME and the S-GW. Recall that the signaling rates Si_f, accompanying flow f, are proportional to the flow data rate d_f.

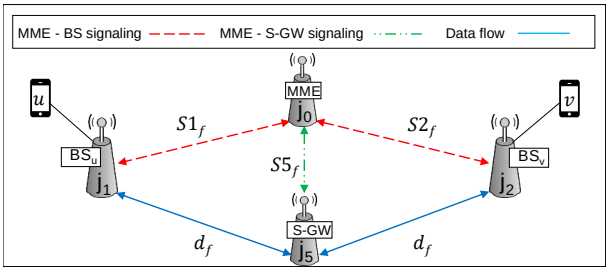

Fig. 3. Data and signaling traffic paths, and their corresponding bit rates d_f and Si_f, for a flow f, when there is one S-GW in the network ($\mathcal{P}_1$).

In this case, we define the following minimization problem:

$$\min \sum_{l \in \mathcal{L}} C_l \qquad s.t. \tag{1}$$

$$\sum_{j \in \mathcal{J}} W_j = 1 \tag{2}$$

$$\sum_{j \in \mathcal{J}} X_{u,j} = 1\ , \quad \forall u \in \mathcal{U} \tag{3}$$

$$\sum_{u \in \mathcal{U}} \frac{X_{u,j}}{\mathcal{K}_j^{DL} \cdot R_{u,j}^{DL}} \sum_{f \in \mathcal{F}/u \in f} d_f \leq 1\ , \quad \forall j \in \mathcal{J} \tag{4}$$

$$\sum_{u \in \mathcal{U}} \frac{X_{u,j}}{\mathcal{K}_j^{UL} \cdot R_{u,j}^{UL}} \sum_{f \in \mathcal{F}/u \in f} d_f \leq 1\ , \quad \forall j \in \mathcal{J} \tag{5}$$

$$C_l^{S1} = \sum_{f \in \mathcal{F}} \sum_{j_1 \in \mathcal{J}} X_{u,j_1} \sum_{j_0 \in \mathcal{J}} W_{j_0} \left(Z_{j_1,j_0}^l + Z_{j_0,j_1}^l \right) S1_f, \quad \forall l \in \mathcal{L} \tag{6}$$

$$C_l^{S2} = \sum_{f \in \mathcal{F}} \sum_{j_2 \in \mathcal{J}} X_{v,j_2} \sum_{j_0 \in \mathcal{J}} W_{j_0} \left(Z_{j_2,j_0}^l + Z_{j_0,j_2}^l \right) S2_f, \quad \forall l \in \mathcal{L} \tag{7}$$

$$\sum_{j \in \mathcal{J}} G_j = 1 \tag{8}$$

$$C_l^d = \sum_{j \in \mathcal{J}} G_j \sum_{f \in \mathcal{F}} \Big(\sum_{j_1 \in \mathcal{J}} X_{u,j_1} Z_{j_1,j}^l + \sum_{j_2 \in \mathcal{J}} X_{v,j_2} \cdot Z_{j,j_2}^l \Big) d_f, \quad \forall l \in \mathcal{L} \tag{9}$$

$$C_l^{S5} = \sum_{f \in \mathcal{F}} \Big(\sum_{j \in \mathcal{J}} G_j \sum_{j_0 \in \mathcal{J}} W_{j_0} \cdot Z_{j_0,j}^l \cdot S5_f \Big)\ , \quad \forall l \in \mathcal{L} \tag{10}$$

$$C_l = C_l^d + C_l^{S1} + C_l^{S2} + C_l^{S5} \tag{11}$$

The bandwidth consumed on a backhaul link l by all the flows is denoted C_l, computed in Eq. 11. C_l is the sum of the bandwidth consumed by all data and signaling traffic on that link. Our objective is to minimize the total bandwidth consumed on the backhaul, on all links by all flows, as formulated in the objective function in Eq. 1. Constraints in Eq. 2 and Eq. 3 state that there is one and only one MME in the network, and that a UE is associated to one and only one BS, respectively. Constraints in Eq. 4 and Eq. 5 state that the total flows received from UEs associated to a BS on the DL should not exceed the DL BS capacity, and the total flows sent by UEs associated to a BS on the UL should not exceed the UL BS capacity, respectively. Eq. 8 states that there is one and only one S-GW in the network.

The data path of a flow f goes from j_1 to j_2, passing through the only S-GW at j_5 (Fig. 3). In this case, a rate d_f is consumed by f on each link l on the routing path between j_1 and j_5, and between j_2 and j_5, that is, on each link l with $Z_{j_1,j_5}^l = 1$, and each link l with $Z_{j_5,j_2}^l = 1$. Eq. 9 computes the value of C_l^d, which is the total bandwidth consumed by the data traffic of all the flows on a link l.

Signaling bit rates $S1_f$, $S2_f$, and $S5_f$ are consumed on each link l that belongs to the routing path between MME j_0 and BS j_1, MME j_0 and BS j_2, and MME j_0 and S-GW j_5, respectively. In Eq. 6, Eq. 7, and Eq. 10, we compute C_l^{S1}, C_l^{S2}, and C_l^{S5}, respectively representing the total bandwidth consumed by the signaling traffic, of bit rates $S1_f$, $S2_f$, and $S5_f$, accompanying all flows f on a single link l.

B. $\mathcal{P}_\mathcal{J}$: All BSs co-located with S-GWs

In this scenario, each BS in the network is co-located with a S-GW. No attachment decision is needed since user attachment follows user association, such that a UE attaches to the BS to which it is associated on the RAN. The problem only outputs the MME optimal placement.

Fig. 4 shows the data and signaling traffic paths for a flow between UEs u and v, respectively associated to BSs j_1 and j_2, co-located with their corresponding S-GWs. If u (resp. v) is associated to BS j_1 (resp. j_2), then u (resp. v) is attached to S-GW j_1 (resp. j_2). We denote by $S3_f$ and $S4_f$ the bit rates of the signaling traffic between the MME, on the one side, and S-GWs j_1 and j_2, respectively, on the other side.

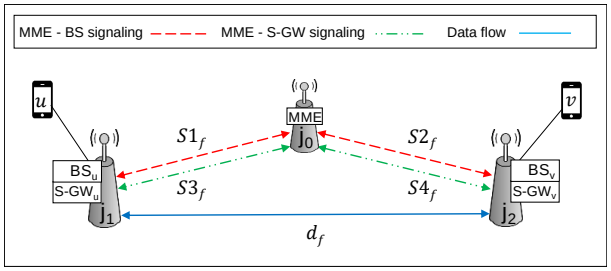

Fig. 4. Data and signaling traffic paths between BSs, and their corresponding bit rates d_f and Si_f, for a flow f, when there are $|\mathcal{J}|$ S-GWs ($\mathcal{P}_\mathcal{J}$).

In $\mathcal{P}_\mathcal{J}$, the objective function is the same as the one in Eq. 1. The constraints from Eq. 2 to Eq. 7 remain unchanged. However, the data traffic cost and the signaling between the MME and the S-GW to which a user is attached are formulated differently. The data path of a flow f goes directly from j_1 to j_2 (Fig. 4). In Eq. 12, we compute C_l^d, the total bandwidth consumed by the data traffic of all the flows f on each link l. In Eq. 13 and Eq. 14 we compute C_l^{S3} and C_l^{S4}, representing the total bandwidth consumed on each link l by the signaling between MME j_0 and S-GW j_1 with a rate $S3_f$, and between MME j_0 and S-GW j_2 with a rate $S4_f$, respectively. Hence, the constraints from Eq. 8 to Eq. 11 are replaced by the following ones:

$$C_l^d = \sum_{f\in\mathcal{F}} \Big(\sum_{j_1\in v} X_{u,j_1} \sum_{j_2\in\mathcal{J}} X_{v,j_2} \cdot Z^l_{j_1,j_2} \cdot d_f \Big) \ , \ \forall l \in \mathcal{L} \tag{12}$$

$$C_l^{S3} = \sum_{f\in\mathcal{F}} \Big(\sum_{j_1\in\mathcal{J}} X_{u,j_1} \sum_{j_0\in\mathcal{J}} W_{j_0} \cdot Z^l_{j_0,j_1} \cdot S3_f \Big) \ , \ \ \forall l \in \mathcal{L} \tag{13}$$

$$C_l^{S4} = \sum_{f\in\mathcal{F}} \Big(\sum_{j_2\in\mathcal{J}} X_{v,j_2} \sum_{j_0\in\mathcal{J}} W_{j_0} \cdot Z^l_{j_0,j_2} \cdot S4_f \Big) \ , \forall l \in \mathcal{L} \tag{14}$$

$$C_l = C_l^d + C_l^{S1} + C_l^{S2} + C_l^{S3} + C_l^{S4} \tag{15}$$

C. $\mathcal{P}_o$: Optimized S-GW placement

In this scenario, each BS can be co-located with a S-GW function. However, each UE can attach to any S-GW, not necessarily the one corresponding to the BS it is associated to on the RAN. In this case, the number of S-GWs with users attached to them, their placement, and the MME placement are all optimized with the objective of minimizing backhaul bandwidth consumption. Fig. 5 shows the data and signaling paths for a flow between u and v, respectively associated to j_1 and j_2, and attached to S-GWs j_3 and j_4, that may or may not be the same as j_1 and j_2.

In $\mathcal{P}_o$, the objective function is the same as in Eq. 1. Likewise, the constraints from Eq. 2 to Eq. 7 remain unchanged, while the constraints from Eq. 8 to Eq. 11 are replaced by:

$$\sum_{j\in\mathcal{J}} Y_{u,j} = 1 \ , \ \ \forall u \in \mathcal{U} \tag{16}$$

$$C_l^d = \sum_{f\in\mathcal{F}} d_f \Bigg(\sum_{j_1\in\mathcal{J}} X_{u,j_1} \sum_{j_3\in\mathcal{J}} Y_{u,j_3} \cdot Z^l_{j_1,j_3} + \sum_{j_4\in\mathcal{J}} Y_{v,j_4} \Big(\sum_{j_3\in\mathcal{J}} Y_{u,j_3} \cdot Z^l_{j_3,j_4} + \sum_{j_2\in\mathcal{J}} X_{v,j_2} \cdot Z^l_{j_4,j_2} \Big) \Bigg) \ , \ \ \forall l \in \mathcal{L} \tag{17}$$

$$C_l^{S3} = \sum_{f\in\mathcal{F}} \Big(\sum_{j_3\in\mathcal{J}} Y_{u,j_3} \sum_{j_0\in\mathcal{J}} W_{j_0} \cdot Z^l_{j_0,j_3} \cdot S3_f \Big) \ , \ \ \forall l \in \mathcal{L} \tag{18}$$

$$C_l^{S4} = \sum_{f\in\mathcal{F}} \Big(\sum_{j_4\in\mathcal{J}} Y_{v,j_4} \sum_{j_0\in\mathcal{J}} W_{j_0} \cdot Z^l_{j_0,j_4} \cdot S4_f \Big) \ , \forall l \in \mathcal{L} \tag{19}$$

$$C_l = C_l^d + C_l^{S1} + C_l^{S2} + C_l^{S3} + C_l^{S4} \tag{20}$$

Since we optimize user attachment, we define the attachment vector Y, such that $Y_{u,j} = 1$ if and only if UE u is attached to S-GW j. Hence, we add Eq. 16, stating that a UE is attached to one and only one S-GW. Since the BS to which a UE is associated and the S-GW to which it is attached are not necessarily co-located, the data path from BS j_1 to BS j_2 passes through S-GW j_3 then S-GW j_4 (Fig. 5). We compute the bandwidth consumption C_l^d caused by the data traffic of all flows on a link l in Eq. 17. In Eq. 18 and 19, we compute C_l^{S3} and C_l^{S4}, which are the costs incurred by all flows on each link l from the signaling between MME j_0 and S-GW j_3, to which one UE is attached, and between MME j_0 and S-GW j_4, to which the other UE is attached, respectively.

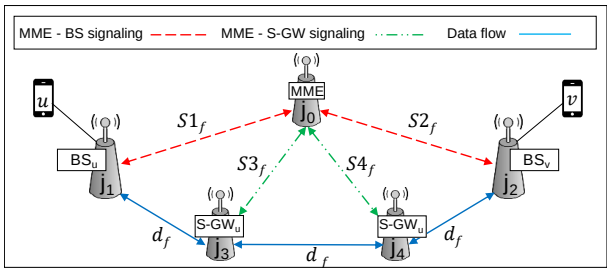

Fig. 5. Data and signaling traffic paths, and their corresponding bit rates d_f and Si_f, for a flow f, when the number of S-GWs is optimized ($\mathcal{P}_o$).

VI. Numerical Results

We consider as an example the network topology in Fig. 6, with 5 BSs deployed in an area of 1 unit square. Tests are conducted on 400 network snapshots. Each snapshot consists of a different combination of a user distribution and a flow distribution, with a total of 35 UEs, each having on average 3 simultaneous bidirectional flows. To study scenarios where the RAN and backhaul are most loaded, we consider that all bidirectional flows are symmetric, such that a flow f and its counterpart in the other direction f' have $d_f = d_{f'} = 1$ Mb/s.

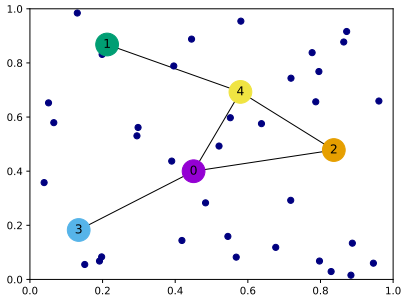

Fig. 6. Network topology.

The MME-BS and MME-S-GW signaling rates, namely $S1_f$, $S2_f$, $S3_f$, $S4_f$, and $S5_f$, are considered equal. As explained in Sec. III, we avoid using pre-defined values of signaling rates to avoid limiting our study to a specific use case. For all Si_f, we vary σ representing the percentage of signaling with respect to the data traffic. For vector X, we adopt an association policy such that each UE associates to the BS from which it gets the maximum signal to interference and noise ratio. For the routing on inter-BS links, we adopt a routing policy such that the shortest path in terms of number of hops is selected between two BSs. All optimization problems are solved with the commercial solver "CPLEX" [11].

For the topology in Fig. 6, the solutions of the three problems return either BS 0 or BS 4 as the optimal placement of the MME (depending on user/flow distribution of the tested snapshot), for all signaling traffic values. For the particular user distribution in Fig. 6, BS 4 is the optimal position of the MME. For all snapshots, the optimal solution of $\mathcal{P}_1$ is to

co-locate the only S-GW in the network with the same BS as the MME. In other words, when there is only one instance of each function, the optimal placement is a centralized one.

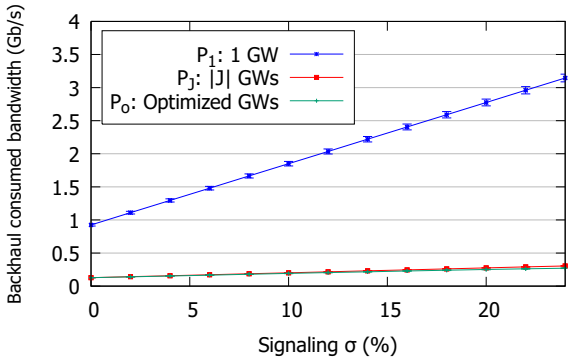

Fig. 7. The total backhaul bandwidth consumption in $\mathcal{P}_1$, $\mathcal{P}_\mathcal{J}$, and $\mathcal{P}_o$

To quantitatively compare the three placement strategies, we show for each of them the total backhaul bandwidth consumption function of the signaling traffic represented by σ, in Fig. 7. We observe that significantly more bandwidth is consumed on the backhaul when only one S-GW is placed in the network ($\mathcal{P}_1$), compared to when all BSs are co-located with S-GWs ($\mathcal{P}_\mathcal{J}$), and when placement is optimized ($\mathcal{P}_o$). This is because data traffic between two BSs is always routed through the S-GW. With only one S-GW in the network, the latter is not necessarily placed on the shortest path between the two BSs, which further increases consumption. This observation is true for all values of signaling traffic represented by σ. Even without signaling traffic, co-locating S-GWs with all the BSs causes a significant economy in the backhaul consumption, reducing the latter by 86% with respect to having one S-GW. These observations confirm that having distributed S-GWs in the network is evidently better than having one S-GW. However, in the distributed scheme, should all BSs be co-located with S-GWs or only a subset?

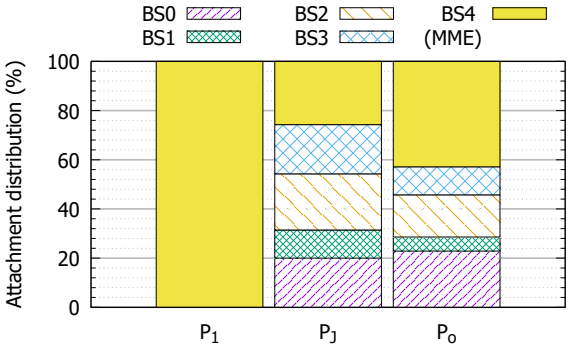

Fig. 8. User attachment distribution in $\mathcal{P}_1$, $\mathcal{P}_\mathcal{J}$, and $\mathcal{P}_o$, for the studied topology and user distribution, with signaling traffic at $\sigma = 6\%$.

From Fig. 7, the answer seems to be that both approaches produce a very similar amount of traffic on the backhaul. To gain further insight, we show in Fig. 8 the user attachment distribution, i.e., the percentage of users attached to each of the BSs, for $\mathcal{P}_1$, $\mathcal{P}_\mathcal{J}$, and $\mathcal{P}_o$. These results correspond to one of the tested network snapshots, with the user distribution shown in Fig. 6, and signaling traffic at $\sigma = 6\%$. We observe that, for $\mathcal{P}_o$, all BSs in the network have users attached to them. This means that, when placement and attachment are optimized, all BSs are co-located with S-GWs, a solution similar to $\mathcal{P}_\mathcal{J}$, which explains the similar results in Fig. 7. However, optimizing the attachment in $\mathcal{P}_o$ changes the user distribution among the BS, with more users attached to BS 4, which also hosts the MME.

These results suggest that, for most users, optimizing attachment is similar to attaching to the S-GW co-located with their BS. The routing function of the Local CN can therefore be easily distributed on all the BS, with negligible loss with respect to an optinal placement.

VII. Conclusion

While virtualizing CN functions in operator networks, with the VNFs running on off-the-shelf servers in distant data centers, is not a novel concept, co-locating these functions in a Local CN with the BS is. In this context, we focused on the Local CN functions placement within a network of multiple interconnected BSs. We compared different placement strategies: centralized, distributed, and optimally distributed. The driving idea behind all strategies is to minimize the backhaul load, since all traffic exchanged between the BSs and the Local CN functions is routed on the inter-BS backhaul links. By evaluating the overall backhaul bandwidth consumption, we showed that distributing S-GWs, such that each BS is co-located with one S-GW, is significantly less costly from a backhaul point of view than having only one centralized Local CN, and performs smilarly to an optimized S-GW placement.

References

[1] H. Mopidevi, D. Rodrigo, O. Kaynar, L. Jofre, B. A. Cetiner, "Compact and broadband antenna for LTE and public safety applications", *IEEE Antennas and Wireless Propagation Letters*, vol. 10, 2011.

[2] F. Yousaf, P. Loureiro, F. Zdarsky, T. Taleb, M. Liebsch, "Cost analysis of initial deployment strategies for virtualized mobile core network functions", *IEEE Comm. Mag.*, 53(12), Dec. 2015.

[3] K. Gomez, L. Goratti, T. Rasheed, L. Reynaud, "Enabling disaster-resilient 4G mobile communication networks", *IEEE Comm. Mag.*, 52(12), Dec. 2014.

[4] J. Oueis, V. Conan, D. Lavaux, R. Stanica, F. Valois, "Overview of LTE isolated EUTRAN Operation for public safety", *IEEE Communications Standards Magazine*, 1(2), July 2017.

[5] A. Baumgartner, V. Reddy, T. Bauschert, "Mobile core network virtualization: a model for combined virtual core network function placement and topology optimization", *Proc. IEEE NetSoft*, Apr. 2015.

[6] T. Taleb, A. Ksentini, "Gateway relocation avoidance-aware network function placement in carrier cloud", *Proc. ACM MSWiM*, Nov. 2013.

[7] J. Oueis, V. Conan, D. Lavaux, H. Rivano, R. Stanica, F. Valois, "Core network function placement in self-deployable mobile networks", *Computer Communications*, vol. 133, Jan. 2019.

[8] B.Aoun, R.Boutaba, Y.Iraqi, G.Kenward, "Gateway placement optimization in wireless mesh networks with QoS constraints", *IEEE JSAC*, 24(11), Nov. 2006.

[9] R. Kreher, K. Gaenger, "LTE Signaling: Troubleshooting and Optimization", *John Wiley & Sons*, 2010.

[10] D. Fooladivanda, C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks", *IEEE Trans. on Wireless Communications*, 12(1), Jan. 2013.

[11] IBM Corp., "IBM ILOG CPLEX Optimization Studio v12.7.0 documentation", 2016.