A Decomposition Framework for Optimal Edge-Cache Leasing

Jonatan Krolikowski, Anastasios Giovanidis, Marco Di Renzo

A Decomposition Framework for Optimal Edge-Cache Leasing

Jonatan Krolikowski, Anastasios Giovanidis, *Member, IEEE*, and Marco Di Renzo, *Senior Member, IEEE*

***Abstract*—Caching popular content at the wireless edge promises performance benefits as well as business perspectives. In this work we study the following arrangement: a Mobile Network Operator (MNO) pre-installs memory on its wireless equipment, sets a price, and invites a unique Content Provider (CP) to invest. The CP leases memory space and places its content; the MNO then associates network users to stations, aiming for offloading CP traffic or not.**

We formulate an optimization problem, which maximizes offloading with minimum leasing costs for the CP. This is an NP-hard mixed-integer non-linear optimization problem. We present an iterative exact solution using Generalized Benders decomposition into a content-related Master problem and a user-association Slave problem. Master is integer linear. Slave is convex for various association policies, including (1) join-the-closest-cache, and (2) cache-aware association. For Slave, we introduce a distributed exact solution, named Generalized Bucket-filling. Extensive simulations illustrate the performance benefits under different association policies. The solution helps to determine the optimal leasing price for the MNO, and the optimal investment budget for the CP. As a general conclusion, both actors profit when the MNO association supports CP decisions.

***Index Terms*—Edge Caching; Mixed-Integer Optimization; Decomposition; User Association; Content Pre-Fetching; Leasing**

I. Introduction

By 2020, wireless data traffic is estimated to reach roughly the 8-fold of its volume of 2015 [3]. Such increase is a challenge to mobile network operators (MNOs) as well as to the content providers (CPs). The MNOs and the CPs have different strategies to cope with these high demands: The MNOs, on the one hand, try to satisfy this increase by densifying the network with new tiers and by allowing cooperation among stations. This results in an increase of wireless traffic that can pose new problems to the wireless backhaul related to congestion. On the other hand, the CPs are on the receiving end of the content requests. Large CPs such as Youtube or Netflix store their data in huge data centers. For such a CP, a steep increase in data demand can be handled by massive infrastructure investments, i.e. upgrade of the data center capacity as well as installation of higher capacity data links to the surrounding network.

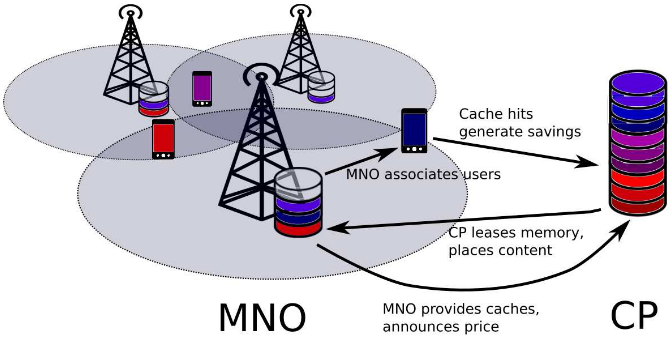

Fig. 1: Cached wireless network run by the MNO, caches leased by the CP.

A recently studied alternative that benefits both actors is to equip wireless nodes with caches (see e.g. [4]–[8]). The main purpose of cache memory installation is to ease backhaul traffic and its processing at the data centers by handling content requests from intermediate caches placed at the edge or inside the core network. This way, the total backbone traffic is reduced and user Quality of Service (QoS) can be improved because cached content is downloaded with less delay from sources closer to the user. Thus, caching is of interest to both the MNOs as well as the CPs.

We consider a scenario in which an MNO has constructed and physically maintains memories at Cached Base Stations (CBSs). The caches are destined to store the multimedia content of a certain CP. We claim that the CP should be in charge of managing the edge-cache memories for the following reasons: The CP usually transfers data to wireless users via secure connections [9], e.g. https [10], so that the intermediate MNO cannot recognize requests and serve them from local caches. Furthermore, content placement policies may depend on spatio-temporal popularity data that only the CP can gather and use.

The MNO profits from edge-caching though the congestion reduction, but also requires a financial recompensation for its storage installation investment. It makes three strategic decisions:

MNO-1) How much storage space to install at each CBS?
MNO-2) Which price to set for the leasing of one cache unit?
MNO-3) Which user association policy to pursue?

The user association policy is critical as will be shown next.

The CP has its own data centers to store its own content, so that it needs a good reason to make use of the MNO caching infrastructure. Its motivation comes from savings in the CP's operation through cache utilitzation: Each time a user-request

Received: December 10, 2017. Revised: April 8, 2018.

Jonatan Krolikowski (corresponding author) and Marco Di Renzo are with Laboratoire des Signaux et Systèmes (L2S), CNRS-CentraleSupélec-Université Paris-Sud, Université Paris Sud, Université Paris-Saclay, 3, rue Joliot Curie, 91192, Gif-sur-Yvette, France, Email: jonatan.krolikowski@l2s.centralesupelec.fr, marco.direnzo@l2s.centralesupelec.fr

Anastasios Giovanidis is with Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France, Email: anastasios.giovanidis@lip6.fr

This is an extended version of preliminary results presented in WiOpt-CCDWN 2017, Paris, France [1] and INFOCOM 2018, Honolulu, HI, USA [2]. The research is funded by LABEX-DigiCosme through the project CONTAIN.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2018.2844986, IEEE Journal on Selected Areas in Communications

for some CP-owned content is served through an edge cache memory, it can be delivered in high quality. At the same time, traffic towards the CP's data center is reduced. For the CP to harvest these savings, it should compensate the MNO financially.

Acknowledging such roles for the CP and the MNO, in this work *we take the point of view of the CP*, which reacts to the MNO's strategic decisions. The CP wants to lease sufficient cache space and place its content so that the savings exceed the cache leasing costs. The two types of decisions that the CP takes are:

CP-1) How much cache space to lease from the MNO at each CBS?

CP-2) Which content to place into the leased caches?

For the decisions to be optimal, they need to be based on the estimated spatio-temporal popularity of the CP's content. Although demand statistics evolve over time, one can consider time windows during which the system parameters can be assumed static. For example, the pre-fetching of content can take place periodically during off-peak hours [11]. The interactions between CP and MNO are illustrated in Fig. 1.

The MNO's user association policy plays a crucial role in deriving benefits from edge-caching. On the one hand, signal strength is vital for wireless transmission. On the other hand, there is no caching gain when users are associated to CBSs not caching the requested content. Due to the association policy, the interaction between the CP and the MNO goes beyond a simple market relation of supply and demand. To clarify this, we consider the following two possibilities.

Policy 1: Standard Association (CLOSEST). The MNO associates each wireless user to the geographically closest CBS. This is the conventional policy in cellular networks that only takes signal quality (through distance) into account. Here, the interaction between CP and MNO is restricted to leasing and placement because the user association is independent of CP actions.

*Policy 2: Cache-aware Association (*OPT*).* Such MNO association policy allows for more user requests to be served by the caches. If users can be associated to any single wireless node among those with sufficient signal quality, then each user has potential access to the union of sets of files cached in all covering stations. It is then beneficial for the CP to cache different sets of content in neighboring nodes. The MNO should associate users to CBSs in such a way that content requests are matched to the cached content. Here, association needs to be a function of the placement decisions. A further discussion on this matter follows in Remark 1, Section III-D.

The contributions of this work are summarized as follows:

• We formulate a mixed integer Network Utility Maximization (NUM) problem that aims to maximize the CP's caching benefits given the user association policy. The leasing and caching decisions are discrete, taken by the CP. The user association decision variables are fractional and can be determined by MNO-CP interaction.

• As solution technique, we introduce generalized Benders decomposition of the NUM into Master and Slave subproblems which converges to the global optimum. One of its main advantages, aside optimality, is that it allows the separation of the user association problem (Slave) from the cache leasing and content placement problem (Master) in an iterative solution process. The technique is completely original for solving NUM edge-caching problems with non-linear utilities.

• The Master sub-problem is mixed-integer-linear and numerically tractable by commercial solvers. The Slave subproblem is convex under various association policies, including CLOSEST and OPT. We introduce a distributed solution for the Slave, based on its Augmented Lagrangian. The novel algorithm that calculates the optimal user association is computationally very efficient, and is named Generalized Bucket-Filling.

• We provide extensive evaluation of the optimal leasing and content placement for linear and concave objective functions. The evaluated metrics are the (weighted) hit ratio and the network throughput; the latter depends on channel conditions. The hit probability or load-balancing benefits depend on the chosen association policy. Traffic offloading is more profitable when the MNO takes content placement into account (OPT), compared to content-agnostic strategies (CLOSEST).

• Our evaluations can derive the optimal price that the MNO should set for its revenue maximization. Furthermore, they can suggest appropriate investment budget for the CP to attain a target hit-ratio.

The remainder of this paper is organized as follows: In Section II, we survey relevant literature to our problem. The system model is developed in Section III where we also state the general mixed integer NUM problem. The general solution based on Benders decomposition is provided in Section IV. Special formulations of the problem with linear or separable concave objective functions as well as different association policies (CLOSEST, OPT) are presented in Section V. The section also includes alternative performance metrics, such as wireless throughput. Section VI presents the Generalized Bucket-Filling algorithm for solving the convex Slave in a distributed way, and provides a proof of optimality. Extensive numerical evaluation of problem variations and analysis of the findings is given in Section VII. Finally, Section VIII concludes our work.

II. Related Literature

There is extensive literature on the advantages of caching in wireless networks. An important body of work concerns the joint edge-cache placement and wireless user association problem. One can identify two different content placement approaches: dynamic caching [6], [8], [12]–[18], and prefetching [4], [5], [7], [10], [11], [19]–[21]. This work falls into the second category. Within the prefetching literature, some works have considered probabilistic cache placement [5], [22] while others study it in a deterministic way [4], [8], [23]. User association is either performed to the closest station [7] or more involved policies are applied that allow users to access content that is cached in all covering CBSs [4], [5], [8], [23].

More specifically, the original FemtoCaching problem (Shanmugam et al. [4]) assumes a bipartite connectivity graph of potential user associations. The aim is to place content such

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2018.2844986, IEEE Journal on Selected Areas in Communications

that user delay is minimized. However, the work does not consider the traffic loads at the femto caches and no specific user association policy is proposed. Baştuğ et al. [7] randomly place CBSs on the plane and users are associated to the closest station. They use metrics of outage probability and average delivery rate to analyse the performance of caching the most popular content. Poularakis, Iosifidis, and Tassiulas maximize in [11] the hit ratio by means of integer optimization. They introduce a bandwidth constraint limiting the number of users that can be connected to each cellular station. In their work, an approximation scheme for hit ratio maximization is provided. Naveen et al. [19] provide an optimal placement and user association scheme with fractional content placement. Deghan et al. [20] also develop an approximation algorithm for the content placement problem minimizing network delay. Their model controls whether users should be routed through a cached or an uncached path. The users on a cached path are always associated to the closest cache storing the content. In [5], Błaszczyszyn and Giovanidis develop a probabilistic content placement policy which maximizes the hit ratio by profiting from multi-coverage. Fully distributed content placement strategies that use LRU variations for multi-coverage are proposed in [15]. In previous work [1], we optimize the user association assuming that content is already placed in caches. Tuholukova, Neglia and Spyropoulos [24] investigate optimal content placement for joint transmission schemes in small cell networks. Other works on CoMP include [25] and [26]. Liu et al. [21] develop a distributed content placement algorithm for different transmission schemes. Alfano et al. [27] include power control issues in this general problem.

The business aspects of edge-caching with its potentials and limitations are comprehensively discussed by Paschos et al. [9]. Cache leasing schemes have been approached from a competitive [28], [29] and a cooperative [8], [30] perspective. In the latter work, Poularakis et al. propose offloading of backhaul traffic to local caches jointly optimizing caching incentive, content placement and routing policies. However, their Lagrangian based solution does not converge to the global optimum due to weak duality (see p. 143 in [31]) and other solution techniques are necessary to solve the problem optimally. In [10] and [17], the authors investigate blind cache splitting between several CPs. The partitioning of the caches remains under the control of the internet service provider, while the CPs are allowed to establish secure connections between caches and users. Liu et al. [21] develop a cache leasing system for small base stations within the framework of contract theory without considering optimal content placement. In the context of ICN, joint cache partitioning and resource allocation has been proposed recently [32].

For the solution of our optimization problem we will use Generalized Benders decomposition. In the context of Content Delivery Networks, Bektaş et al. [33] use linear Benders decomposition for joint placement and routing problems with binary variables. The main difference to our work is that we treat a wireless network and we solve mixed-integer non-linear NUM problems that also includes leasing aspects. The decomposition in our paper has a natural business interpretation. Other works in the wireless networking literature that use the Benders decomposition technique in a different context include [34] and [35]. For an application to optical networks, see [36].

III. System Model and Problem Statement

A. Cache Leasing and Content Placement

We consider a cellular communications network with a finite set $\mathcal{M}$ of CBSs. Each CBS m is equipped with k_m memory units of size b_{MU} (in MBytes, e.g. 1000) which the CP can lease. Leasing and placement decisions are taken at the beginning of a long time window (and stay fixed throughout) during which content popularity statistics are assumed static. Denoting the decision variable of how many cache units to lease (see CP-1) at m by $z_m \in \mathbb{Z}_{\geq 0}$, the bounded availability of memory gives the constraint set

$$z_m \leq k_m, \ \ \forall\, m \in \mathcal{M}. \tag{1}$$

The vector of the cache leasing variables is $\mathbf{z} = (z_m)_{m\in\mathcal{M}}$.

Having leased cache space at the CBSs, the CP places content from a finite object catalog $\mathcal{F}$ into the caches (see CP-2). The decision to store content f in the cache of m will set the variable $x_{m,f}$ to 1, otherwise $x_{m,f} = 0$. The vector of content placement variables is $\mathbf{x} = (x_{m,f})_{m\in\mathcal{M}, f\in\mathcal{F}}$. Each file has a given file size b_f (in MBytes), and all file-sizes are known. The limited capacity of the leased cache space gives the second constraint set

$$\sum_{f\in\mathcal{F}} b_f x_{m,f} \leq b_{\text{MU}} z_m, \ \ \forall\, m \in \mathcal{M}. \tag{2}$$

For convenience, we define the set of feasible tuples of leasing and placement vectors as

$$\mathcal{X} := \left\{(\mathbf{x},\mathbf{z}) \in \{0,1\}^{|\mathcal{M}||\mathcal{F}|} \times \mathbb{Z}_{\geq 0}^{|\mathcal{M}|} \;\middle|\; (1),(2)\right\}.$$

B. Wireless Environment and User Association

Coverage Cells: Our communications model is the following: Each CBS has a planar 2D coverage cell. Users covered by a CBS receive a radio signal strong enough to be potentially associated to it. Coverage cells may overlap, thus offering the users multiple options for service from covering CBSs. However, we do not allow simultaneous service by more than one station, i.e. cooperative service is not possible.

Network Regions: The network area is partitioned into a set of regions. All positions in each region are assumed to experience the same radio conditions with respect to fading and interference. Furthermore, the MNO has a user association policy Π that allows for users in region s to potentially be associated to any CBS in $\mathcal{M}(s) \subseteq \mathcal{M}$, $|\mathcal{M}(s)| \geq 1$. (Π1) With the traditional policy join-the-closest-station (Closest), $\mathcal{M}(s)$ is just the closest covering station. For the Opt policy, to be defined later, $\mathcal{M}(s)$ is the set of covering CBSs. In general, for policy Π, the set of regions is denoted by $\mathcal{S}^{\Pi}$ (see Fig. 2).

Content Popularity: For each region $s \in \mathcal{S}^{\Pi}$ and each content $f \in \mathcal{F}$, the expected number of users in s requesting f is considered to be known. It is denoted by $N_{s,f}$. The content popularity vector is $\mathbf{N} = (N_{s,f})_{s\in\mathcal{S}^{\Pi}, f\in\mathcal{F}}$. It can be measured or estimated by a combination of user location and mobility statistics, mobile application usage, as well as the history of past cache queries, see [37].

User Association Variables and Constraints: In order to make optimal decisions, the CP has two types of information at its disposal: the popularity vector $\mathbf{N}$ and the MNO's user association policy Π. Knowledge of $\mathbf{N}$ and Π allows the CP to take decisions based on the expected association of users with CBSs. We assume that the statistics of content popularity are considered static during a time window. Decisions for content placement are taken at the beginning of this window and remain fixed throughout. Since at the beginning of the window, future user positions are yet unknown, we introduce expected association variables. The implemented protocol should associate users to CBSs with frequency equal to the optimal expected value.

In the context of this work, we are only interested in *the users who find their request cached at the CBS they are associated to (cache-hit traffic)*. The association vector of *cache-hit* users to the CBSs is $\mathbf{y} = (y_{m,s,f})_{m\in\mathcal{M},s\in\mathcal{S}^{\Pi}(m),f\in\mathcal{F}}$, where $y_{m,s,f}$ represents the expected user traffic from region s requesting content f and associated with CBS m. $\mathcal{S}^{\Pi}(m)$ is the subset of regions whose users can potentially be associated to m according to Π. The vector $\mathbf{y}$ has fractional non-negative entries.

User association is unique in the sense that a single user cannot be served by two or more CBSs simultaneously. The total population $N_{s,f}$ can be distributed among the CBSs $\mathcal{M}(s)$, and some of it is potentially not associated to any CBS at all. Thus,

$$\sum_{m\in\mathcal{M}(s)} y_{m,s,f} \leq N_{s,f}, \;\; \forall\, s \in \mathcal{S}^{\Pi}, f \in \mathcal{F}. \tag{3}$$

This constraint allows for possible splitting of the population $N_{s,f}$ among the CBSs in $\mathcal{M}(s)$. The set of assignment vectors feasible to this constraint set is denoted by

$$\mathcal{Y}^{\Pi} := \left\{\mathbf{y} \in \mathbb{R}_{\geq 0}^{\sum_{m\in\mathcal{M}}|\mathcal{S}^{\Pi}(m)||\mathcal{F}|} \mid (3)\right\}.$$

Since we are only interested in cache-hit traffic, $y_{m,s,f}$ can only be nonzero if $x_{m,f} = 1$, i.e. if object f is cached in station m. Since no more than the total population requesting content f in s can be included in $y_{m,s,f}$, the following constraint set is valid:

$$y_{m,s,f} \leq N_{s,f} x_{m,f}, \;\; \forall\, m \in \mathcal{M}, s \in \mathcal{S}^{\Pi}(m), f \in \mathcal{F}. \tag{4}$$

This constraint set relates MNO association variables with CP cache placement decisions.

C. CP savings

The CP uses the general *savings function* $\mathrm{h}(\cdot)$ to measure user association $\mathbf{y}$. This function represents the savings (in €) obtained when users are associated with caches that store the requested content, thus avoiding use of its data centers. In this paper, we consider any monotonously increasing, continuously differentiable and concave savings function. Two choices for $\mathrm{h}(\cdot)$ are particularly of interest:

i) In case that the CP is solely interested in maximizing the hit ratio, it can choose $\mathrm{h}(\cdot)$ as a *linear* function.

ii) Choosing $\mathrm{h}(\cdot)$ as the sum of *strictly concave* functions (one function per CBS), the CP can include aspects such as soft resource requirements and load-balancing.

The discussion over particular choices of $\mathrm{h}(\cdot)$, which result in problems with different objectives, is postponed to Section V. The particular choices allow additional communication conditions (e.g. fading and interference) to be included in h.

D. MNO policy

The way the MNO associates users to CBSs determines the association vector $\mathbf{y}$, and depends on the MNO's user association policy Π. Examples for such policies are:

1) CLOSEST: Association to the closest covering CBS.

2) OPT: Association maximizing the CP's savings function.

Observe that (4) implies that cache-hits (vector $\mathbf{y}$) depend on the placement $\mathbf{x}$ in all cases. The resulting association vector is denoted by $\mathbf{y}^{\Pi}(\mathbf{x})$. If $\Pi =$ CLOSEST, the association entry $y_{m,s,f}$ is positive only for users that find their content cached. However, association actions do not depend on placement $\mathbf{x}$.

On the other hand, if $\Pi =$ OPT, the MNO fully cooperates with the CP in the sense that it always adapts its association vector $\mathbf{y}$ to the placement $\mathbf{x}$ such that the CP's savings function h is maximized. This is achieved by splitting traffic among CBSs given multi-coverage.

Remark 1. *For practical purposes, there are several ways in which the MNO can associate users in a cache-aware manner while maintaining user privacy. The MNO can initially associate each user to a large set of covering stations but only serve the user from a CBS having the content. Alternatively, associate each user to a single CBS and redirect afterwards to a station with the content. In both cases, after initial association, the CP becomes aware of the user request and communicates the appropriate serving station to the MNO.*

For the two policies $\Pi =$ OPT and $\Pi =$ CLOSEST, the set of association regions $\mathcal{S}^{\Pi}$ is different as explained in III-B (User Association). We illustrate this difference in the example of Fig. 2. For both policies, the association vector is the optimal solution to the User Association problem

$$\text{(UA-}\Pi\text{)} \qquad \mathbf{y}^{\Pi}(\mathbf{x}) = \underset{\mathbf{y}\in\mathcal{Y}^{\Pi}}{\arg\max} \qquad \mathrm{h}(\mathbf{y}) \qquad \text{s.t.} \qquad (4).$$

This problem is convex, thus always tractable. Note, however, that for CLOSEST, the above problem is trivial. The solution is simply the covered users within the Voronoi cell of each CBS that request cached content. We will come back to a general solution of the UA for OPT in Section VI.

E. Problem statement

The objective of the CP is to lease cache memory at the CBSs and place content into it such that the relation of its expected savings to the leasing cost is optimal. As mentioned, the savings are given by the function $\mathrm{h}(\cdot)$ that takes as input the user association vector $\mathbf{y}^{\Pi}(\mathbf{x})$ where $\mathbf{x}$ is the content placement action and Π is the MNO's association policy. The leasing costs at each CBS m are the product of leased units

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2018.2844986, IEEE Journal on Selected Areas in Communications

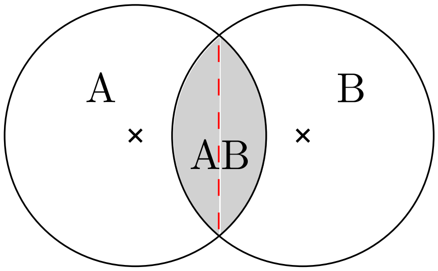

Fig. 2: In case $\Pi = \textsc{Opt}$, there are three regions A, B and AB. Users in region A and B can only be associated to their uniquely covering CBSs, respecively. Users in region AB can potentially be associated to any of the two CBSs. If $\Pi = \textsc{Closest}$ (dashed line), there are two regions: A and the left part of AB contain traffic entirely associated to the left CBS, B and the right part of AB contain traffic belonging to the right CBS.

z_m times the price per unit q_m that is set by the MNO. Through this price, the CP is charged for making use of cache memory. An additional fee for the appropriate user association and content delivery can be included. Formally, the CP seeks a feasible tuple of vectors $(\mathbf{x}, \mathbf{z}) \in \mathcal{X}$ that maximizes the objective function $\mathrm{h}(\mathbf{y}^{\Pi}(\mathbf{x})) - \sum_{m\in\mathcal{M}} q_m z_m$.

The CP's Cache Leasing and Content Placement problem (CLCP) can be formulated as the Non-Linear Mixed-Integer Problem (NLMIP)

$$\text{(CLCP)} \max_{\substack{(\mathbf{x},\mathbf{z})\in\mathcal{X}\\ \mathbf{y}\in\mathcal{Y}^{\Pi}}} \quad \mathrm{h}(\mathbf{y}) - \sum_{m\in\mathcal{M}} q_m z_m$$
$$\text{s. t.} \quad y_{m,s,f} \leq N_{s,f} x_{m,f}, \ \forall\, m, s, f,$$

where m is a CBS, s is a planar network region and f is a data file.

Special Case: For zero costs and unit file size under the $\textsc{Closest}$ policy with linear savings function h, it is optimal to store k_m locally Most Popular Content in each CBS m.

F. Complexity

Even with a linear savings function h and without taking cache leasing into account, the CLCP problem is NP-hard.

Proposition 1. *The CLCP problem is NP-hard.*

Proof. Assuming that h can be evaluated in polynomial time, a certificate can be checked in polynomial time, thus the problem is in NP.

Now, we show a polynomial-time reduction from the Helper Decision problem (HD) presented in [4]. The reduction identifies users (in HD) with regions (in CLCP) and helpers (HD) with CBSs (CLCP). An instance of HD is transformed into a CLCP instance in the following way: Setting $b_f = b_{\mathrm{MU}} = 1$ and $k_m = M$ and all prices q_m to 0 eliminates the variables z_m and provides that (1) and (2) are equivalent to the capacity constraint in HD.

Choose $\mathrm{h}(\mathbf{y}) = \sum_{s\in\mathcal{S}} \tilde{w}_s \sum_{f\in\mathcal{F}} \sum_{m\in\mathcal{M}(s)} y_{m,s,f}$ where the $\tilde{w}_s$ correspond to the weights in HD. Since $\tilde{w}_s > 0$ for all s by definition in HD, and since $y_{m,s,f} \leq N_{s,f} x_{m,f}$ as well as (4) for all m, s, f, then whenever a CBS covering a region s stores the requested content f, all user traffic from the region will be associated with some CBS. Thus, $\sum_{m\in\mathcal{M}(s)} y_{m,s,f} = N_{s,f}$ if $x_{m,f} = 1$ for some $m \in \mathcal{M}(s)$ and 0 otherwise. Choosing $N_{s,f}$ (CLCP) as P_f (HD) for all regions s shows that, defined this way, the objective function of CLCP is equivalent to the objective function of HD.

This is a polynomial time transformation which concludes the proof. Note that in [4] NP-completeness is proved for catalog size two. □

IV. Solution

The CLCP problem is very difficult to be solved even numerically by existing software, due to its high complexity. It is a mixed-integer problem with non-linear objective. We thus need to proceed analytically. The solution technique that resolves this problem is Generalized Benders decomposition by Schrijver [38] and Geoffrion [39] which converges to the global optimum. This method can decompose our problem in a way that removes the non-linearity from the discrete problem. The decomposed discrete problem, called Master, is linear and thus simpler to deal with. As a second by-product there appears a continuous convex subproblem called Slave, which can be solved by standard techniques (e.g. Lagrangian). Observe that due to Proposition 1, our solution algorithm cannot be polynomial even with linear savings function unless P = NP. However, a state-of-the-art MIP solver can be used for the iterative solution of Master. Its performance is shown in Section VII. In what follows, we give an overview over Generalized Benders decomposition applied to CLCP.

CLCP can be decomposed into two problems called Master and Slave. Master decides about cache leasing and content placement in the prefetching phase. Slave computes the optimal user association for a fixed content placement in the delivery phase. We obtain

$$\text{(Master)} \qquad \max_{(\mathbf{x},\mathbf{z})\in\mathcal{X}} \qquad \mathrm{h}(\mathbf{y}(\mathbf{x})) - \sum_{m\in\mathcal{M}} q_m z_m,$$

where $\mathrm{h}(\mathbf{y}(\mathbf{x}))$ is the objective value of

$$\text{(Slave)} \qquad \mathbf{y}(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}^{\Pi}} \qquad \mathrm{h}(\mathbf{y})$$
$$\text{s. t.} \qquad (4).$$

Note that the Master problem can be treated by the CP, the Slave by the MNO. $\mathcal{X}$ is discrete and finite and $\mathcal{Y}^{\Pi}$ is compact and convex. Slave is the UA-Π problem from Section III-D which is generally non-linear. Thus, Master cannot be solved directly. It can be solved, however, by following an iterative procedure that deals with this problem by solving a sequence of Slave problems for different values of $\mathbf{x}$ (and $\mathbf{z}$). The solutions to Slave are used to construct (linear) Benders cuts that constitute approximations to Slave. The Benders cuts are iteratively introduced as constraints of the *Surrogate Problems* which are linear approximations to Master. With each iteration, the approximation improves until the optimal solution of Master is found.

A. Benders Cuts

Let $\{(\mathbf{x}^t, \mathbf{z}^t) \in \mathcal{X} \mid t = 1, \ldots, T\}$ be a set of vector tuples feasible to Master for some $T \geq 0$. Let $\mathbf{y}^t := \mathbf{y}(\mathbf{x}^t)$ denote a corresponding vector that optimizes Slave for given $\mathbf{x}^t$. Let

$\boldsymbol{\lambda}^t = (\lambda^t_{m,s,f})_{m\in\mathcal{M},s\in\mathcal{S}^\Pi(m),f\in\mathcal{F}}$ be the vector of Lagrangian multipliers corresponding to the constraints (4).

Slave is a convex problem. Thus, the duality theorem of convex programming implies

$$\begin{aligned} \mathrm{h}(\mathbf{y}(\mathbf{x})) \leq \quad & \mathrm{h}(\mathbf{y}^t) \\ + & \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}^\Pi(m)} \sum_{f\in\mathcal{F}} \lambda^t_{m,s,f}(N_{s,f}x_{m,f} - y^t_{m,s,f}) \end{aligned}$$

for all feasible vectors $\mathbf{x}$. This upper bound to Slave is called *Benders cut*. Reformulated, we get

$$\mathrm{h}(\mathbf{y}(\mathbf{x})) \leq \Gamma^t + (\mathbf{v}^t)'\mathbf{x}, \tag{5}$$

where $\Gamma^t := \mathrm{h}(\mathbf{y}^t) - \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}^\Pi(m)} \sum_{f\in\mathcal{F}} \lambda^t_{m,s,f} y^t_{m,s,f}$ and $(\mathbf{v}^t)'$ is the transpose of $\mathbf{v}^t = (v^t_{m,f})_{m\in\mathcal{M},f\in\mathcal{F}}$ with $v^t_{m,f} = \sum_{s\in\mathcal{S}^\Pi(m)} \lambda^t_{m,s,f} N_{s,f}$.

B. Surrogate Problem

For a set of T Benders cuts (for some $T \geq 1$), we obtain an upper bound to the original problem CLCP by solving the *Surrogate* IP

$$\begin{aligned} (\text{SUR-}T) \max_{\substack{(\mathbf{x},\mathbf{z})\in\mathcal{X} \\ \gamma\in\mathbb{R}_{\geq 0}}} \quad & \gamma - \sum_{m\in\mathcal{M}} q_m z_m \\ \text{s. t.} \quad & \gamma \leq \Gamma^t + (\mathbf{v}^t)'\mathbf{x} \qquad t = 1,\ldots,T. \end{aligned}$$

The optimal objective value to the surrogate problem is denoted by $\textsc{sur}^T$. The optimal solution consists of $\mathbf{x}^{T+1}$, $\mathbf{z}^{T+1}$ and γ^{T+1}. Note that the auxiliary variable γ, together with the Benders cuts, approximates Slave linearly. This way, SUR-T avoids the non-linearity which creates the difficulty for solving Master.

C. Benders Iteration and Convergence

Generalized Benders decomposition is an iterative process. We start in step 0 with an initial feasible tuple of leasing and placement vectors $(\mathbf{x}^0, \mathbf{z}^0) \in \mathcal{X}$ and without any Benders cuts. At the start of step $T \geq 0$, we have the current leasing and placement $(\mathbf{x}^T, \mathbf{z}^T) \in \mathcal{X}$ and T Benders cuts.

Slave: We solve Slave with input $\mathbf{x}^T$ to obtain the optimal association vector $\mathbf{y}^T$. Clearly, since the triple $\mathbf{x}^T$, $\mathbf{z}^T$ and $\mathbf{y}^T$ are feasible to the original problem CLCP, its corresponding objective value provides a lower bound to the optimal value of CLCP. Additionally, we compute the Lagrangian multipliers $\boldsymbol{\lambda}^T$ and the corresponding $(T+1)$-th Benders cut (5).

Surrogate: With the Benders cut we obtain the surrogate MIP SUR-$(T+1)$. Its optimal solution is the feasible leasing and placement vectors $(\mathbf{x}^{T+1}, \mathbf{z}^{T+1}) \in \mathcal{X}$. The objective value $\textsc{sur}^{T+1}$ is an upper bound to CLCP.

This process iterates. At any step T, $\textsc{sur}^T$ is the current upper bound (note that $\textsc{sur}^{T+1} \leq \textsc{sur}^T$ after every step T), while the current lower bound is provided by the best found solution $\max_{t\in\{0,\ldots,T\}} \mathrm{h}(\mathbf{y}^t) - \sum_{m\in\mathcal{M}} q_m z^t_m$. The process terminates in the globally optimal cache leasing and content placement vectors $\mathbf{z}^*$ and $\mathbf{x}^*$ when the upper and lower bounds coincide. Convergence is guaranteed from the proof of Theorem 2.4 in [39] and the fact that the domain $\mathcal{X}$ of the Master is finite.

In every step T, an instance of the Surrogate problem needs to be solved. In general, the Surrogate problem is not easily tractable as to the following proposition.

Proposition 2. *For $T \in \mathbb{N}$ and general parameters $\Gamma^t \in \mathbb{R}$ and $\mathbf{v}^t \in \mathbb{R}^{|\mathcal{M}||\mathcal{F}|}$, $t = 1,\ldots,T$, $q_m \in \mathbb{R}$ for $m \in \mathcal{M}$ as well as domain $\mathcal{X}$, the Surrogate problem SUR-T is NP-hard.*

Proof. Objective function and constraints can be evaluated in polynomial time, thus the problem is in NP.

Now, we reduce SET COVER to SUR-T. In the SET COVER decision problem, the question is if there is a subset $\mathcal{S}$ of a collection of sets $\mathcal{F} \subseteq 2^{\mathcal{T}}$ with cardinality less than or equal to S such that the $\bigcup_{\mathcal{S}} = \mathcal{T}$ where $\mathcal{T}$ is the universe . For the reduction, we identify the Benders cuts t with the elements in the universe $\mathcal{T}$. The files f are in bijection to the sets in the family $\mathcal{F}$. Note that then, the SET COVER decision problem can be written as: Are there

$$\begin{aligned} \gamma \in \mathbb{R}, \mathbf{x} \in \{0,1\}^{|\mathcal{F}|} \quad & \text{with} \\ \gamma \geq 1 & \\ \gamma \leq \sum_{f\in\mathcal{F}} \mathbb{1}_{t\in f} x_f \quad & \forall\, t \\ \sum_{f\in\mathcal{F}} x_f \leq S, & \end{aligned}$$

where $\mathbb{1}_{t\in f} = 1$ if $t \in f$ and 0 otherwise? Here, x_f is a binary variable taking value 1 if and only if f is element of $\mathcal{S}$ and γ represents the lowest frequency of any element of $\mathcal{T}$ in $\mathcal{S}$.

This is an instance of the SUR-T decision problem for only one CBS (omitting the index m): The price is chosen as $q_m = 0$. The memory capacity is chosen as $k_m = S$, the file sizes as well as memory are all unit size ($b_f = b_{\text{MU}} = 1$). Then, the variable z_m can be omitted since in the optimum, $\sum_{f\in\mathcal{F}} x_f = z_m$. The constraint $\sum_{f\in\mathcal{F}} x_f \leq S$ is the combination of the constraints (1) and (2). Constraint set $\gamma \leq \sum_{f\in\mathcal{F}} \mathbb{1}_{t\in f} x_f$ comes from choosing $\Gamma^t = 0$ for all t and $v^t_{m,f} = \mathbb{1}_{t\in f}$.

The reduction is performed in polynomial time. This concludes the proof. □

Despite the fact that SUR-T is NP-hard, state of the art MIP solvers such as CPLEX are capable of solving SUR-T in reasonable runtime. Approximation algorithms for SUR-T are a topic for future research.

D. Implementation Considerations

Here, we provide a short discussion about information exchange between the entities solving Slave and SUR-T towards the global optimum. The SUR-T problem that solves the leasing and content placement aspect requires information on the price per cache unit. For each Benders cut, the vector of Lagrangian multipliers (dual prices) of the user association together with the optimal objective value of Slave need to be communicated. Knowledge of the spatial popularity statistics in this phase is reasonably assumed.

On the other hand, the Slave problem as formulated needs to know the exact utility h of the CP together with a subset of the popularity statistics which is related to the content the CP aims to cache. Hence, although the content placement and user association decisions are treated separately in two intertwined optimization problems, Slave does require sensitive information from the CP.

To overcome the implicit conflict of interest, a first suggestion can be that the MNO informs the CP about its general association policy Π to cooperate or not, together with the cache price. The CP can then solve both problems and after having identified the optimal association vector it can give association suggestions to the MNO each time a potential cache-related user emerges (as proposed in Remark 1). Obviously, such approach requires a more integrated interaction between the two actors. A second proposal could be that the MNO solves Slave by using estimates on the CP savings function and popularity data. The resulting association solution communicated to the CP via the Lagrangian multipliers will generate a different set of Benders cuts for the CP. Obviously, in this case the final solution is suboptimal. An interesting extension of the current research would be to mathematically investigate the quality of such solution.

V. Special Cases

The savings function h introduced in Section III-C maps the user association vector $\mathbf{y}$ onto the savings $\mathrm{h}(\mathbf{y})$ (in €). In the following, we give examples of specific expressions for $\mathrm{h}(\cdot)$ and explain what each example implies for the solution. We denote the cache-hit traffic volume associated to CBS m through vector $\mathbf{y}$ by

$$v_m(\mathbf{y}) = \sum_{s\in\mathcal{S}^{\Pi}(m)} \sum_{f\in\mathcal{F}} y_{m,s,f}. \tag{6}$$

A. Linear Savings (Case i in Simulations)

As a first case, the cache-hit user traffic of CBSs is linearly mapped onto monetary benefits for the CP, i.e.

$$\mathrm{h}^{\mathrm{L}}(\mathbf{y}) = c \sum_{m\in\mathcal{M}} v_m(\mathbf{y}), \tag{7}$$

where c is the savings (in €) per cache hit. Since the popularity vector is constant, the savings value is proportional to the hit ratio. The latter is simply calculated by $\mathrm{h}^{\mathrm{L}}(\mathbf{y})/(c\sum_{s\in\mathcal{S}^{\Pi}}\sum_{f\in\mathcal{F}} N_{s,f})$.

With a linear savings function, the slave problem becomes easily tractable for any association policy Π, including Closest and Opt: Given a CP vector $\mathbf{x}$, the MNO can freely distribute users among the covering CBSs $\mathcal{M}(s)$ which have f cached. The association to any CBS contributes equally to the savings. If no $m \in \mathcal{M}(s)$ caches f, then $y_{m,s,f} = 0$, hence no cache hit from region s for file f.

B. Separable Concave Savings (Case ii in Simulations)

As a second case, we introduce the sum of strictly concave functions, one per CBS, taking as argument the associated traffic volume. The strict concavity of the utility functions implies diminishing returns for user traffic in every CBS. This way, the MNO has the incentive to associate users with underused CBSs while the overuse of CBSs is disincentivized. This choice for h can model physical resource limitations on each CBS that prohibit the good service of users when their volume becomes very high.

Formally, we define utility functions $\mathrm{U}_m(\cdot)$ for every $m \in \mathcal{M}$ which are monotonously increasing, strictly concave and continuously differentiable. The input of the utility functions is the cache-hit traffic volume at the CBS. The savings function is the sum of all utility functions. We obtain

$$\mathrm{h}^{\mathrm{N}}(\mathbf{y}) = c \sum_{m\in\mathcal{M}} \mathrm{U}_m(v_m(\mathbf{y})), \tag{8}$$

where c is the scaled utility of cache-hits (in €).

The load can be balanced among the CBSs by guaranteeing fairness with regards to associated volume. Some notions of fairness are max-min, $\alpha-$ and proportional fairness. Each of them is achieved by appropriate choice of the utility functions (see [40], [41]). E.g. for proportional fairness, utilities are chosen as (weighted) logarithms.

For any association policy Π, the slave problem with savings function as in (8) is a convex problem.

C. Weighted Savings

In the previous section, there was no differentiation between wireless users in the objective function. Here, we introduce weights $w_{m,s,f} \geq 0$ that are specific to users from region s requesting content f associated to CBS m. This generalization allows to include costs and benefits from associating certain user groups to particular stations. The weighted traffic volume at CBS m is

$$v_m^{\mathbf{w}}(\mathbf{y}) = \sum_{s\in\mathcal{S}^{\Pi}(m)} \sum_{f\in\mathcal{F}} w_{m,s,f}\, y_{m,s,f}, \tag{9}$$

where $\mathbf{w}$ is the vector of weights. A weighted version of the linear $h^{\mathrm{L},\mathbf{w}}(\mathbf{y})$ and concave $h^{\mathrm{N},\mathbf{w}}(\mathbf{y})$ savings function can then be introduced.

a) Prioritized Caching: If the weights $w_{m,s,f}$ are proportional to the file sizes b_f, files of larger size that create more burden to the backbone are favorized to be cached. If they are inversely proportional, files of smaller size are preferred. Another work with similar objectives is [14].

b) FemtoCaching: The users in [4] are equivalent to the regions as defined in our work. By choosing $w_{m,s,f}$ as the delay weights (see [4, maximization (3)]) together with zero leasing costs, FemtoCaching is a special case of linear weighted CLCP (see also proof of Proposition 1 in Section III).

D. Wireless Throughput (Case iii in Simulations)

The weights can represent the downlink throughput between a user and a CBS (other work with the same objective is [25]). For such a weights choice, the channel quality between m and s is a constant value $h_{m,s}$ that depends on a reference distance and the path loss exponent. The emitted power level of m is denoted by p_m and the noise level by σ^2. Then, the signal-to-interference-plus-noise ratio (SINR) of users in region s when associated to covering CBS m is

$$\mathrm{SINR}_s(m) = p_m h_{m,s}\Bigg(\sum_{\substack{\tilde{m}\text{ covers } s\\ \tilde{m}\neq m}} p_{\tilde{m}} h_{\tilde{m},s} + \sigma^2\Bigg)^{-1}, \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2018.2844986, IEEE Journal on Selected Areas in Communications

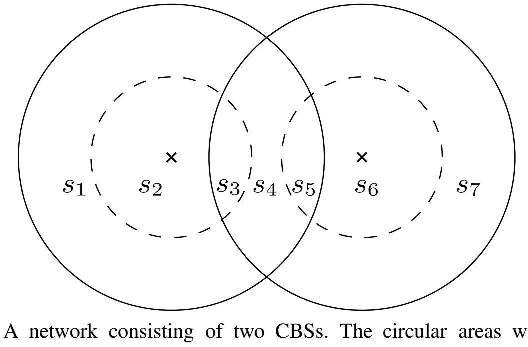

Fig. 3: A network consisting of two CBSs. The circular areas with solid line depict the areas around each CBS, where the received signal emitted by this station is above a certain threshold. For each CBS, we choose here to differentiate between two zones of signal strength (strong, weak), separated by the dashed lines. In the regions where there is disc overlap, users experience *interference* from the CBS they are not associated to. We can identify in this way 7 different regions with different overlaps.

where we assume that the interference from CBSs not covering s is negligible. For the downlink transmission from CBS m to region s, the throughput is equal to

$$w_{m,s} = B\log_2(1+\mathrm{SINR}_s(m)) \quad \text{[in bits/sec]} \tag{11}$$

where B [Hz] is a chunk of bandwidth allocated to each served user. The total service bandwidth per CBS is equal to the product of B [Hz] times the users routed to the CBS. Using the value (11) as weights, CLCP takes into account that it is favorable for a CBS to serve users with good radio conditions in order to use its resources effectively. Note that, we can also define other weights that depend on file f, for example $w_{m,s,f} = g(w_{m,s}/b_f)$, where $g(\cdot)$ is some increasing function and b_f is the file size. Such expression would then evaluate throughput over the requested file-size, giving larger weight w to (and thus favoring) smaller file sizes. Fig. 3 shows an example of network regions with corresponding wireless performance weights (SINR values one station's signal as beneficial and the other's as interference). We would like to emphasize that the arbitrarily many levels of signal strength on the coverage area of each station can be introduced by appropriately redefining the area partition $\mathcal{S}$. The modelling tradeoff is between the precision of communications aspects and the runtime of the optimization process.

VI. Distributed Solution Algorithm for Slave

We present a distributed solution to the user association problem when the most general utility function is chosen as

$$\mathrm{h}^{\mathrm{N},\mathbf{w}}(\mathbf{y}) = c\sum_{m\in\mathcal{M}} \mathrm{U}_m(v_m^{\mathbf{w}}(\mathbf{y})). \tag{12}$$

The special case of unweighted utilities as in (8) has been considered in previous work [1].

Following the Slave problem formulation, cache leasing and content placement decisions are assumed to be fixed. Only users located in a covered region $s\in\mathcal{S}(m)$ and requesting cached content with $x_{m,f}=1$ can be associated to cache m. By eliminating all variables $y_{m,s,f}$ that are forced to be 0 for Slave, the constraints (4) are redundant and are omitted.

1) Dual method for the Augmented Lagrangian: We solve the user association problem using the dual method on the Augmented Lagrangian (see [42], Section 3.4.4) relaxing constraints (3). To achieve a distributed solution, the regular Lagrangian is not appropriate since the objective is not strictly concave in the primal variables and hence the primal solution is not unique. This creates conflicts when different stations compete for the same users in the solution process of the primal subproblem, i.e. taking distributed decisions may not be globally feasible. However, convergence can be guaranteed for the Augmented Lagrangian

$$\begin{aligned} L^{(\varrho)}(\mathbf{y},\boldsymbol{\lambda}) = c&\sum_{m\in\mathcal{M}} \mathrm{U}_m(v_m^{\mathbf{w}}(\mathbf{y}_m)) \\ &-\sum_{s\in\mathcal{S}}\sum_{f\in\mathcal{F}}\lambda_{s,f}(N_{s,f}-\sum_{m\in\mathcal{M}(s)} y_{m,s,f}) \\ &-\frac{\varrho}{2}\sum_{s\in\mathcal{S}}\sum_{f\in\mathcal{F}}(N_{s,f}-\sum_{m\in\mathcal{M}(s)} y_{m,s,f})^2, \end{aligned} \tag{13}$$

where $\boldsymbol{\lambda} := (\lambda_{s,f}), s\in\mathcal{S}, f\in\mathcal{F}$ is the dual vector with $\lambda_{s,f}\geq 0$ and $\varrho > 0$ is a penalizing factor. In order to maximize (13) in a distributed way, we apply Diagonal Quadratic Approximation (DQA) to find the primal solution for given dual values (see [1] for details). In every DQA iteration, a separate primal problem needs to be solved.

2) Separated Primal Solution: Before stating the separated primal problem, we simplify the notation for this section. For convenience, we define the pool of potential users for the CBS m as $\mathcal{Q}(m) := \{(s,f) \mid s\in\mathcal{S}(m), f\in\mathcal{F}, x_{m,f}=1\}$. The index q will be used as equivalent to the double index s,f in this section, e.g. we write N_q instead of $N_{s,f}$. Furthermore, since the problem is separated by CBS, we omit the index m, writing y_q instead of $y_{m,s,f}$, $\mathbf{y}$ instead of $\mathbf{y}_m$, and $\mathcal{Q}$ instead of $\mathcal{Q}(m)$. The separated primal problem is then

$$\begin{aligned} \mathbf{y}^* &= \operatorname*{arg\,max}_{0\leq\mathbf{y}\leq\mathbf{N}} c\,\mathrm{U}(\sum_{q\in\mathcal{Q}} w_q y_q) + \sum_{q\in\mathcal{Q}}\lambda_q y_q - \frac{\varrho}{2}\sum_{q\in\mathcal{Q}}(\bar{N}_q - y_q)^2 \\ &= \operatorname*{arg\,max}_{0\leq\mathbf{y}\leq\mathbf{N}} c\,\mathrm{U}\Big(\sum_{q\in\mathcal{Q}} w_q y_q\Big) - \sum_{q\in\mathcal{Q}}\big[(\varrho/2)y_q^2 - a_q y_q\big] \end{aligned} \tag{14}$$

where $\bar{N}_q$ is the remaining user population not served by other CBSs in the previous DQA step (might be negative, see [1] for details), and $a_q := \lambda_q + \varrho\bar{N}_q$. The equation to (14) comes from the development of the quadratic term and from omitting the additive constants which do not affect the optimal choice of values for the variables. In the following, we will denote the objective function of (14) by $g(\mathbf{y})$.

Problem (14) is a convex optimization problem. Methods for solving this type of problem such as the gradient descent method are well known. However, their speed of convergence can be an issue. Here, we present a problem specific solution method that does not depend on convergence and solves (14) exactly and efficiently. Our method is based on the insight that the optimal solution lies within a one-dimensional subspace of its $|\mathcal{Q}|$-dimensional domain. The following theorem characterizes this subspace.

Proposition 3. *Let $\mathbf{y}^*$ be defined as in* (14). *Then, there exists $\nu^* \geq 0$ such that for all $q \in \mathcal{Q}$*

$$y_q^* = \begin{cases} 0, & \nu^* \leq \beta_q \\ N_q, & \nu^* \geq (w_{\tilde{q}}/w_q)N_q + \beta_q \\ f_q(\nu^*), & \textit{otherwise} \end{cases} \tag{15}$$

with

$$\beta_q = \frac{a_{\tilde{q}} - (w_{\tilde{q}}/w_q)a_q}{\varrho}. \tag{16}$$

In the above, $\tilde{q} := \arg\max_{q \in \mathcal{Q}} a_q/w_q$ and $f_q(\nu) = (w_q/w_{\tilde{q}})\,\nu - (w_q/w_{\tilde{q}})\beta_q$.

In other words, the optimal vector $\mathbf{y}^*$ can be determined by finding the optimal value ν^*. This fact considerably reduces the complexity of finding $\mathbf{y}^*$.

The value $\nu \geq 0$ can be interpreted as the *water level* in a scenario in which $|\mathcal{Q}|$ *buckets of varying width* $w_q/w_{\tilde{q}}$ and height $(w_{\tilde{q}}/w_q)N_q$ are positioned at different bottom levels. When the water level is below or exactly at bottom level β_q of bucket q, the bucket is empty. Otherwise, the bucket is filled up to level ν unless the water level is higher than the upper edge of the bucket at $(w_{\tilde{q}}/w_q)N_q + \beta_q$. In that case, the bucket is filled to its capacity of N_q. In between, the volume in the bucket is exactly $f_q(\nu)$. This observation leads to an algorithm solving (14) efficiently: Starting from water level $\nu = 0$, let the water level increase. The water volumes represent the variables in $\mathbf{y}$. The algorithm (see Algorithm 1 for detailed presentation) terminates when the optimal value ν^* is reached. This is true when the maximum of the objective function of (14) is reached or when all buckets are full. An example for such an arrangement of buckets is shown in Fig. 4 a). The proof of Proposition 3 can be found in the Appendix.

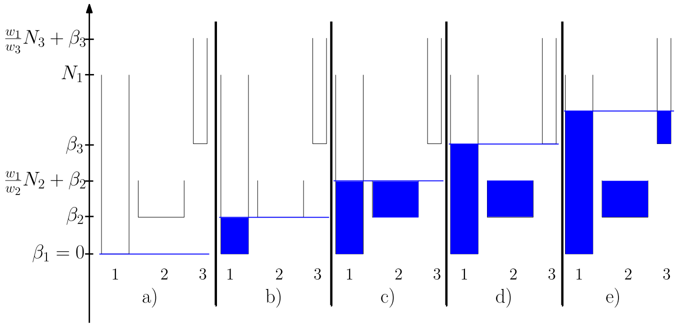

Fig. 4: Illustration of the Bucket-filling algorithm. Each bucket q is placed with its bottom at level β_q and has a width of $w_q/w_{\tilde{q}}$, a height of $(w_{\tilde{q}}/w_q)N_q$ and unit depth. The buckets are filled to a common level – or until they are full. The algorithm halts when further increasing the water level ν starts decreasing the objective value.

We can now describe the novel Generalized Bucket-filling algorithm that efficiently finds the optimal solution to UA-primal-m (see Algorithm 1).

Observe that step 2 can be done in $O(|\mathcal{Q}|\log(|\mathcal{Q}|))$ operations with a sorting algorithm such as quicksort. Thus, it dominates steps 1 and 3 that each need $O(|\mathcal{Q}|)$ operations. Step 5 is executed up to $2\,|\mathcal{Q}|$ times, since every region-file tuple is activated and deactivated no more than one time each. Assuming that f_q and $\partial/\partial y_q\ g(\mathbf{y})$ can be evaluated in $O(1)$, the runtime of the loop in steps 5 and 6 is $O(|\mathcal{Q}|)$. This shows that the overall runtime is dominated by sorting in step 2 and thus is $O(|\mathcal{Q}|\log(|\mathcal{Q}|)) = O(\hat{S}\hat{F}\log(\hat{S}\hat{F}))$ where $\hat{S}$ is the largest number of regions covered by any CBS and $\hat{F}$ the largest number of files that can be cached by any CBS.

Algorithm 1 Generalized Bucket-Filling (solves UA-primal-m with objective function (12))

```
1: Choose $\tilde{q} \in \mathcal{Q}$ such that $a_{\tilde{q}}/w_{\tilde{q}} \geq a_q/w_q$ for all $q$.
2: Sort $\mathcal{Q}$ by $\beta_q$ non-decreasingly.
3: Initialize the sets of *active* region-files as $\mathcal{Q}^A = \{q \in \mathcal{Q} \mid \beta_q = \beta_{\tilde{q}}\}$, *inactive* region-files $\mathcal{Q}^I := \mathcal{Q} \setminus \mathcal{Q}^A$ and *deactivated* region-files as $\mathcal{Q}^D := \emptyset$.
4: Set $\nu := 0$.
5: Increase $\nu$ until
     • $\nu = \beta_q$ for some inactive $q \in \mathcal{Q}$, then $\mathcal{Q}^A := \mathcal{Q}^A \cup \{q\}$ and $\mathcal{Q}^I := \mathcal{Q}^I \setminus \{q\}$
     • $\nu = (w_{\tilde{q}}/w_q)N_q + \beta_q$ for some active $q$, then $\mathcal{Q}^A := \mathcal{Q}^A \setminus \{q\}$ and $\mathcal{Q}^D := \mathcal{Q}^D \cup \{q\}$ or
     • $\partial/\partial y_q\ g(\mathbf{y}) = 0$ for all $q \in \mathcal{Q}^A$, $\partial/\partial y_q\ g(\mathbf{y}) \leq 0$ for all $q \in \mathcal{Q}^I$, and $\partial/\partial y_q\ g(\mathbf{y}) \geq 0$ for all $q \in \mathcal{Q}^D$, where $\mathbf{y} = (y_q)$ with $y_q = f_q(\nu)$ for $q \in \mathcal{Q}^A$, $y_q = 0$ for $q \in \mathcal{Q}^D$ and $y_q = N_q$ for $q \in \mathcal{Q}^I$.
6: If the last condition is fulfilled, return $\mathbf{y}$. Otherwise, go to step 5.
```

This algorithm is executed for every CBS in every DQA iteration which is why its efficiency is paramount.

VII. Experiments and Numerical Evaluation

A. Environment

We simulate cellular networks in an urban environment and calculate the optimal cache leasing and content placement for 6 cases which differ in savings function and user association policy: The savings function h is chosen as: i) Linear as in (7). ii) The sum of utility functions as in (8) where all utility functions U_m are chosen as the natural logarithm to achieve proportional fairness for the user traffic at the CBSs. iii) The weighted sum of utilities as in (12) where the weights are defined as in (11) and the utility functions are identities. This way, the sum of utilities is equal to the network throughput. For each of the three cases of savings functions, the MNO's user association policy is a) the MNO-CP cooperative policy Opt or b) the conventional policy Closest. In each case, we simulated 100 random sets of CBS positions as a Poisson Point Process (PPP). This means that their total number in each run is a random Poisson realization, and their positions are uniformly distributed in the simulation window. The density of the PPP is $0.8\frac{\text{CBS}}{\text{km}^2}$ for the cases i) (linear savings) as well as iii) (weighted linear savings) and $0.6\frac{\text{CBS}}{\text{km}^2}$ for the cases ii) (log-savings). This implies an average minimal distance of 560m and 650m between the CBS positions, respectively. For the cases i) and iii), the evaluation window has size 5000×5000m^2, while the cases ii) were evaluated in a 3000×3000m^2 window. The expected number of CBSs in the evaluated windows is 20 for case i) and iii) and 5.4 for case ii). In both cases, a larger area was simulated

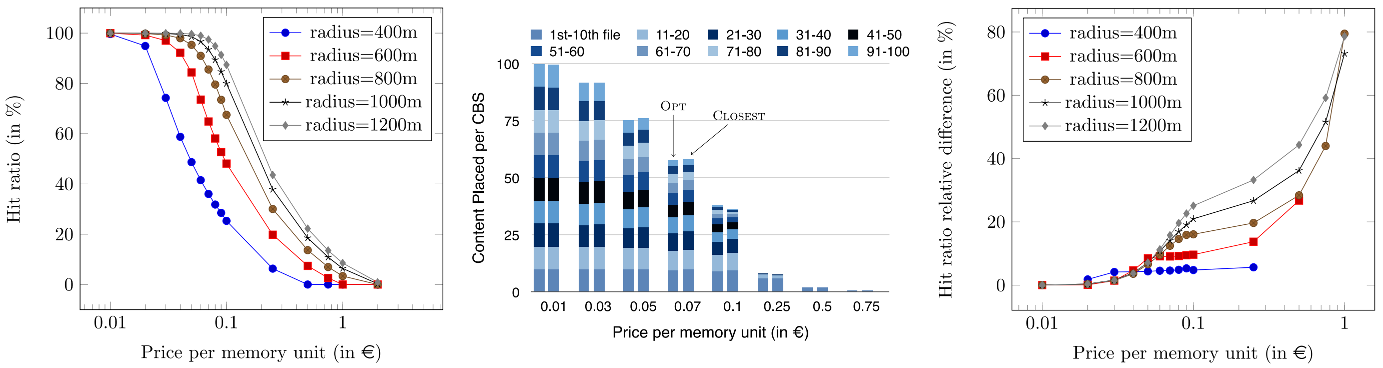

Fig. 5: Linear savings function. (a): Hit ratio in relation to price for different coverage radii with OPT association. (b): Cache lease and placement of popular files depending on cache unit price. Each left column represents the case of OPT association, each right column with CLOSEST association. (c): Relative difference between the hit ratio achieved by OPT to CLOSEST over cache unit price for different coverage radii.

to avoid edge effects. The coverage radius varies from 400m to 1200m. The MNO price per unit size cache memory at all CBSs varies (0.01€-2.00€). The user population is distributed uniformly over the network with a density of 30 users per km^2. The total simulated file catalog contains 100 objects. The content popularity follows the Zipf distribution with parameter 0.6 unless explicitly stated otherwise. The available cache size from the MNO is set to the catalog size so that only the pricing influences cache leasing decisions.

B. Implementation

All simulations have been performed using a native JAVA simulation environment. User association corresponding to the solution of the slave problem in Sections IV and V is entirely done by optimization algorithms that we developped in the lab, outlined in Section VI. The surrogate problem SUR-T in Section VI is solved using the state of the art mixed-integer problem solver IBM CPLEX 12.7.0 in combination with IBM ILOG CPLEX Optimization Studio. The experiments have been performed on a machine with a 2.40 GHz 16-core processor and 48 GB RAM.

C. Results for Linear Savings Function (case i)

At first, we present the simulation results for linear savings function. On average, the optimal solution was obtained after 2 Benders iterations. We emphasize case i.a) which performs OPT user association and compare it with case i.b) CLOSEST. Fig. 5(a) illustrates how the hit ratio in case i.a) depends on the price per cache unit for different coverage radii. For all radii, the hit ratio decreases with increasing prices. With lowest price (0.01€/Unit), the CP leases in each CBS the entire memory available, so the hit ratio is 100%. As the price increases, the CP leases less units, and the hit ratio is reduced. This happens more quickly in networks with smaller coverage areas because there are less users covered by each CBS and also less coverage overlap area. When the price reaches a high level (2€), the cost from leasing cache memory exceeds the benefit from cache hits and the hit ratio drops to 0% for all coverage radii. The differences between the curves in Fig. 5(a) diminish with higher radii where multi-coverage is already high enough.

The CP's leasing and placement decisions for networks with coverage radius 1000m for the policies OPT (i.a) and CLOSEST (i.b) are shown in Fig. 5(b). For each price, there are two columns: The lefthand-side column represents the i.a) case, the righthand-side column i.b). The height of each column is the average amount of cache units per CBS which are leased for the respective price. The subdivisions of each column represent the popularity of the files stored in the leased memory space: The bottom part are the ten most popular files, the second-to-bottom part are the files of popularity rank 11 to 20 and so on. For the lowest cache price (0.01€), all 100 available units are leased: For both assignment policies, the amount of leased cache memory decreases with increasing price. For all prices, Fig. 5(b) shows that the less popular files are represented more frequently with OPT than with CLOSEST, especially for prices $\geq$ 0.1€. There is more diversity of visible content with the OPT placement.

Fig. 5(c) directly compares OPT with CLOSEST. For all prices the hit ratio achieved by OPT (case i.a) is higher than the one achieved by the CLOSEST policy (case i.b). The relative hit ratio differences are higher when the coverage area of the CBSs is higher. For higher prices, the CLOSEST hit ratio is close to 0, therefore the relative differences can become very high.

While Fig. 5 shows the caching benefits for the CP, the costs it has to pay in return to the MNO (in case OPT) are depicted in Fig. 6(a). The CP costs equal the MNO income. This amount can be calculated by multiplying the number of leased cache units with the price per unit. The maximum of the curve can be clearly identified for each radius. *This is the operational point for the MNO when the latter aims for maximum income.* The maxima are higher for larger coverage areas, while the difference in income decreases with increasing radius. Furthermore, the higher the coverage radius, the higher the cache leasing price at which the maximum is achieved.

The relation between hit ratio and MNO income can be seen in Fig. 6(b): The x-axis displays the hit ratio achieved, the y-axis shows the income the MNO earns. Again, the income is higher in networks with larger coverage area. The maximum income for all simulated networks can be found for an achieved hit ratio between 80-90%. Conversely, if the

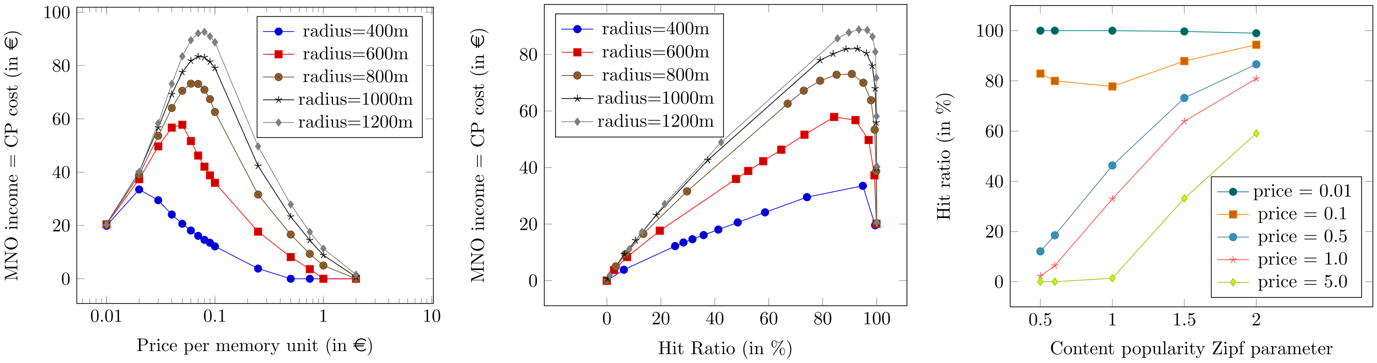

Fig. 6: Linear savings function. (a): Income of MNO in relation to price per cache unit for different coverage radii with OPT. (b): MNO income/CP investment over hit ratio for different coverage radii with OPT: Achieving the hit ratio on the x-axis results in the MNO income on y-axis. (c): Hit ratio in relation to Zipf parameter for different prices with OPT.

CP decides to invest a certain sum, it can maximally achieve the rightmost of the two corresponding hit ratio values under the condition that the MNO chooses the pricing strategy most favorable to the CP.

The experimental results presented until now are based on a Zipf parameter of 0.6. However, a varying Zipf parameter influences the results: Fig. 6(c) shows that the higher the price, the lower the hit ratio for any Zipf parameter in case i.a). This is due to the fact that lower price implies more leased units for the CP. For all prices (except the lowest one which achieves a hit ratio of near 100% throughout), the hit ratio increases with increasing Zipf parameter: With higher Zipf parameter, the population share requesting the most popular files becomes higher, thus caching popular files becomes more profitable. Also, for the same price, the leased cache memory is more effective with higher Zipf parameters. The lower the Zipf parameter, the more pronounced the differences in hit ratio between different cache prices since the benefits from overlapping coverage are bigger when content popularity is more even.

D. *Results for Log Savings Function per CBS (case ii)*

Here, we present the experimental results of case ii) in which the savings function is the sum of logarithms. The optimal solution was obtained after 8 Benders iterations on average. Fig. 7(a) shows the hit ratio for varying cache unit price both for the OPT (different coverage radii) and the CLOSEST policies. Due to the specific choice of the logarithmic savings function of case ii) the CLOSEST association gives identical cache leasing and content placement results for all coverage radii. For every coverage radius and every cache unit price, the OPT policy achieves a higher hit ratio than the CLOSEST policy. Furthermore, the higher the coverage radius in case ii.a), the higher the hit ratio. The hit ratio improvement can reach over 15 percentage points using OPT. With increasing prices, caching becomes less profitable and the hit ratio decreases.

The advantage to the hit ratio of the OPT policy (ii.a) can be explained by the optimal content placement shown in Fig. 7(b). Each pair of columns represents cache leasing and content placement for a certain unit price. Each left column represents the decisions according to OPT, each right column the decisions according to CLOSEST. The height is the average amount of cache units leased per CBS. The inner sections of the columns represent the content placement in all of the CBSs: The lowest section are the 10 most popular files, the second lowest the files ranked 11 to 20 and so on. It can be seen that particularly for low cache unit prices, the diversity of cached content is higher in case ii.a) than in case ii.b). For the lowest price (0.01€), CLOSEST provides only content from the more popular half of the catalog, while OPT places content from the tail of the catalog as well.

The main purpose of choosing the specific savings function in ii.a) is, however, the balancing of traffic load among the CBSs in order to avoid excess of resources by user overflow which will lead to service dissatisfaction. Fig. 7(c) shows that the additional load (from the increase in hit ratio using OPT, see Fig. 7(a)) is distributed to the less loaded CBSs. The two upper (solid) lines in the graph represent the maximum load of a CBS in relation to the overall covered population per CBS both in the cases ii.a) and ii.b). The two lower (dashed) lines are the minium loaded CBS. The maximum loaded CBSs in both ii.a) and ii.b) coincide as the figure shows. The minimum loaded CBS of ii.a) is higher than the ii.b), showing that excess users coming from the higher hit ratio are associated to the less loaded stations.

The three plots show that the OPT policy achieves an increase in hit ratio (good for both the CP and the MNO) while at the same time diversifying the cached content (good for the user) and avoiding an overload of CBSs (good for everybody).

E. *Policy Comparison*

In order to illustrate the true benefits of OPT over CLOSEST (with linear savings) as well as over other content placement policies, spatially inhomogeneous content popularity should be considered. We evaluate here a traffic scenario in which the network window is symmetrically divided such that on each side, the popularity of files follows a Zipf distribution with parameter 0.6, but the ten most popular files on one side are swapped with the second most popular decile on the other side to give locally differing popularity distributions.

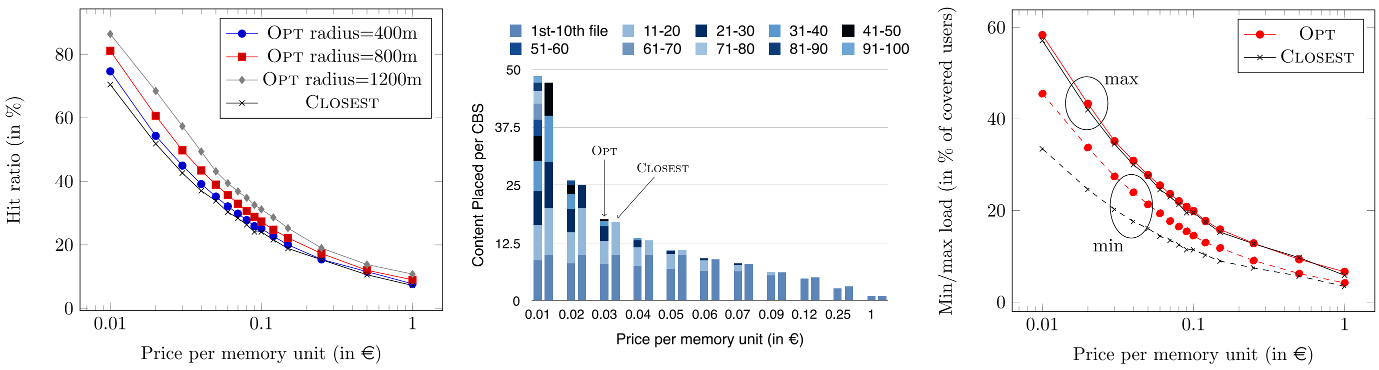

Fig. 7: Logarithmic savings function. (a): Hit ratio in relation to price for different coverage radii for cases OPT and CLOSEST. (b): Cache lease and placement of popular files depending on cache unit price. Each left column represents OPT, each right column is CLOSEST. (c) Cache lease and placement of popular files depending on cache unit price. Each left column represents OPT, each right column is CLOSEST.

We fix the cache sizes uniformly and do not consider the leasing mechanism, enabling comparisons with known content placement policies that do not take cache leasing into account.

Fig. 8(a) shows that for a large range of cache sizes, OPT provides a 50% relative gain in hit ratio over CLOSEST. Compared to other placement policies such as geographic caching (GEO, see [5]), caching of globally Most Popular Content (MPC) and uniformly random caching (RANDOM), the OPT policy provides considerable gains in each case. This result works strongly in favour of our suggestion that the MNO and CP should closely collaborate for user-association decisions.

F. Sensitivity to Traffic Estimation Errors

The actual content request distribution can diverge from the statistics on which the cache leasing and content placement decisions are based. We evaluate the effect of a diverging Zipf parameter of the popularity distribution on the hit ratio of CP decisions based on parameter 0.6 (linear savings).

From Fig. 8(b) it can be seen that if the Zipf parameter was underestimated for leasing and placement, the hit ratio will be higher than expected as well. Conversely, an overestimated Zipf parameter leads to a decrease in the obtained hit ratio.

G. Results for Throughput Maximization (Weighted Linear Savings Function, case iii)

Finally, we present the simulation results for throughput maximization. On average, the optimal solution was obtained after 2 Benders iterations. The CBSs have a coverage radius of 1000m and a power of $p = 1W$. Every station has two coverage zones, one in the inner half of the coverage radius, one in the outer half. The channel $h_{m,s}$ is calculated assuming a path loss exponent of 4. For the calculation of the SINR (see (10)), the distance of any user in the zone with strong signal ($\leq$ 500m distance to the CBS) is assumed to be 500m, in the weak signal zone it is assumed to be 1000m. The noise is $\sigma^2 = 10^{-12}$W.

For each network, three different frequency reuse scenarios are simulated. In case of full frequency reuse ("1-freq"), each user uniformly gets assigned the bandwidth of B = 1MHz. All neighboring CBSs induce interference to each other. In a second scenario "2-freq", every station can operate on half of the total spectrum (random coin toss). To potentially serve the same number of users as in 1-freq, every associated user is served on half of the previous bandwidth, i.e. B = 0.5MHz. But in this case, the benefit is that only neighboring stations with the same half of the spectrum interfere. For "3-freq", the same principle is applied for a division of the spectrum into 3 orthogonal parts.

For this run of experiments, the savings function is the linear weighted sum of traffic with the weights chosen as user throughputs (9). The valuation of the savings is set to c = 1€/MBits which implies that any achieved throughput of more than 1MBits per invested Euro brings the CP into profit. Fig. 8(c) shows the total throughput per invested Euro ("caching efficiency") over cache leasing price (between 0.1€ and 1.6€) for the user association policies OPT and CLOSEST, each for the different frequency reuse scenarios. Observe that OPT yields a significantly higher caching efficiency than CLOSEST throughout. The relative difference between the two user association policies is greatest in the 2-freq scenario whereas the absolute system performance is maximised with 1-freq. This is due to the fact that for 2-freq, center users receive less bandwidth than in 1-freq even though interference is low. Interference reduction and system performance of 2-freq and 3-freq can be improved, however, by assigning the parts of the spectrum to CBSs in a more intelligent way than uniformly randomly. Depending on frequency reuse, the MNO can offer profit through caching ($\geq$ 1MBits/€) to the CP even when the cache-price is set high.

VIII. CONCLUSIONS

In this work, we propose a business model in which an MNO leases edge caches to a CP. The CP's objective is to maximize its savings through offloading of traffic from its data centers to the wireless caches while limiting the cache leasing costs. The optimality of the CP decisions depends on the MNO's user association policy as well as its pricing strategy.

The problem is modelled as a NLMIP taking the perspective of the CP. Network topology, association policy, pricing as

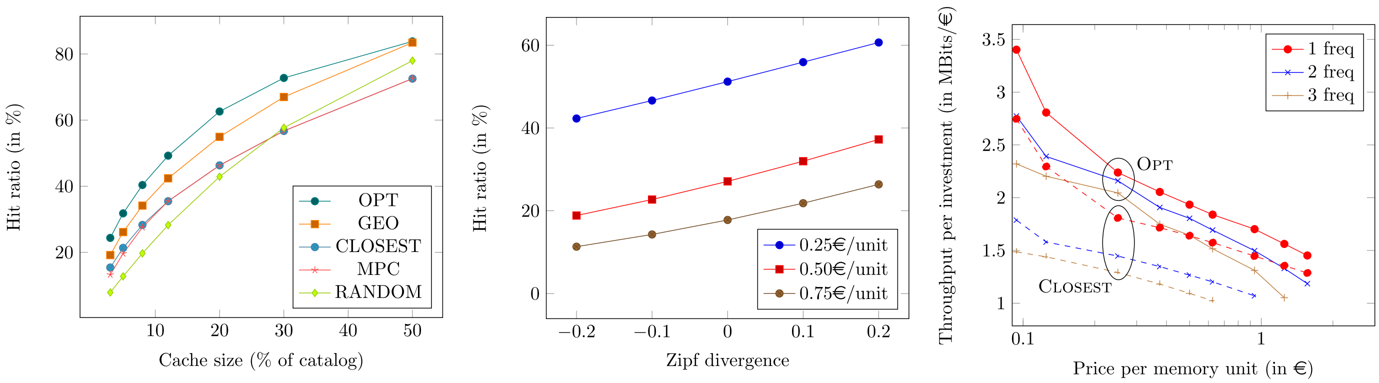

Fig. 8: (a) Hit ratio for different content placement policies with fixed cache sizes and spatially inhomogeneous traffic with linear savings function. (b) Hit ratio when the Zipf parameter of the actual content popularity diverges from the Zipf parameter on which the leasing and placement decisions are based. (c) Throughput optimization by weighted savings function: Throughput per cache investment.

well as CP savings function are general. Radio conditions and wireless node resource constraints can be included. An optimal solution to the general problem is found by applying Generalized Benders decomposition. The solution method converges to the global optimum and allows for each of the players to take separate actions iteratively.

Extensive experiments for random network topologies allow to compare the optimal CP decisions for different MNO association policies, cache prices, as well as CP savings functions. In all versions of the problem, we have identified a unique price that maximizes the MNO revenue. It depends on how much the CP valorizes traffic offloading achieved by the edge caches. This information is included in the CP's choice of the savings function. Another main conclusion is that MNO association policies which adhere to CP actions and exploit multi-coverage opportunities achieve higher offloading benefits for a given monetary investment. All these results suggest that the CP and MNO can jointly develop cooperative business models related to caching, that lead to considerable economic as well as operational benefits for both parties.

In future research, the model can be extended to include several CPs that compete for one or multiple MNO storage resources. Each MNO can either aim for the social optimum e.g. sum hit-rate maximization, or can target its own profit maximization. In each case, different pricing strategies should be considered.

IX. Appendix

Proof of Proposition 3. Note first that $f_q(y)$ is invertible with

$$f_q^{-1}(y) = (w_{\tilde{q}}/w_q)y + \beta_q \tag{17}$$

and f_q^{-1} is non-decreasing. The value $f_q^{-1}(y)$ represents the reference water level ν if y is the non-zero water volume in bucket q. Let $\hat{y}_q := w_q y_q$ for all $q \in \mathcal{Q}$. Then, the following problem is equivalent to (14):

$$\hat{\mathbf{y}}^* = \underset{0 \leq \hat{\mathbf{y}} \leq \hat{\mathbf{N}}}{\arg\max} \quad \hat{g}(\hat{\mathbf{y}})$$

with $\hat{g}(\hat{\mathbf{y}}) := \mathrm{U}\left(\sum_{q\in\mathcal{Q}} \hat{y}_q\right) - \sum_{q\in\mathcal{Q}}\left[\frac{\varrho}{2w_q^2}\hat{y}_q^2 - \frac{a_q}{w_q}\hat{y}_q\right]$ and $\hat{\mathbf{N}} := (w_q N_q), q \in \mathcal{Q}$. Note that $N_q > 0$ for all q and thus the upper and the lower bound of y_q cannot be fulfilled with equality simultaneously. The KKT conditions can thus be written in a compact way: There exist $\gamma_q \in \mathbb{R}$ for all $q \in \mathcal{Q}$ such that

$$\frac{\partial}{\partial \hat{y}_q}\hat{g}(\hat{\mathbf{y}}^*) - \gamma_q = 0 \tag{18}$$

for all $q \in \mathcal{Q}$ where

$$\gamma_q = \begin{cases} \geq 0, & \text{if } \hat{y}_q^* = w_q N_q \iff y_q^* = N_q \\ \leq 0, & \text{if } \hat{y}_q^* = 0 \iff y_q^* = 0 \\ = 0, & \text{otherwise} \end{cases} \tag{19}$$

and

$$\frac{\partial}{\partial \hat{y}_q}\hat{g}(\hat{\mathbf{y}}) = \mathrm{U}'\left(\sum_{q\in\mathcal{Q}} \hat{y}_q\right) - \frac{\varrho}{w_q^2}\hat{y}_q + \frac{a_q}{w_q}.$$

For any $p, q \in \mathcal{Q}$, (18) implies that the following equation holds:

$$-\frac{\varrho}{w_q^2}\hat{y}_q^* + \frac{a_q}{w_q} - \gamma_q = -\frac{\varrho}{w_p^2}\hat{y}_p^* + \frac{a_p}{w_p} - \gamma_p$$

since the term $\mathrm{U}'\left(\sum_{q\in\mathcal{Q}} \hat{y}_q^*\right)$ is contained in both derivatives. A calculation resubstituting $y_p^* = \hat{y}_p^*/w_p$ and $y_q^* = \hat{y}_q^*/w_q$ results in $y_q^* = f_q \circ f_p^{-1}(y_p^*) + (w_q/\varrho)(\gamma_p - \gamma_q)$. Now, we will find a $p \in \mathcal{Q}$ such that $\nu^* := f_p^{-1}(y_p^*)$ yields (15). Then,

$$y_q^* = f_q(\nu^*) + \frac{w_q}{\varrho}(\gamma_p - \gamma_q) \tag{20}$$

Case 1: $\exists p \in \mathcal{Q}$ with $0 < y_p^* < N_p$. Define $\nu^* := f_p^{-1}(y_p^*)$. Note that, in this case, $\gamma_p = 0$ such that for all $q \in \mathcal{Q}$:

$$y_q^* = f_q(\nu^*) - \frac{w_q}{\varrho}\gamma_q. \tag{21}$$

With this equation, we deduce (15) for any $q \in \mathcal{Q}$: Firstly, if $0 < y_q^* \leq N_q + \beta_q$, then $\gamma_q = 0$ follows from (19) and $y_q^* = f_q(\nu)$ holds due to (21). Secondly, if $y_q^* = 0$, then $\gamma_q \leq 0$ by (19). Then (21) implies that $f_q(\nu^*) \leq y_q^* = 0$ which by definition of f_q is true if and only if $\nu^* \leq \beta_q$. Lastly, if $y_q^* = N_q$, then $\gamma_q \geq 0$ by (19) and with (21): $f_q(\nu^*) \geq y_q^* = N_q$. By definition, this holds if and only if $\nu^* \geq (w_{\tilde{q}}/w_q)N_q + \beta_q$.

Case 2: $\nexists q \in \mathcal{Q}$ with $0 < y_q^* < N_q$, but $\exists q \in \mathcal{Q}$ with $y_q^* = 0$. Choose p such that $f_p^{-1}(0) = \beta_p =: \nu^*$ is minimal among all such lower-bound region-files, i.e. $p = \arg\min_{q\in\mathcal{Q},y_q^*=0}\beta_q$. It suffices to show that $f_q(\nu^*)\leq 0$ if $y_q^*=0$ and $f_q(\nu^*)\geq N_q$ if $y_q^*=N_q$. Let firstly $q\in\mathcal{Q}$ with $y_q^*=0$. Then

$$f_q(\nu) = f_q(\beta_p) = (w_q/w_{\tilde{q}})(\beta_p-\beta_q)\leq 0$$

by definition of p. Secondly, let $q\in\mathcal{Q}$ with $y_q^*=N_q$. Note that (19) implies that $\gamma_p\leq 0$ and $\gamma_q\geq 0$. Thus, (20) implies that $f_q(\nu^*)\geq N_q$ which holds if and only if $\nu^*\geq(w_{\tilde{q}}/w_q)N_q+\beta_q$.

Case 3: $y_q^* = N_q$ for all $q\in\mathcal{Q}$. Choose $p := \arg\max_{q\in\mathcal{Q}} f_q^{-1}(N_q)$ and $\nu := f_p^{-1}(N_p) =$. Then, since f_q is non-decreasing for $q\in\mathcal{Q}$,

$$f_q(\nu^*)\geq f_q(y_q^*) = f_q(N_q) = (w_{\tilde{q}}/w_q)N_q+\beta_q.$$

It remains to be shown that $\nu^*\geq 0$ in all cases. Note that $\nu = f_p^{-1}(y_p^*)$ for some $p\in\mathcal{Q}$, f_p^{-1} is non-decreasing (see (17)) and $y_p^*\geq 0$. Then

$$\begin{aligned}\nu = f_p^{-1}(y_p^*) &\geq f_p^{-1}(0) = (w_{\tilde{q}}/w_p)\beta_p\\ &= w_{\tilde{q}}\frac{a_{\tilde{q}}/w_{\tilde{q}} - a_p/w_p}{\varrho}\geq 0.\end{aligned}$$

The last step is true due to the choice of $\tilde{q} = \arg\max_{q\in\mathcal{Q}} a_q/w_q$. □

References

[1] J. Krolikowski, A. Giovanidis, and M. Di Renzo. Fair distributed user-traffic association in cache equipped cellular networks. In *WiOpt-CCDWN*, pages 1–6, 2017.

[2] J. Krolikowski, A. Giovanidis, and M. Di Renzo. Optimal cache leasing from a mobile network operator to a content provider. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, USA, April 2018.

[3] Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 white paper. White Paper, 2 2016.

[4] K. Shanmugam, N. Golrezaei, A.G. Dimakis, A.F. Molisch, and G. Caire. Femtocaching: Wireless content delivery through distributed caching helpers. *Information Theory, IEEE Transactions on*, 59(12):8402–8413, Dec 2013.

[5] B. Błaszczyszyn and A. Giovanidis. Optimal geographic caching in cellular networks. In *Communications (ICC), 2015 IEEE International Conference on*, pages 3358–3363. IEEE, 2015.

[6] S. Elayoubi and J. Roberts. Performance and cost effectiveness of caching in mobile access networks. In *Proceedings of the 2nd ACM Conference on Information-Centric Networking*, ACM-ICN '15, pages 79–88, New York, NY, USA, 2015. ACM.

[7] E. Baştuğ, M. Bennis, and M. Debbah. Cache-enabled small cell networks: Modeling and tradeoffs. In *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, pages 649–653, Aug 2014.

[8] K. Poularakis, G. Iosifidis, I. Pefkianakis, L. Tassiulas, and M. May. Mobile data offloading through caching in residential 802.11 wireless networks. *IEEE Transactions on Network and Service Management*, 13(1):71–84, 2016.

[9] G. Paschos, E. Baştuğ, I. Land, G. Caire, and M. Debbah. Wireless caching: Technical misconceptions and business barriers. *IEEE Communications Magazine*, 54(8):16–22, 2016.

[10] A. Araldo, G. Dan, and D. Rossi. Stochastic dynamic cache partitioning for encrypted content delivery. In *Internet Teletraffic Congress (ITC) 2016*, 2016.

[11] K. Poularakis, G. Iosifidis, and L. Tassiulas. Approximation algorithms for mobile data caching in small cell networks. *IEEE Transactions on Communications*, 62(10):3665–3677, 2014.

[12] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas. Placing dynamic content in caches with small population. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, April 2016.

[13] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini. Unravelling the impact of temporal and geographical locality in content caching systems. *IEEE Transactions on Multimedia*, 17(10):1839–1854, Oct 2015.

[14] G. Neglia, D. Carra, and P. Michiardi. Cache policies for linear utility maximization. In *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*, pages 1–9, 2017.

[15] A. Giovanidis and A. Avranas. Spatial multi-LRU caching for wireless networks with coverage overlaps. *SIGMETRICS Perform. Eval. Rev.*, 44(1):403–405, June 2016.

[16] M. Dehghan, B. Jiang, A. Seetharam, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman. On the complexity of optimal request routing and content caching in heterogeneous cache networks. *IEEE/ACM Transactions on Networking*, 25(3):1635–1648, June 2017.

[17] M. Dehghan, W. Chu, P. Nain, and D. Towsley. Sharing LRU cache resources among content providers: A utility-based approach. *arXiv:1702.01823*, 2017.

[18] N. Gast and B. Van Houdt. TTL approximations of the cache replacement algorithms LRU(m) and h-LRU. *Performance Evaluation*, 117:33 – 57, 2017.

[19] K.P. Naveen, L. Massoulie, E. Baccelli, A. Carneiro Viana, and D. Towsley. On the interaction between content caching and request assignment in cellular cache networks. In *ACM*, AllThingsCellular '15, pages 37–42. ACM, 2015.

[20] M. Dehghan, A. Seetharam, Bo Jiang, Ting He, Th. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman. On the complexity of optimal routing and content caching in heterogeneous networks. In *IEEE INFOCOM*, 2015.

[21] T. Liu, J. Li, F. Shu, M. Tao, W. Chen, and Z. Han. Design of contract-based trading mechanism for a small-cell caching system. *IEEE Transactions on Wireless Communications*, 16(10):6602–6617, Oct 2017.

[22] K. Avrachenkov, J. Goseling, and B. Serbetci. A low-complexity approach to distributed cooperative caching with geographic constraints. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):27:1–27:25, June 2017.

[23] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas. Exploiting caching and multicast for 5g wireless networks. *IEEE Transactions on Wireless Communications*, 15(4):2995–3007, April 2016.

[24] A. Tuholukova, G. Neglia, and T. Spyropoulos. Optimal cache allocation for femto helpers with joint transmission capabilities. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2017.

[25] J. Liu, B. Bai, J. Zhang, and K. B. Letaief. Cache placement in fog-rans: From centralized to distributed algorithms. *IEEE Transactions on Wireless Communications*, 16(11):7039–7051, Nov 2017.

[26] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris. Cooperative caching and transmission design in cluster-centric small cell networks. *IEEE Transactions on Wireless Communications*, 16(5):3401–3415, May 2017.

[27] G. Alfano, M. Garetto, and E. Leonardi. Content-centric wireless networks with limited buffers: When mobility hurts. *IEEE/ACM Transactions on Networking*, 24(1):299–311, Feb 2016.

[28] F. De Pellegrini, A. Massaro, L. Goratti, and R. El-Azouzi. A pricing scheme for content caching in 5g mobile edge clouds. In *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 193–198, Oct 2016.

[29] M. Mangili, F. Martignon, S. Paris, and A. Capone. Bandwidth and cache leasing in wireless information-centric networks: A game-theoretic study. *IEEE Transactions on Vehicular Technology*, 66(1):679–695, Jan 2017.

[30] V. G. Douros, S. E. Elayoubi, E. Altman, and Y. Hayel. Caching games between content providers and internet service providers. *Performance Evaluation*, 2017.

[31] D. Bertsimas and R. Weismantel. *Optimization over integers*. Athena Scientific, 2005.

[32] W. Chu, M. Dehghan, J. C. S. Lui, D. Towsley, and Z.-L. Zhang. Joint Cache Resource Allocation and Request Routing for In-network Caching Services. *arXiv:1710.11376*, October 2017.

[33] T. Bektaş, J. Cordeau, E. Erkut, and G. Laporte. Exact algorithms for the joint object placement and request routing problem in content distribution networks. *Computers & Operations Research*, 35(12):3860–3884, 2008.

[34] L. P. Qian, Y. J. A. Zhang, Y. Wu, and J. Chen. Joint base station association and power control via benders' decomposition. *IEEE Transactions on Wireless Communications*, 12(4):1651–1665, April 2013.

[35] H. Lin and H. Üster. Exact and heuristic algorithms for data-gathering cluster-based wireless sensor network design problem. *IEEE/ACM Transactions on Networking*, 22(3):903–916, June 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2018.2844986, IEEE Journal on Selected Areas in Communications

[36] A. Elwalid, D. Mitra, and Q. Wang. Cooperative data-optical internetworking: Distributed multi-layer optimization. In *INFOCOM*. IEEE, 2006.

[37] K.-W. Lim, S. Secci, L.Tabourier, and B. Tebbani. Characterizing and predicting mobile application usage. *Computer Communications*, 95:82 – 94, 2016. Mobile Traffic Analytics.

[38] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[39] A. M. Geoffrion. Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10:237–260, 1972.

[40] F. Kelly. Charging and rate control for elastic traffic. *Transactions on Emerging Telecommunications Technologies*, 8(1):33–37, 1997.

[41] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.*, 8(5):556–567, October 2000.

[42] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

Jonatan Krolikowski is a PhD student at the Laboratoire des Signaux & Systèmes, CentraleSupélec, Université Paris-Saclay, France, since October 2015. The subject of his PhD thesis is ”Optimal Content Management and Dimensioning in Wireless Networks”. Prior, he worked as a developer of optimization algorithms for the logistics management and consulting company 4flow (Berlin). He has obtained the degrees MSc and BSc in Mathematics from TU Berlin in 2014 and 2013, respectively. In 2007, he obtained a Magister (Masters equivalent) in Media and Communication Studies from FU Berlin. His main interest is in both the theory and practice of efficient algorithms for challenging network optimization problems.

Anastasios Giovanidis (M) is a permanent researcher of the French National Center for Scientific Research (CNRS, CR1), affiliated since 2016 with the Computer Science Department of the Sorbonne University (LIP6 laboratory). Prior to that he was a member of the Telecom ParisTech CNRS-LTCI laboratory. He has been a postdoctoral fellow, first with the Zuse Institute Berlin, Germany and later with INRIA, Paris, France. He obtained in 2010 the Dr.-Ing. degree in Mobile Communications from the Technical University of Berlin, Germany and in 2005 the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece. He has served as the general co-Chair of WIOPT 2017 and co-organiser of the CCDWN 2018 caching-related workshop. His research interests are performance evaluation and optimisation, currently applied to edge-caching problems and cooperative wireless networks.

Marco Di Renzo (SM) was born in L’Aquila, Italy, in 1978. He received the Laurea (cum laude) and Ph.D. degrees in electrical engineering from the University of L’Aquila, Italy, in 2003 and 2007, respectively, and the Habilitation à Diriger des Recherches (Doctor of Science) degree from University Paris-Sud, France, in 2013. Since 2010, he has been a Chargé de Recherche CNRS (CNRS Associate Professor) in the Laboratory of Signals and Systems (L2S) of Paris-Saclay University - CNRS, Centrale-Supélec, Univ Paris Sud, Paris, France. He serves as the Associate Editor-in-Chief of IEEE Communications Letters, and as an Editor of IEEE Transactions on Communications, and IEEE Transactions on Wireless Communications. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society and IEEE Communications Society, and a Senior Member of the IEEE. He is a recipient of several awards, including the 2013 IEEE-COMSOC Best Young Researcher Award for Europe, Middle East and Africa, the 2013 NoE-NEWCOM# Best Paper Award, the 2014-2015 Royal Academy of Engineering Distinguished Visiting Fellowship, the 2015 IEEE Jack Neubauer Memorial Best System Paper Award, the 2015-2018 CNRS Award for Excellence in Research and Ph.D. Supervision, the 2016 MSCA Global Fellowship (declined), the 2017 SEE-IEEE Alain Glavieux Award, the 2018 IEEE ICNC Silver Contribution Award, and 6 Best Paper Awards at IEEE conferences (2012 and 2014 IEEE CAMAD, 2013 IEEE VTC-Fall, 2014 IEEE ATC, 2015 IEEE ComManTel, 2017 IEEE SigTelCom).