Trade–offs in Cache–enabled Mobile Networks

Davit Harutyunyan*, Abbas Bradai§ and Roberto Riggio*

*FBK CREATE-NET, Italy; Email: d.harutyunyan,rriggio@fbk.eu
§XLIM Institute, University of Poitiers, France; Email: abbas.bradai@univ-poitiers.fr

***Abstract*—Mobile data traffic demand has been growing at an unprecedented rate in the last few years. Cache–enabled mobile edge computing is known to be one of the most promising techniques to accommodate the traffic demand and alleviate the congestion at the backhaul links. However, due to limited cache capacity at the eNBs, at some parts of the network, congestion of the backhaul links and the radio resources is still possible. Thus, efficient approaches are needed in order to cache content at eNBs as well as to leverage the utilization of the resources in the mobile network while trying to avoid their congestion.**

In this paper, we study the trade–offs between the radio resource utilization and the backhaul link utilization in cache–enabled mobile networks. Initially, we show the trade–offs by formulating a mobility–aware joint content caching, user association, and resource allocation problem as an Integer Linear Programming problem and proposing a greedy heuristic to solve the large instances of the problem. We then propose an approach to compute radio resource and backhaul link costs, and by using the costs, we formulate a joint user association and resource allocation problem aiming at preventing network congestion, assuming that the cached content is given. The results reveal that around 10% more users get an association to the network by using the proposed algorithm.

***Index Terms*—LTE, Mobile Edge Caching, Resource Allocation, User Mobility.**

I. Introduction

Cisco's analysis shows that the global mobile data traffic grew 63% in 2016, ranging from 44% growth in the North America to 96% in the Middle East and Africa. According to its forecast, the global mobile data traffic is expected to increase from 7 exabytes in 2017 up to 49 exabytes in 2021 [1]. Whereas, Ericsson's forecast shows that by 2022 the global mobile data traffic will reach up to 71 exabytes out of which around 75% accounting for video traffic [2]. This traffic increase will take place not only due to a huge number of mobile subscriptions (around one million more subscriptions in 2020 compared to the ones in 2017) but also due to vertical applications such as Vehicle–to–Everything (V2X) communication, e-health, etc.

This galloping trend forces Mobile Network Operators (MNOs) to perform costly upgrades in order to meet this data traffic demand. Cache–enabled Mobile Edge Computing (MEC) is known to be one of the promising ways in order to boost mobile network capacity. Since the video traffic demand will predominate other sorts of mobile traffic demands [2], it is vital to apply MEC techniques in order to save huge backhaul link resources which, otherwise, would be highly loaded [3]. Since local cache capacity at evolved Node Bs (eNBs) is limited, the efficient content (e.g., file, service) caching requires careful considerations. For example, in order to effectively select the content that is to be locally cached at the eNBs, in our opinion, it is important to take into account also users mobility, apart from their content preference.

On the other hand, although caching the high–demanded content undoubtedly curtails the load at backhaul links, one cannot claim that it totally prevents the backhaul link congestion, whose utilization along with the utilization of radio resources[1] depends on the spatio–temporally fluctuating traffic demand, which has been recently snowballing at a rapid pace. Backhaul links and PRBs at some eNBs may be congested while, at the same time, those resources may be underutilized at other eNBs. It is, therefore, of a great importance to have some techniques at disposal for leveraging the PRB utilization and the backhaul link utilization. For example, if the file demand that has been cached at an eNB is very high such that it results in many User Equipments (UEs) being associated with the eNB, leading to a congestion of its PRB utilization, MNOs may decide to associate some of the UEs to less utilized neighbor eNBs that may not have the requested file cached. This would increase the PRB and the backhaul link utilization of the neighbor eNBs while alleviating the load of the former eNB, thereby, providing the possibility of serving more UEs.

With the advent of LTE networks, the focal point for MNOs has been shifted from increasing the data rate experienced by UEs towards more supporting high Quality of Service (QoS). Bringing the content to the mobile networks edge (i.e., closer to users) also plays a pivotal role in enhancing the QoS of UEs by curtailing the round–trip time for service provisioning. Ideally, in order to guarantee a high QoS for the UEs in mobility, MNOs should always find an optimal UE–eNB association and allocate enough PRBs to the UEs. Otherwise with low QoS, the video quality, for example, may degrade. The picture may appear distorted and in a slow motion, while the audio and video portions may be unsynchronized. As a consequence, the subscribers frequently experiencing such problems may decide to change their mobile network provider. In this work, it is our assumption that the host eNBs always allocate PRBs to the UEs enough to meet the throughput demand of the consumed services without degrading their quality. The contribution of this paper is threefold.

- First, assuming that the UEs always achieve their requested throughput, we study the trade–offs between PRB

Research leading to these results received funding from the European Unions H2020 Research and Innovation Action under Grant Agreement H2020–ICT–644843 (5G–ESSENCE Project).

[1] Radio resources are considered in terms of Physical Resource Blocks (PRB) a pair of which is the smallest resource element assigned by the LTE eNB schedulers to UEs. Their quantity at eNBs depends on the number of sectors, LTE channels and the number of carriers. Hereafter, radio resources will be referred as PRBs.

utilization and backhaul link utilization by formulating and solving a joint content caching, UE association and resource allocation problem as an Integer Linear Programming (ILP) problem. In order to address the complexity of the proposed algorithm, we also propose a greedy heuristic to solve the problem.
- Second, we study the effect of the cache capacity and the file repository size on the resource utilization.
- Third, assuming the cached content is given, we compute the cost of PRB and backhaul bandwidth based on their utilization and, by using these costs, we propose an algorithm to jointly associate UEs and allocate resources, having the goal of avoiding congestion of those resources.

The rest of this paper is structured as follows. The related work is discussed in Sec. II. The problem statement along with the substrate network and UEs association request models are introduced in Sec. III. The problem formulation is presented in Sec. IV. The numerical results are reported in Sec. V. Finally, Sec. VI draws the conclusions.

II. Related Work

A considerable amount of literature has been published on joint user association and cache placement problem, optimizing various aspects of wireless communication and mobile computing in cache–enabled mobile networks. The major part of these works focuses on minimizing users' download delay [4], [5], [6]. A multiuser joint task offloading and resource optimization problem is studied in [4]. A heuristic decision offloading algorithm is proposed aiming at maximizing the utility metric, which characterizes users' Quality of Experience (QoE) in terms of their task completion time and energy consumption. Another joint user association and content placement problem is studied in [5] aiming to minimize the weighted sum of user download delay and caching cost, with the latter being proportional to the size of the content. The problem is formulated as a non-linear optimization problem, transformed into convex subproblems and solved using Kuhn–Munkres algorithm. In reference [6], a joint design and optimization of the content caching and user association problem is formulated as an integer non–linear programming problem aiming to minimize the average download delay subject to cache and backhaul link capacity constraints. In order to reduce the complexity of the problem, the authors decompose the original problem into an assignment problem and simple integer linear subproblems.

With the goal of maximizing the probability of files being fetched by the local cache, an optimal cache placement problem is studied in [7]. The problem is modeled as a discrete Markov chain with k–step corresponding to users random walk. It is then transformed into a binary integer programming problem and solved using Branch & Bound algorithm to find the optimal solution, and a greedy approximation algorithm to find a near–optimal solution.

Another group of studies focuses on maximizing users data rate and backhaul savings [8], [9]. In [8], the authors study the trade–offs between users' association and backhaul load reduction in cache–enabled mobile networks. Initially, an exact algorithm is proposed for joint optimization of content placement and users' association, aiming at maximizing the weighted sum of users' data rate and backhaul savings. Then, the authors propose an iterative algorithm that optimizes the content placement at each base station assuming fixed user association and optimizes user association assuming fixed caching policy. In reference [9], the authors formulate a user–cell association problem as a one–to–many matching game and propose a distributed algorithm to solve it. Unlike the previous work, the authors use users' speed and direction of arrival in order to estimate the duration in which a user will remain under the coverage of a cell.

Some other works aim to load balance users traffic in the network [10], [11], [12]. The authors of [10] propose a distributed user–traffic association algorithm with the goal of load balance the traffic at cached base stations under the assumption that the files are cached a priori. The load balancing is achieved by setting a soft limit on the maximum traffic associated with cached base stations. A joint user association and resource allocation problem is also studied in [11]. Having the objective of balancing users' total data rate and the data rate retrieved from the cache, the authors formulate a mixed integer non–linear problem and solve it using a dual composition method.

However, very few works consider users' mobility in their optimization problem, which in our opinion plays a pivotal role in a cache placement problem since cache contents are likely to change depending on the users' mobility. Moreover, to the best of our knowledge, this is the first work that studies the trade–offs between radio resources at eNBs and backhaul link resources making sure that users always get their demanded traffic. Additionally, this work employs a PRB and a backhaul link cost selection approach in order to avoid congesting the aforementioned resources in the network.

III. Network Model

This section formally states the problem and details the mobile network and the UE association request models.

A. Problem Statement

Figure 1 depicts the reference network scenario for the joint content caching, UE association and resource allocation problem. Consider an LTE network composed of two eNBs, which are equipped with a cache having limited capacity, and a core network, which serves as a gateway to access all the content files. For simplicity, it is assumed that the files requested by UEs have the same size and that each eNB has a fixed amount of memory. If the files have different sizes, they can be divided into blocks with equal sizes [11]. Thus, the cache capacity can be expressed in terms of the number of files that can be stored, which is obtained by dividing the cache memory size by a file size. In this reference network, it is assumed that each eNB has a cache with the capacity of storing two files. Suppose the UE is connected to eNB1 requesting the content one ($C1$). Since we consider UEs mobility, at some point in time, the UE may be located in an area where eNB2 may provide a better channel condition than eNB1. Assuming that $C1$ content is not cached at eNB2, the UE can either stay connected to eNB1 consuming more PRBs, or the UE can handover connection to eNB2, and therefore, although use a fewer PRBs, consume backhaul link resources. Hence, a trade–off exists between the backhaul link utilization

TABLE I: Substrate network parameters

Variable	Description
G_{net}	Mobile network graph.
N_{net}	Set of all eNBs in G_{net}.
N_{enb}	Set of ordinary eNBs in G_{net}.
N_{cdn}	Set of eNBs collocated with CDN in G_{net}.
E_{net}	Set of backhaul links in G_{net}.
$\omega_{prb}^{net}(m)$	Number of PRBs available at the eNB $m \in N_{net}$.
$\omega_{ccp}^{net}(m)$	The number of files that the eNB $m \in N_{enb}$ can cache.
N_F	Set of all files.
$N_f(m)$	Set of files cached at the ordinary eNB $m \in N_{enb}$.
$loc(m)$	Geographical location of the eNB $m \in N_{net}$.
$\delta(m)$	Coverage radius of the eNB $m \in N_{net}$ (in meters).
Λ_{bwt}	Cost for each Mbps of bandwidth resource.
Λ_{prb}	Cost for each PRB.

and the PRB utilization at the eNBs, and, depending on their utilization, MNOs can pick the preferable option and control UEs association and resource allocation accordingly.

The problem of joint content caching, UEs association and resource allocation can be formally stated as follows:

Given: a small cluster of an operational LTE–A network with eNBs, the cache capacity per eNB, the transport network topology with the capacity of each link, number of UEs and their requested contents with the traffic demand per content.

Find: UEs associations with the network resource allocation along with the content to be cached at each eNB.

Objective: minimize the weighted sum of the PRB utilization at eNBs and the backhaul link utilization.

B. Mobile Network Model

Let $G_{net} = (N_{net}, E_{net})$ be an *undirected* graph modelling the mobile network, where $N_{net} = N_{cdn} \cup N_{enb}$ is the union of the set of $n_1 = |N_{cdn}|$ eNBs, which are collocated with the Content Distribution Network (CDN) in the core network (called CDN eNBs) and $n_2 = |N_{enb}|$ the set of ordinary eNBs that have a cache with a fixed amount of memory. Notice that, as opposed to the ordinary eNBs that have a limited memory which allows a limited number of files to be cached, the CDN eNBs have all the files that UEs may request since they are collocated with the CDN. Lastly, let $E_{net} \subseteq N_{cdn} \times N_{enb}$ be the set of backhaul links. An edge $e^{nm} \in E_{net}$ if and only if a connection exists between $n, m \in N_{net}$. Two weights, $\omega_{prb}^{net}(m)$ and $\omega_{ccp}^{net}(m)$ are assigned to each eNB $m \in N_{net} : \omega_{prb,ccp}^{net}(m) \in \mathbb{N}^+$ representing, respectively, the number of PRBs at eNB m and the cache capacity expressed in terms of the number of files that the eNB m can cache. Each eNB is associated with a coverage radius of $\delta(m)$, in meters, indicating the coverage area of the eNB m, and a geographic location $loc(m)$, as x, y coordinates, which are derived by converting the real locations (i.e., longitude and latitude) of the eNBs of the considered operational LTE–A network. Another weight $\omega_{bwt}^{net}(e^{nm})$ is assigned to each link $e^{nm} \in E_{net} : \omega_{bwt}^{net}(e^{nm}) \in \mathbb{N}^+$ representing the capacity (in Gbps) of the wireless link connecting the eNBs n and m. Table I summarizes the mobile network parameters.

C. UEs Association Request Model

UEs association requests are modelled as *undirected* graphs $G_{req} = (N_{req}, E_{req})$ where $N_{req} = N_{ue} \cup N_{uef}$ is the union of the set of $n_1 = |N_{ue}|$ UEs and $n_2 = |N_{uef}|$ the set of their requested files, while $E_{req} \subseteq N_{ue} \times N_{uef}$ is the set of virtual links between UEs and their requested files. Notice that the requested files can be fetched either from the local cache of the host eNB, if it has cached the requested file, or from a CDN, which in our scenario is collocated with a CDN eNB. In the former case, no backhual bandwidth is consumed since the host eNB is the same eNB that serves the requested file. Conversely, in the latter case, backhual bandwidth is required to map the virtual link onto a backhaul path, connecting the host eNB with the CDN eNB, and fetch the file since the host eNB does not have the requested file in the local cache.

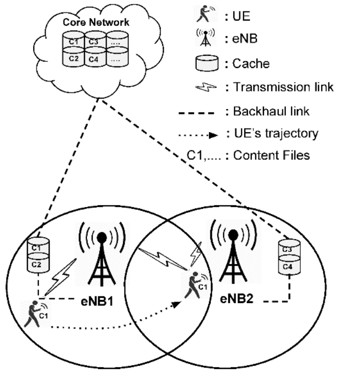


Fig. 1: A sample scenario of users mobility in a cache–enabled mobile network.

Each UE $u \in N_{ue}$ in the request has two weights $N_{uef}(u)$ and $\omega_{tp}^{u}(f)$ indicating, respectively, the set of the requested files and the requested throughput per requested file $f \in N_{uef}(u)$ [5]. Notice that it is our assumption that the throughput requested by the UEs for their files is always satisfied. Given the throughput demand of the file f requested by the UE u, the PRB demand $\omega_{prb}^{u}(m, f)$ for the file f to be provided by the eNB $m \in N_{net}$ can be computed as follows:

$$\omega_{prb}^{u}(m, f) = \frac{\omega_{tp}^{u}(f) T_{sbf}}{2 N_{sbc}^{m} N_{ofdms}^{m} N_{modb}^{m} N_{ant}^{m}}$$

where 2 is the number of PRBs in a subframe that has duration of T_{sbf} (1ms). N_{sbc}^{m} and N_{ofdms}^{m} are, respectively, the number of subcarriers (12) and the number of OFDM symbols per subcarrier (7) in a PRB, while N_{ant}^{m} is the number of MIMO streams. Notice how these parameters are unequivocally defined for a given version of the LTE standard. N_{modb}^{m} is the number of modulated bits per symbol. For example, if a 64–QAM modulation is used, then $N_{modb}^{m} = 6$. Several models linking channel quality and MCS can be found in literature [13]. However, since the focus of this paper is on the formulation of the content caching and the user association problems, the selection of a particular channel model, although important, takes a secondary role. As a result, in the numerical evaluation, we will leverage on a simple MCS estimation model which is based on the distance between the UE and the host eNB. Each UE $u \in N_{ue}$ is also associated with a geographic location $loc(u)$, as x, y coordinates. This information together with the locations of the eNBs and their coverage radius is used to find candidate eNBs that can host the UEs. Table II summarizes the UE request parameters.

TABLE II: UE request parameters

Variable	Description
G_{req}	UE association request graph.
N_{uef}	Set of files requested for all UEs in G_{req}.
$N_{uef}(u)$	Set of files requested for the UE $u \in N_{ue}$.
N_{ue}	Set of UEs in G_{req}.
E_{req}	Set of all virtual links in G_{req}.
$E_{req}(u)$	Set of virtual links of the UE $u \in N_{ue}$.
$loc(u)$	Geographical location of the UE $u \in N_{ue}$.
$\omega^u_{prb}(m,f)$	PRBs needed from eNB $m \in N_{net}$ for file $f \in N_{uef}(u)$.
$\omega^u_{prb}(m)$	PRBs needed from eNB $m \in N_{net}$ for UE $u \in N_{ue}$.
$\omega^u_{tp}(f)$	Throughput request by file $f \in N_{uef}(u)$ of UE $u \in N_{ue}$.
$\omega^u_{bwt}(e')$	Bandwidth demand of link $e' \in E_{req}(u)$ of UE $u \in N_{ue}$.

TABLE III: Binary decision variables $\{0, 1\}$

Variable	Description
Φ^u_m	Shows if the UE $u \in N_{ue}$ has been mapped to the eNB $m \in N_{net}$.
$\Phi^{u,f}_m$	Shows if the file $f \in N_{uef}(u)$ requested by the UE $u \in N_{ue}$ has been fetched from the eNB $m \in N_{net}$.
$\hat{\Phi}^{u,f}_m$	Shows if the file $f \in N_{uef}(u)$ requested by the UE $u \in N_{ue}$ that was mapped on the eNB $m \in N_{net}$ has been fetched from the CDN eNB that is different from the host m eNB.
$\Phi^{u,e'}_e$	Shows if the virtual link $e' \in E_{req}(u)$ of the UE $u \in N_{ue}$ has been mapped to the substrate link $e \in E_{net}$.

IV. Problem Formulation

Upon arrival of a UE association request, the substrate network must decide if it can be accepted and mapped or if it shall be rejected. This association problem can be considered as a virtual network embedding problem [14], which is *NP*–hard and has been studied extensively in the literature [15], [16]. The embedding process consists of two steps: the node embedding and the link embedding. In the node embedding step, each virtual node (i.e., UEs, UEs' requested files) in the request is mapped to a substrate node (i.e., an ordinary eNB, a CDN eNB). In the link embedding step, each virtual link (i.e., the link between UEs and their requested files) is mapped to a single substrate path (i.e., a path between the host eNBs). In both cases, nodes and links constraints must be satisfied.

A. ILP Formulation

Before formulating the ILP problem, we first need to find the candidate eNBs for each UE in the UE association request. Considering the location $loc(u)$ of the UE $u \in N_{ue}$ along with the location $loc(m)$ and the coverage radius $\delta(m)$ of the eNBs $\forall m \in N_{net}$, a cluster of candidate eNBs $\Omega(u)$ for the UE $u \in N_{ue}$ can be defined as follows:

$$\Omega(u) = \Big\{ m \in N_{net} | dis(loc(m), loc(u)) \leq \delta(m) \Big\} \quad (1)$$

We can now formulate the ILP problem. The objective function (formula (2)) of this ILP problem is to minimize the weighted sum of the PRB utilization and the backhaul link utilization. The weights/costs Λ_{bwt} and Λ_{prb} provide the possibility for the MNOs to control the use of the PRBs and backhaul links of the eNBs aiming at preventing partial radio resource and/or backhaul link congestion in the network.

$$min\Big(\sum_{u\in N_{ue}} \sum_{m\in N_{net}} \sum_{f\in N_{uef}(u)} \Lambda_{prb}(m)\Phi^{u,f}_m \omega^u_{prb}(m,f) + \\ + \sum_{u\in N_{ue}} \sum_{e\in E_{net}} \sum_{e'\in E_{req}} \Lambda_{bwt}(e)\Phi^{u,e'}_e \omega^u_{bwt}(e') \Big) \quad (2)$$

We will now detail the constraints used in this ILP problem. Constraint (3) ensures that each UE is associated to one eNB that belongs to its cluster of candidates (4).

$$\sum_{m\in N_{net}} \Phi^u_m = 1 \quad \forall u \in N_{ue} \quad (3)$$

$$\sum_{m\in N_{net}\setminus\Omega(u)} \Phi^u_m = 0 \quad \forall u \in N_{ue} \quad (4)$$

Constraint (5) makes sure that each file $f \in N_{uef}(u)$ requested by the UE $u \in N_{ue}$ is fetched by only one eNB that can be either the host eNB or a CDN eNB (6).

$$\sum_{m\in N_{net}} \Phi^{u,f}_m = 1 \quad \forall u \in N_{ue}, \quad \forall f \in N_{uef}(u) \quad (5)$$

$$\Phi^u_m - \Phi^{u,f}_m - \hat{\Phi}^{u,f}_m = 0 \quad (6)$$

$$\forall u \in N_{ue}, \quad \forall f \in N_{uef}(u), \quad \forall m \in N_{net}$$

Constraint (7) enforces for each virtual link between the UE $u \in N_{ue}$ and its requested file $f \in N_{uef}(u)$ be a continuous path established between the eNB hosting the UE and the eNB providing the requested file. $E^{\star m}_{net}$ is the set of the backhual links that originate from any eNB and directly arrive at the eNB $m \in N_{net}$, while $E^{m\star}_{net}$ is the set of the backhaul links that originates from the eNB $m \in N_{net}$ and arrive at any eNB directly connected to m.

$$\sum_{e\in E^{\star m}_{net}} \Phi^{e^{u,f}}_e - \sum_{e\in E^{m\star}_{net}} \Phi^{e^{u,f}}_e = \begin{cases} -1 & \text{if } m = u \\ 1 & \text{if } m = f \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$\forall m \in N_{net}, \quad \forall e^{u,f} \in E_{req}$$

Virtual links can be mapped to a backhaul link in the mobile network as long as the backhaul link has enough capacity to meet the bandwidth demand of the virtual links (8).

$$\sum_{u\in N_{ue}} \sum_{e'\in E_{req}(u)} \omega^u_{bwt}(e')\Phi^{u,e'}_e \leq \omega^{net}_{bwt}(e) \quad \forall e \in E_{net} \quad (8)$$

The cache capacity at each eNB determines the number of files that can be cached, assuming that the files have equal size. Constraint (9) makes sure that the quantity of the cached files is less or equal to the cache capacity of the host eNB.

$$\sum_{f\in N_F} \Phi_f \leq \omega^{net}_{ccp}(m) \begin{cases} \Phi_f = 1 & \text{if } \sum_{u\in N_{ue}} \Phi^{u,f}_m \geq 1 \\ \Phi_f = 0 & \text{otherwise} \end{cases} \quad (9)$$

$$\forall m \in N_{net}$$

Algorithm 1: *eNB Caching Algorithm*

```
Data: $(G_{net}, G_{req})$
Result: Cached Content at the eNBs.
Phase I: Find the list of candidates;
for $u \in N_{ue}$ do
    • For the UE $u$, find the candidate eNBs;
end
Phase II: Cache the popular content;
repeat
    for $f \in N_F$ do
        for $m \in N_{enb}$ do
            $ratio = \frac{PRB(m,f)}{BANDWIDTH(m,f) \cdot QUANTITY(m,f)}$;
        end
    end
    • Start file caching from the file at the eNB that has the minimum ratio;
    • Delete the demand for the file that has been cached from those UEs
      that request that file and have as a candidate eNB the eNB that has just
      cached the file.
until Cache of each eNB is full;
```

Algorithm 2: *UE Association Algorithm*

```
Data: $(G_{net}, G_{req}, N_f)$
Result: UEs Association and Resource Allocation.
repeat
    • Compute PRB and backhaul link costs;
    for $u \in N_{ue}$ do
        • Associte the UE $u$ to the candidate eNB that after association its
          requested file(s) can be fetched with the minimal cost;
        for $f \in N_{uef}(u)$ do
            • Fetch the file $f$ requested by the UE $u$ from the eNB which
              would serve the file with the minimal cost;
            • Allocate PRB and backhaul bandwidth resources;
            • Update network resources;
            • Compute PRB and backhaul link resource utilization.
        end
    end
until All batch requests are mapped;
```

Constraints (10) and (11) pertain to the PRB utilization. Specifically, constraint (10) forces that the PRBs be provided by the eNB that the UE has been associated with, regardless of the eNB that provides the requested file. In other words, even if the requested file is fetched from the CDN eNB, which is different from the host eNB, PRBs must be allocated by the host eNB. Lastly, constraint (11) ensures that the eNBs can associate UEs as long as they have enough PRBs to meet their requested file throughput demand.

$$\omega^u_{prb}(m)\Phi^u_m - \sum_{f \in N_{uef}(u)} (\Phi^{u,f}_m + \hat{\Phi}^{u,f}_m)\omega^u_{prb}(m,f) = 0 \quad (10)$$

$$\forall u \in N_{ue}, \quad \forall m \in N_{net}$$

$$\sum_{u \in N_{ue}} \sum_{f \in N_{uef}(u)} \omega^u_{prb}(m,f)(\Phi^{u,f}_m + \hat{\Phi}^{u,f}_m) \leq \omega^{net}_{prb}(m) \quad (11)$$

$$\forall m \in N_{net}$$

The described ILP formulation is for the joint content caching, UE association and resource allocation problem. Notice that for the joint UE association and resource allocation problem, the ILP problem formulation is the same with the only difference that the files are cached a priori. Therefore, the ILP has no need to select which files to cache at the eNBs since it is already given. This information exempts the need for applying constraint (9). Whereas, the rest of the constraints along with the objective function are left the same.

B. Heuristic

The ILP problem becomes computationally intractable as the size of the mobile network, the cache capacity at the eNBs and the number of UEs, making association request, increase. For example, the ILP algorithm takes 10 minutes on Intel Core i7 laptop (3.0 GHz CPU, 16 Gb RAM) using the Matlab ILP solver (intlinprog) to associate 50 UEs (a single batch UE association request) and provide their requested content by the mobile network composed of 1 CDN eNB and 6 eNBs that have cache capacity enough for storing 3 files. In order to address this scalability issue, we also propose a heuristic that is able to embed the same batch UE association request in less than a second.

Let us make the following notations before describing the heuristic. Let $n_1 = |N_{ue}|$ and $n_2 = |N_{net}|$ be the number of, respectively, UEs and eNBs. Then, let c_1 and c_2 be the number of, respectively, candidate eNBs of the UEs and the files cached at the eNBs. Finally, let $k = |E_{net}|$ be the number of the links between ordinary eNBs and the CDN eNB.

The proposed heuristic is composed of two parts: eNB caching (Alg. 1) and UE association (Alg. 2). Both of the algorithms are employed in order to solve the joint content caching, UE association and resource allocation problem. Whereas, only the UE association algorithm is used to solve the joint UE association and resource allocation problem, providing cached files at the eNBs as an input. The eNB caching algorithm consists of two phases and aims at finding the optimal content to be cached at the eNBs. In the first phase, for each UE, the algorithm loops over each eNB and creates a list of candidate eNBs by considering the coverage of the eNBs and the distance between the UE and the eNBs. The required time in order to complete this step is $O(n_1 n_2)$. The second phase aims at finding the popular content to be cached at the eNBs. Initially, the quantity, the PRB requirement in order to support the overall throughput demand, and the bandwidth demand on the backhaul links is computed for each requested file at each eNB. Then, a composite metric, which is the ratio between the PRB demand and the product of the bandwidth demand and the quantity of the requested file, is computed for each file of each eNB. The file caching starts from the eNB that has the minimum ratio for any file and that file is cached at that eNB. This is followed by deleting the cached file demand from the UEs that are under the coverage area of the eNB that has just cached the file. This process is repeated until the cache capacity of the eNBs is filled with the cached files. This phase takes $O(n_2 c_2 [n_1 c_1 + n_2])$ time.

Once the content caching is over, the information about the cached content is fed to the UE association algorithm, which initiates the UE association process. Before mapping the UE batch association request, the PRB and the backhaul link costs are computed, respectively, for each eNB and backhaul link by considering their utilization. The costs of these resources are directly proportional to their utilization. In other words, the less is the utilization of a PRB/link at an eNB, the cheaper is the cost of the PRB/link at that eNB. For each UE in each batch association request, the algorithm considers all its candidate eNBs and computes the association cost, expressed as the sum of the costs of the required PRBs and the backhaul bandwidth, in order to map the files requested by the UE to their corresponding candidate eNBs. The algorithm then associates the UE with the cheapest candidate eNB.

After the UE has been associated, the algorithm considers the files requested by the UE and checks the availability of each file at the cache of the host eNB. If the file is available, it is fetched from the host eNB; otherwise, it is fetched from the CDN eNB through the shortest path, which is picked by using Dijkstra's shortest path algorithm. After the file mapping, the required PRB and bandwidth resources are allocated, and the network resources are updated. Lastly, the utilization of PRBs and the backhaul links is computed for each eNB, and the described process is repeated until all the UE association requests of all the batches are mapped. This step takes $O(n_1 c_1 [\log_{10} n_2 (c_1 + 1) + 1])$ time.

V. Evaluation

The goal of this section is to compare the ILP–based and the heuristic algorithms. We shall first describe the simulation environment in our study. We will then report on the outcomes of the numerical simulations carried out in a simulator implemented in Matlab®.

A. Simulation Environment

A small cluster (i.e., 7 eNBs) of an operational LTE–A network in the city center of Yerevan (Armenia) is considered in our simulations. The cluster provides mobile coverage in an area of $3Km^2$. The CDN resides with the eNB deployed in the center. All other ordinary eNBs are connected to the core network by means of 1 Gbps wireless microwave links. The number of sectors per eNB, as well as the number of carriers/cells per sector, vary in the set of $\{3, 4\}$ and $\{1, 2, 3\}$, respectively. Three LTE channels, 20MHz, 15MHz and 10MHz, are employed in the network. For simplicity, it is assumed that omnidirectional antennas are employed and that the eNBs support 2×2 MIMO configuration. However, the overall channel bandwidth capacity of each eNB corresponds to the real value computed as the sum of the LTE channel bandwidths available at each sector of the considered eNB.

Initially, we consider mobile users and we compute the optimal file caching, considering their requirements. It is assumed that the users are moving in random directions with different speeds [17]. More specifically, in our model users speed are randomly picked in $[3, 5, 10]$ Km/h set. 10 simulation runs are considered each corresponding to a single time sample of users movement. Due to scalability issue of the ILP–based algorithm, the simulations are considered for from 50 to 300 UEs, which make association requests with each specifying maximum two files and the required throughput per file. For simplicity, during each simulation run, it is assumed the UEs requirements are known and do not change over time, regardless of their location. Although capturing UEs changing requirements would make the scenario more realistic, it would also unnecessarily complicate the model without adding any value to the main message of this study. UEs randomly request files from the file repositories with the size of 10, 15 and 20. The local cache capacity is expressed in terms of the number of files that can be cached, assuming the files have equal size, and the simulations are run for cache capacities enough for storing 1, 2 and 3 files.

The results plotted in Fig. 2 and Fig. 3 are for the average of 10 simulations for both ILP and heuristic algorithms for cheap PRB (*CP-I* and *CP-H*) and for cheap link (*CL-I* and *CL-H*) cases. In these simulations, the cost of cheap PRB (Λ_{prb}) for both algorithms is selected as a half of the cost of a backhaul link. Whereas, the cost of a cheap link (Λ_{bwt}) is selected as a half of the cost of a PRB. It is important to mention that these values are selected for the sake of demonstrating the trade–off between the backhaul link utilization and the PRB utilization.

In order to show the effect of computing and using appropriate PRB and backhaul link costs, we also conduct a simulation in which the MNO sequentially receives 100 batches of UE association requests (each composed of 5 UE association requests), assuming the eNBs have already cached the content.

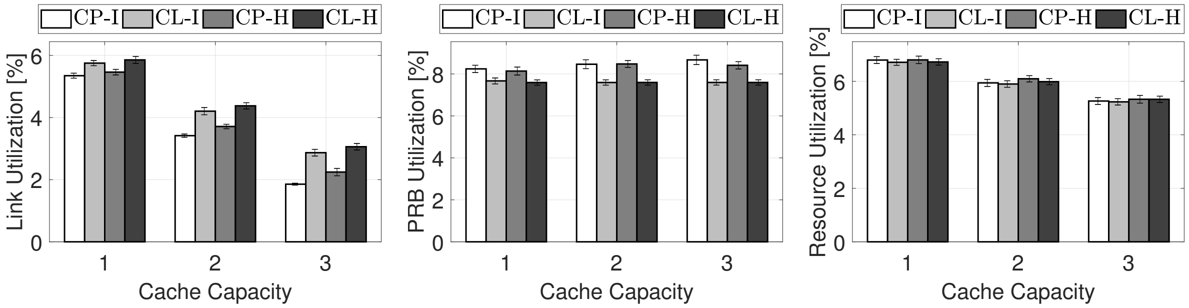
B. Simulation Results

Figure 2 and Fig. 3 show the backhaul link, the PRB and the overall resource utilization as a function of, respectively, the cache capacity and the file repository size at the eNBs for a fixed file repository size for a single batch of UE association request composed of 50 UEs. In Fig. 2a, it can be observed that the link utilization reduces for the ILP and for the heuristic algorithms in both cases with the increase in the cache capacity at the eNBs. This is justified by the fact that with the cache capacity increase, the eNBs are able to cache more files, and therefore, the probability that the files requested by UEs can be directly fetched from the host eNBs increases, which, in turn, reduces the backhaul link utilization. It can also be observed that for both algorithms the backhaul link utilization is more when its cost Λ_{bwt} is cheaper from the PRB cost Λ_{prb}.

Figure 2b displays the PRB utilization as a function of the cache capacity. It can be seen that there is no significant difference in the PRB utilization when the cache capacity increases. This is because regardless of the cache capacity, the UEs have to be provided the required number of PRBs in order to meet their traffic demand. For both algorithms, we can notice that the PRB utilization in the case of cheap PRBs is more than that in the case of cheap links.

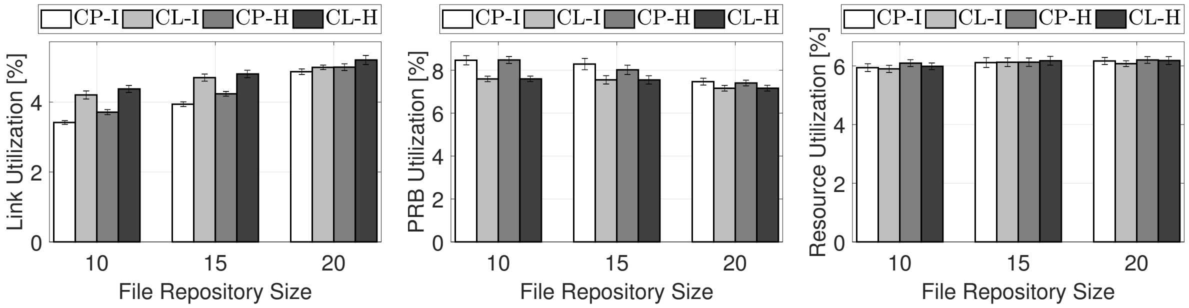
In order to compare the ILP algorithm with the heuristic, let us now analyze the sum of the link resource utilization and the PRB utilization for different cache capacities (see Fig. 2c). Irrespective of the cache capacity at the eNBs, we can observe that the overall resource utilization for ILP algorithm for both cheap PRB and cheap link cases is less or equal to the overall resource utilization in the case of the heuristic. This negligible difference in the resource utilization witnesses the fact the heuristic is able to find solutions very close to the optimal ones found by the ILP algorithm.

Figure 3a captures the effect of the file repository size on the link utilization. We can observe that the link utilization increases along with the increase in the file repository size since the more is the variety of the files that can be requested/cached, the less is the probability that the requested files are available at the local cache of the eNBs. This, in turn, increases the probability of the files being fetched from the CDN eNB located in the core network rather than from the host eNB, resulting in an increase in the backhaul link utilization.



(a) Link utilization vs. cache capacity.
(b) PRB utilization vs. cache capacity.
(c) Resource utilization vs. cache capacity.

Fig. 2: Average link utilization, PRB utilization and overall resource utilization as a function of the cache capacity of the eNBs in the joint content caching, user association and resource allocation problem.



(a) Link utilization vs. file repository size.
(b) PRB utilization vs. file repository size.
(c) Resource utilization vs. file repository size.

Fig. 3: Average link utilization, PRB utilization and overall resource utilization as a function of the file repository size in the joint content caching, user association and resource allocation problem.

The PRB utilization for different file repository sizes is illustrated in Figure 3b. We can observe that the PRB utilization slightly decreases with the file repository size increase. Since the probability of the requested files being available at the local cache of the eNBs reduces with the file repository size, the UEs have no need for consuming many more PRBs in order to reach to a farther eNB candidate that might otherwise have the requested file at the local cache. Therefore, the UEs prefer to use fewer PRBs in the cases of both algorithms. Moreover, we can observe that the PRB utilization is more for both algorithms when the PRBs cost cheaper than the backhaul links. Finally, let us analyze the overall resource utilization as a function of the file repository size depicted in Fig. 3c. We can observe that similar to Fig. 2c, the heuristic for both cheap PRB and cheap link cases achieves resource utilization values very close to the optimal ones found by the ILP algorithm, leading to nearly equal utilization of the overall resources.

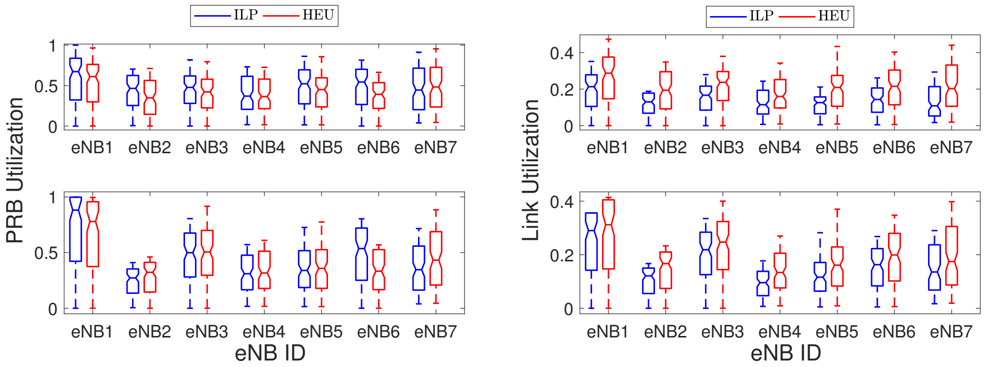
Thus, both Fig. 2 and Fig. 3 clearly demonstrate the trade–off between PRB and backhaul bandwidth. Specifically, we observe that the cheaper is the cost of a PRB from the cost of a Mbps backhaul bandwidth, the higher is the PRB utilization, but also the lower is the backhaul link utilization at the eNBs. Similarly, the cheaper is the single unit of backhaul bandwidth utilization cost from the cost of a PRB, the higher is the utilization of the backhaul link; however, the lower is the PRB utilization at the eNBs. Hence, setting appropriate backhaul bandwidth and PRB utilization costs enables MNO to balance the utilization of these resources with the ultimate goal of avoiding their congestion at the eNBs. One possible way to implement this pricing mechanism is to employ cell range expansion techniques such as the one using RSRP bias [18], which allows UEs to camp on the eNBs having *weak* RSRP, and therefore, offloading the eNBs having congested resources.

TABLE IV: Execution Time

Number of UEs	50	100	150	200	250	300
CP-I (sec.)	5.93	30.37	90.37	144.01	237.93	304.33
CL-I (sec.)	5.9	30.49	89.42	145.7	258	326.1
CP-H (sec.)	0.11	0.17	0.31	0.34	0.39	0.43
CL-H (sec.)	0.12	0.2	0.29	0.34	0.39	0.47

The main motivation for proposing the heuristic is to address the scalability issue of the ILP algorithm. In order to get an insight into how fast and scalable is the proposed heuristic, let us analyze Table IV, which shows the execution time of both algorithms for cheap PRB and cheap link cases for a different number of UE association requests. The reported values are for the studied mobile network in which the file repository size is 10, and each eNB has a cache capable of storing a single file. It can be observed that the execution time increases dramatically with the number of UEs for the ILP algorithm for both cases, while the execution time increase is negligible for the heuristic. For example, in order to associate 300 UEs to the network meeting their file and traffic demand, the required time for the ILP algorithm is 5 minutes, while the heuristic can achieve this association in less than a half second.

In order to ascertain the effect of the PRB and the backhaul link cost selection on the utilization of the said resources, let us analyze Fig. 4, which illustrates the distribution of the PRB



(a) PRB utilization distribution at eNBs.

(b) Link utilization distribution at eNBs.

Fig. 4: Distribution of PRB and backhaul link resource utilization for 100 batches of UE association requests for the cases of using the computed PRB and backhaul link costs (upper box plots), and the default costs (the lower box plots) in the joint user association and resource allocation problem.

utilization and the backhaul link utilization for all eNBs after sequentially mapping 100 batches of UE association requests in the joint UE association and resource allocation problem. We remind the reader that each batch request in this problem is composed of 5 UEs, and the cached content at the eNBs is given. Fig. 4a displays the distribution of the PRB utilization at eNBs for both ILP and heuristic algorithms with the upper box plot corresponding to the case when the PRB and the backhaul link costs are computed after each batch association and used for next batch associations, and the lower box plot for the case in which the default costs are always used ($\Lambda_{prb} = \Lambda_{bwt} = 1$). It can be observed that computing and using PRB and the backhaul link costs results in a more uniform utilization of those resources at the eNBs for both the ILP and the heuristic algorithms compared to the case of using the default costs. For example, we can see the PRB utilization at eNB1 reaches up to 100% in the default cost case (the lower box plot); whereas, at the same time, the PRB utilization is less than 50% at the eNB2, regardless of the employed algorithm. This uneven load distribution in this particular scenario leads to around 10% rejection of the UEs association request compared to the case when Λ_{prb} and Λ_{bwt} are computed and used (the upper box plot). Both the ILP and the heuristic have accepted all the batch association requests (i,e., 500 UEs in total) in the case of using the computed costs. Whereas in the case of using the default costs, both of them have equally associated 455 UEs.

The picture is similar also for the backhaul link utilization depicted in Fig. 4b. It can be observed that a significant difference exists in the backhaul link utilization of the eNBs in the case of using the default cost for the backhaul links (the lower box plot). Conversely, the backhaul links of eNBs are more uniformly utilized in the case of employing the computed link cost, leading to a reduction of the probability of rejecting UE association requests due to insufficient backhaul link capacity. Fig. 4b also illustrates that in the considered scenario the UEs batch association requests never got rejected because of the insufficient backhaul link capacity since the maximum link utilization at the eNBs is around 40%. This is solely due to the fact that in the considered scenario each eNB has only a single backhaul link, which results in each link being used by only one eNB. However, in the case of having multiple hops from the eNBs to the core network where the CDN resides, some of the backhaul links will have to be shared among several eNB, and the problem of uniform backhaul link utilization will become more severe, requiring careful backhaul link mapping in order to avoid as much as possible congesting some of the backhaul links, which otherwise may become a cause for rejecting UE association requests.

VI. Conclusions

Cache–enabled Mobile Edge Computing is perceived to be a promising way to alleviate mobile backhaul load. In order to fully deliver on its promises, however, it calls for efficient content caching approaches at the network edge (i.e. at eNBs). Additionally, approaches are needed to balance the utilization between the backhaul links and PRBs at the eNBs.

In this study, we demonstrate the trade–offs between the backhaul link utilization and the PRB utilization at eNBs by formulation and solving a mobility–aware joint content caching, UE association and resource allocation problem. Assuming content is cached a priori, we then show the effect of computing and using a unit of a backhaul link and a PRB costs at the eNBs by formulating and solving a joint UE association and resource allocation problem. Both problems are formulated as VNE problems and solved using ILP techniques with the objective of minimizing the weighted sum of backhaul link and PRB utilization while making sure that UEs QoS is not compromised. Heuristics are also proposed in order to tackle the scalability issues of the ILP algorithms.

Small–scale simulations show that our algorithm, due to more uniform PRB utilization at the eNBs, got 10% more UEs associated with the network compared to the baseline approach in which no backhaul link and PRB costs were used. Given the imbalanced backhaul link utilization of the baseline approach, we can deduce that in a big–scaled network, the gap in the number of UEs association between these two approaches will grow more. This stems from the fact that the backhaul links of each eNB will be utilized also by other eNBs resulting in a congestion on some backhaul links, which may ultimately entail to rejecting more UEs association requests.

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021," Cisco, Tech. Rep., 2017.

[2] "Ericsson Mobility Report, 2016-2021," Ericsson, Tech. Rep., 2017.

[3] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing - A key technology towards 5G," *ETSI White Paper*, vol. 11, no. 11, pp. 1–16, 2015.

[4] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser joint task offloading and resource optimization in proximate clouds," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3435–3447, 2017.

[5] H. Chen, Q. Chen, R. Chai, and D. Zhao, "Utility function optimization based joint user association and content placement in heterogeneous networks," in *Proc. of IEEE WCSP*, Nanjing, China, 2017.

[6] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *IEEE Access*, vol. 4, pp. 8625–8633, 2016.

[7] I. Keshavarzian, Z. Zeinalpour-Yazdi, and A. Tadaion, "A clustered caching placement in heterogeneous small cell networks with user mobility," in *Proc. of IEEE ISSPIT*, Abu Dhabi, UAE, 2015.

[8] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," in *Proc. of IEEE ICASSP*, Shanghai, China, 2016.

[9] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Proc. of IEEE WiOpt*, Hammamet, Tunisia, 2014.

[10] J. Krolikowski, A. Giovanidis, and M. Di Renzo, "Fair distributed user-traffic association in cache equipped cellular networks," in *Proc. of IEEE WiOpt*, Paris, France, 2017.

[11] G. Ren, H. Qu, J. Zhao, S. Zhao, and Z. Luan, "A distributed user association and resource allocation method in cache-enabled small cell networks," *China Communications*, vol. 14, no. 10, pp. 95–107, 2017.

[12] D. Harutyunyan, S. Herle, D. Maradin, G. Agapiu, and R. Riggio, "Traffic-aware user association in heterogeneous LTE/WiFi radio access networks," in *Proc. of IEEE/IFIP NOMS*, Taipei, Taiwan, 2018.

[13] J. Du, L. Zhao, J. Feng, J. Xin, and Y. Wang, "Enhanced PSO based energy-efficient resource allocation and CQI based MCS selection in LTE-A heterogeneous system," *China Communications*, vol. 13, no. 11, pp. 197–204, 2016.

[14] A. Fischer, J. F. Botero, M. Till Beck, H. De Meer, and X. Hesselbach, "Virtual network embedding: A survey," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 4, pp. 1888–1906, 2013.

[15] M. Chowdhury, M. R. Rahman, and R. Boutaba, "ViNEYard: Virtual network embedding algorithms with coordinated node and link mapping," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 206 –219, February 2012.

[16] D. Harutyunyan and R. Riggio, "Flex5G: Flexible Functional Split in 5G Networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 961–975, 2018.

[17] J. Xu, Y. Zhao, and X. Zhu, "Mobility model based handover algorithm in LTE-Advanced," in *Proc. of IEEE ICNC*, Honolulu, Hawaii, USA, 2014.

[18] P. Tian, H. Tian, J. Zhu, L. Chen, and X. She, "An adaptive bias configuration strategy for range extension in LTE-advanced heterogeneous networks," 2011.