Joint User Association and Caching in Wireless Heterogeneous Networks with Backhaul

Yuezhou Liu[1], Alireza Alizadeh[2], Mai Vu[2], and Edmund Yeh[1]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA
[2]Department of Electrical and Computer Engineering, Tufts University, Medford, MA, USA

***Abstract*—We consider a mobile network consisting of both the wireless access network and the backhaul network. All base stations in the access network and gateways in the backhaul network are equipped with caches, so that routing costs for serving content requests can be reduced by caching the requested content items closer to the users. In this case, user association in the wireless access network must be aware of both the quality of wireless channels and the content caching strategy. In this paper, we propose a framework that jointly optimizes wireless user association and content caching in both access and backhaul networks. The resulting problem is NP-hard. We propose a polynomial-time algorithm based on convex approximation and pipage rounding that produces a solution within a constant factor of $1 - 1/e$ from the optimal. Simulation results show that the proposed joint algorithm outperforms schemes that combine cache-independent user association methods with traditional caching strategies (e.g. LRU) in terms of minimizing the aggregate routing cost and backhaul traffic while achieving a high data sum rate in the access network.**

I. Introduction

In recent years, mobile traffic has experienced explosive growth due to the proliferation of mobile devices and demands for high-volume media content. To keep pace with this growth, mobile caching is proposed as a promising solution. By storing content items closer to mobile users (UEs), requests are served at the edge, which helps to improve response time, reduce backhaul traffic, and alleviate server congestion. At the same time, the heterogenous network (HetNet) emerges as an effective way to increase the capacity and coverage of wireless access networks. In HetNets, together with traditional macro base stations (MBSs) working at sub-6 GHz band, dense deployment of short-range small base stations (SBSs) operating at mmWave frequency bands occurs at distances much closer to UEs, thus enabling higher data rate access.

In mobile caching, one typically equips BSs with storage devices. For example, proactive content caching at SBSs is used to overcome capacity-limited backhaul links and minimize the delivery delay [1]. Caching at gateways can also reduce the amount of user traffic that must go to the internet through backhaul links, thus further increases the effective bandwidth [2]. At the access network, because of the dense deployment of SBSs, each UE is likely to be within the range of multiple base stations (BSs) and can associate (connect) with any of them to fetch content items, necessitating mechanisms for optimal user association. The traditional method of associating UEs with the BS having highest signal to interference plus noise ratio (SINR), *Max-SINR*, can lead to unbalanced BS loads. Load balancing user association has been studied for LTE and 5G networks for maximizing transmission rates in [3]–[5].

With the availability of mobile caching, cache-aware user association becomes necessary to take content availability into consideration. Several works have studied joint user association and caching at BSs. These include optimizing the fractions of content items served by different BSs and caching policy to minimize the average serving time [6]. The problem is studied for maximizing user data rates and backhaul savings in [7]. Joint user association and caching at unmanned aerial vehicles (UAVs) that minimizes content acquisition delay is studied in [8], where UAVs play a similar role as BSs.

None of these existing works, however, consider caching in backhaul nodes such as gateways. Furthermore, though some papers formulate a joint problem of association and caching, the problem is subsequently separated into two sub-problems (a caching problem with fixed association and an association problem with fixed caching) for tractability. This separation is usually sub-optimal and leads to performance loss.

In this paper, we study the joint user association and caching in mobile networks, which includes caching at not only BSs but also gateways. To our best knowledge, this is the first work to consider such a joint optimization. We follow [9] that studies caching and [10] that studies joint optimization of routing and caching to minimize the aggregate expected routing cost, while extend beyond them by including user association in the access network. We design an approximation algorithm which has polynomial time-complexity and produces a solution within a constant factor from the optimal. By simulations, we show that the proposed joint user association and caching algorithm achieves not only the lowest aggregate expected routing cost but also a high data rate compared to existing solutions.

II. System Model and Problem Formulation

We study a heterogeneous access network including cache-enabled MBSs and SBSs, and a backhaul network containing a number of cache-enabled gateways as shown in Fig. 1. Let $\mathcal{K}$ be the set of UEs, $\mathcal{J} = \mathcal{J}_M \cup \mathcal{J}_S$ be the set of BSs, where $\mathcal{J}_M$ is the set of MBSs operating at sub-6 GHz frequency band and $\mathcal{J}_S$ is the set of SBSs working at mmWave band, $\mathcal{N}$ represents the set of gateways in backhaul, and $\mathcal{S}$ is the set of servers located on the Internet. We represent the whole network by a directed and bidirectional graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{K} \cup \mathcal{J} \cup \mathcal{N} \cup \mathcal{S}$,

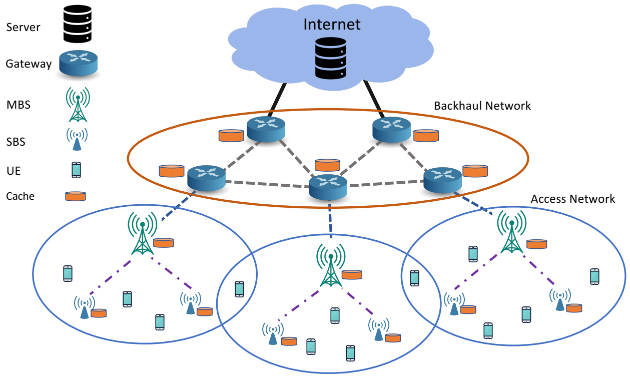

Fig. 1. A heterogeneous mobile network including cache-enabled gateways (GWs), LTE macro BSs (MBSs), 5G small BSs (SBSs), and users (UEs).

and $\mathcal{E}$ is set of bidirectional links connecting the UEs, BSs, gateways, and servers.

UEs are located at random locations in the wireless access network. Each UE is associated with one of the BSs over wireless links. The SBSs are connected to MBSs via fixed fronthaul links while MBSs are connected to gateways via fixed backhaul links. Some gateways are connected to the remote servers by multi-hop wired links. There is a content catalog $\mathcal{C}$ of equally-sized content items to be delivered in the network.[1] Each UE in the network generates content requests for content items at rates depending on their content preferences and their activity levels.

A. User Association and Caching

1) User Association: Each UE is associated with one of the BSs to fetch content items. Let binary association variables

$$\beta_{kj} \in \{0,1\}, \quad k \in \mathcal{K}, j \in \mathcal{J}, \tag{1}$$

be the indicator variables indicating whether UE k is associated with BS j. We call matrix $\beta = [\beta_{kj}]_{k\in\mathcal{K}, j\in\mathcal{J}} \in \{0,1\}^{|\mathcal{K}|\times|\mathcal{J}|}$ the *user association strategy* in the access network. We consider unique user association such that each UE is connected to one and only one BS at a time.[2] Furthermore, each BS j has a load balancing constraint D_j specified by the maximum number of UEs it can serve simultaneously, where $D_j \le M_j$, and M_j is the number of transmit antennas of BS j. Thus, the binary association variables must satisfy

$$\sum_{j\in\mathcal{J}} \beta_{kj} = 1 \quad \text{for all } k \in \mathcal{K}, \tag{2}$$

$$\sum_{k\in\mathcal{K}} \beta_{kj} \le D_j \quad \text{for all } j \in \mathcal{J}, \tag{3}$$

where the first set of constraints represents the unique association and the second set indicates the load balancing constraints.

We note that the association variables are somehow related to the routing variables in [10], as we also select the paths of the user requests when we determine the user association with BSs. However, the association variables have one more constraint, the load constraint, which necessitates a different method (see III-B2).

2) Caching Strategy: We consider cache-enabled MBSs, SBSs, and gateways. Each node $v \in \mathcal{J} \cup \mathcal{N}$ is equipped with a cache that can store $c_v \in \mathbb{N}_+$ content items. Let

$$x_{vi} \in \{0,1\}, \quad \text{for all } v \in \mathcal{J} \cup \mathcal{N}, i \in \mathcal{C}, \tag{4}$$

be the caching variables indicating whether node v stores content i in its cache. We call matrix $X = [x_{vi}]_{v\in\mathcal{J}\cup\mathcal{N}, i\in\mathcal{C}} \in \{0,1\}^{(|\mathcal{J}|+|\mathcal{N}|)\times|\mathcal{C}|}$ the *caching strategy* of the network satisfying the capacity constraint:

$$\sum_{i\in\mathcal{C}} x_{vi} = c_v, \quad v \in \mathcal{J} \cup \mathcal{N}. \tag{5}$$

B. Content Requests

A request (i,k) is determined by the content item $i \in \mathcal{C}$ requested and the requesting UE $k \in \mathcal{K}$. Let $R \subseteq \mathcal{C} \times \mathcal{K}$ be the set of all requests. Requests of different types arrive according to independent Poisson processes with rates $\lambda_{(i,k)} > 0$.

For each content item $i \in \mathcal{C}$, there is a set of *designated servers* $\mathcal{S}_i \subseteq \mathcal{S}$ that store i permanently. A request $(i,k) \in \mathcal{R}$ is routed over a path towards a designated server of i. The path is determined as follows: Given requested content item i and associated BS j, we assume that the path between BS j and the server of content item i is fixed and pre-established by an external routing algorithm, given by $p^{ij} = (p^{ij}_0, p^{ij}_1, \ldots, p^{ij}_M)$, where $p^{ij}_0 = j$, $p^{ij}_m \in \mathcal{V}\backslash\mathcal{K}$ and $(p^{ij}_{m-1}, p^{ij}_m) \in \mathcal{E}$ for $m = 1, \ldots, M$. Then the complete path of request (i,k) is given by $(k, p^{ij}_0, p^{ij}_1, \ldots, p^{ij}_M)$. We assume that every path is well-routed: (*1*) the path contains no cycles; (*2*) the last node of the path is the designated server, i.e., $p^{ij}_M \in \mathcal{S}_i$; and (*3*) no other node in the path is the designated server, i.e., $p^{ij}_m \notin \mathcal{S}_i$ for $m = 1, \ldots, M-1$.

C. Requests Routing Cost

1) Link Cost: In this paper, we assume that UEs and BSs are connected via wireless links, while the links among BSs, gateways and servers are wired links. We associate each link $(u,v) \in \mathcal{E}$ (either wired or wireless link) with a cost $w_{uv} \ge 0$ indicating the routing cost (e.g., delay or financial expense) incurred when transferring a content item across edge (u,v).

In particular, we assume the wireless link cost between BS $j \in \mathcal{J}$ and UE $k \in \mathcal{K}$ is given as

$$w_{jk} = f(c_{jk})$$

where c_{jk} is the capacity of the wireless link between BS j and UE k, and the wireless cost $f(\cdot)$ is a decreasing function in link capacity, so that the cost of transferring a content item over a link with higher capacity is lower. For example, the following wireless cost function describes the transmission plus propagation time delay for sending a content item over wireless link (j,k):

$$w_{jk} = \frac{d}{c_{jk}} + \tau_{jk} \tag{6}$$

[1] We will address unequally-sized content items in future works. This extension involves partition items into equally-sized “chunks” and UEs associate with multiple BSs at a time.

[2] Techniques involving multiple BSs connections such as coordinated multipoint (CoMP) are possible future directions.

where d is the size of content items, and τ_{kj} is the (constant) propagation delay over (j,k).

The link capacity c_{jk} depends on the wireless channel characteristics, interference, and also receiver processing techniques. With linear processing, the base station transmits a signal vector precoded for all its associated UEs. The precoding matrix at the base station is a function of the channel estimate. We consider the following wireless link capacity (instantaneous rate) from BS j to an associated UE k:

$$c_{jk} = W \log_2 \left(1 + \frac{P_{jk} \mathbf{w}_k^* \mathbf{H}_{jk} \mathbf{f}_{jk} \mathbf{f}_{jk}^* \mathbf{H}_{jk}^* \mathbf{w}_k}{I_{jk} + N_0 W \mathbf{w}_k^* \mathbf{w}_k}\right), \qquad (7)$$

where $\mathbf{H}_{jk}$ represents the downlink channel from BS j to UE k, $P_{jk} = P_j / D_j$ is the transmit power from BS j dedicated to UE k, P_j is the total transmit power of BS j, $\mathbf{f}_{jk} \in \mathbb{C}^{M_j \times 1}$ is the linear precoder (transmit beamforming vector) for each UE k associated with BS j, and $\mathbf{w}_k \in \mathbb{C}^{N_k \times 1}$ is the linear combiner (receive beamforming vector) of UE k. N_0 represents the noise power spectral density, W is the system bandwidth, and I_{jk} is the interference power at UE k when connected to BS j, as defined in [5],

$$I_{jk} = \frac{P_j - P_{jk}}{M_j} \mathbf{w}_k^* \mathbf{H}_{jk} \mathbf{H}_{jk}^* \mathbf{w}_k + \sum_{i \in \mathcal{J}, i \neq j} \frac{P_i}{M_i} \mathbf{w}_k^* \mathbf{H}_{ik} \mathbf{H}_{ik}^* \mathbf{w}_k,$$

where the first and second terms in I_{jk} represent the intra-cell and inter-cell interference, respectively. When computing the interference, we assume that BSs equally allocate the transmitting power to the antennas, i.e., each BS j has an precoding matrix $F_j = \frac{1}{\sqrt{M_j}} \mathbf{I}_{M_j}$. The transmitting power of BS j is given by $\frac{P_j}{M_j} \mathrm{Tr}(F_j F_j^*) = P_j$.

2) Routing Cost: A request is routed over its path until reaching a node (either the designated server or an intermediate cache node) that stores the requested content item. The content item is then sent back in a response message over the reverse path to the UE. Compared with the size of a response message that carries the content item, the size of a request message is relatively small. Thus, we assume that the request forwarding costs are negligible. Then the routing cost of a request $(i,k) \in \mathcal{R}$ is

$$C_{(i,k)}(\beta, X) = \sum_{j \in \mathcal{J}} \beta_{kj} \cdot \left[w_{jk} + \sum_{m=1}^{|p^{ij}|-1} w_{p^{ij}_{m+1} p^{ij}_m} \prod_{m'=1}^{m} (1 - x_{p^{ij}_{m'}, i}) \right]. \qquad (8)$$

D. Problem Formulation

We aim to determine the user association strategy and caching strategy that minimize the *aggregate expected routing cost*, defined as

$$C(\beta, X) = \sum_{(i,k) \in \mathcal{R}} \lambda_{(i,k)} C_{(i,k)}(\beta, X). \qquad (9)$$

Let C^0 be the constant:

$$C^0 = \sum_{(i,k) \in \mathcal{R}} \lambda_{(i,k)} \sum_{j \in \mathcal{J}} \left[w_{jk} + \sum_{m=1}^{|p^{ij}|-1} w_{p^{ij}_{m+1} p^{ij}_m} \right], \qquad (10)$$

which is an upper bound to $C(\beta, X)$ given in (9), for any feasible β and X. Then, minimizing the cost (9) is equivalent to maximizing the *association and caching gain*, $G(\beta, X) = C^0 - C(\beta, X)$, given by

$$G(\beta, X) = \sum_{(i,k) \in \mathcal{R}} \lambda_{(i,k)} \sum_{j \in \mathcal{J}} (1 - \beta_{kj}) w_{jk} + F(\beta, X), \qquad (11)$$

where

$$F(\beta, X) = \sum_{(i,k) \in \mathcal{R}} \lambda_{(i,k)} \sum_{j \in \mathcal{J}} \sum_{m=1}^{|p^{ij}|-1} w_{p^{ij}_{m+1} p^{ij}_m} \cdot \left\{ 1 - \beta_{kj} \prod_{m'=1}^{m} \left(1 - x_{p^{ij}_{m'}, i} \right) \right\}. \qquad (12)$$

We formally pose the joint user association and caching optimization problem as follows:

$$\text{Maximize:} \quad G(\beta, X) \qquad (13a)$$

$$\text{subject to:} \quad X \in \mathcal{D}_1^C \text{ and } \beta \in \mathcal{D}_1^A \qquad (13b)$$

where $\mathcal{D}_1^A$ is the set of β satisfying (1)-(3) and $\mathcal{D}_1^C$ is the set of X satisfying (4) and (5). This problem can be reduced to the 2-disjoint set cover problem [1], and is henceforce NP-hard.

III. Joint User Association and Caching

Due to the NP-hardness of problem (13), we turn our attention to efficient approximation algorithms. In this section, we introduce such an algorithm which produces a solution within a constant factor $1 - 1/e$ from the optimal. The algorithm mainly consists of two steps: convex approximation and pipage rounding.

A. Convex Approximation

In the convex approximation step, we relax the constraints in (13b) and find a convex approximation of the objective in (13a). The same approximation technique has been applied to caching [9] and joint optimization of caching and routing [10]. In this section, we show how this technique applies to our scheme.

We consider the linear relaxation of the constraints in (13b). Let y_{vi}, $v \in \mathcal{J} \cup \mathcal{N}$, $i \in \mathcal{C}$ be real-valued caching variables satisfying

$$\begin{aligned} \textstyle\sum_{i \in \mathcal{C}} y_{vi} &= c_v, \quad v \in \mathcal{J} \cup \mathcal{N}, \\ y_{vi} &\in [0,1], \quad v \in \mathcal{J} \cup \mathcal{N}, i \in \mathcal{C}, \end{aligned}$$

and $Y = [y_{vi}]_{v \in \mathcal{J} \cup \mathcal{N}, i \in \mathcal{C}} \in [0,1]^{(|\mathcal{J}| + |\mathcal{N}|) \times |\mathcal{C}|}$. Let ρ_{kj}, $k \in \mathcal{K}$, $j \in \mathcal{J}$ be real-valued association variables satisfying

$$\begin{aligned} \textstyle\sum_{j \in \mathcal{J}} \rho_{kj} &= 1, \quad k \in \mathcal{K}, \\ \textstyle\sum_{k \in \mathcal{K}} \rho_{kj} &\leq D_j, \quad j \in \mathcal{J}, \\ \rho_{kj} &\in [0,1], \quad k \in \mathcal{K}, j \in \mathcal{J}, \end{aligned}$$

and $\rho = [\rho_{kj}]_{k \in \mathcal{K}, j \in \mathcal{J}} \in [0,1]^{|\mathcal{K}| \times |\mathcal{J}|}$. We have $\rho \in \mathcal{D}_2^A$ and $Y \in \mathcal{D}_2^C$ where $\mathcal{D}_2^A$ and $\mathcal{D}_2^C$ are the convex hulls of $\mathcal{D}_1^A$ and $\mathcal{D}_1^C$ respectively. We note that the objective $G(\rho, Y)$, whose form is given in (11), is the sum of a non-negative linear function and a non-convex function $F(\rho, Y)$. The latter, whose form is given by (12), can be further approximated by a concave function $L(\rho, Y)$, given by

$$L(\rho, Y) = \sum_{(i,k)\in\mathcal{R}} \lambda_{(i,k)} \sum_{j\in\mathcal{J}} \sum_{m=1}^{|p^{ij}|-1} w_{p^{ij}_{m+1} p^{ij}_{m}} \cdot \min\left\{1, 1-\rho_{kj} + \sum_{m'=1}^{m} y_{p^{ij}_{m'}, i}\right\}. \quad (14)$$

The above approximations lead to a concave objective $H(\rho, Y)$ and the following convex optimization problem:

Maximize:

$$H(\rho, Y) = \sum_{(i,k)\in\mathcal{R}} \lambda_{(i,k)} \sum_{j\in\mathcal{J}} (1-\rho_{kj}) w_{jk} + L(\rho, Y) \quad (15a)$$

subject to: $Y \in \mathcal{D}_2^C$ and $\rho \in \mathcal{D}_2^A$ (15b)

The functions $L(\rho, Y)$ and $F(\rho, Y)$ satisfy the following inequalities (first used by Goemans and Williamson to solve the MAX SAT problem [11]):

$$(1-1/e)L(\rho, Y) \leq F(\rho, Y) \leq L(\rho, Y).$$

Then, as the first terms of $G(\rho, Y)$ and $H(\rho, Y)$ are the same non-negative linear function, we obtain:

$$(1-1/e)H(\rho, Y) \leq G(\rho, Y) \leq H(\rho, Y). \quad (16)$$

Based on (16), we can derive the following lemma showing that the association and caching gain derived by solving (15) is within a constant factor from the optimal association and caching gain of (13):

Lemma 1. *Let* (β^*, X^*) *and* (ρ^*, Y^*) *be the optimal solutions of* (13) *and* (15) *respectively. Then,*

$$G(\rho^*, Y^*) \geq (1-1/e)G(\beta^*, X^*). \quad (17)$$

Proof. By (16), $G(\rho^*, Y^*) \geq (1-1/e)H(\rho^*, Y^*) \geq (1-1/e)H(\beta^*, X^*) \geq (1-1/e)G(\beta^*, X^*)$, and the lemma follows. □

Problem (15) is a convex optimization problem and, in fact, can be converted to a linear program by introducing auxiliary variables. Thus, it can be solved in polynomial time.

B. Pipage Rounding

Given the real-valued solution (ρ^*, Y^*), where $\rho^* \in \mathcal{D}_2^A$, $Y^* \in \mathcal{D}_2^C$, we show that it is possible to round it to an integer solution $(\beta^{'}, X^{'})$, where $\beta^{'} \in \mathcal{D}_1^A$, $X^{'} \in \mathcal{D}_1^C$, with non-decreased objective value, i.e.,

$$G(\beta^{'}, X^{'}) \geq G(\rho^*, Y^*), \quad (18)$$

by leveraging the so called pipage rounding algorithm [12]. The rounding process consists of two steps:

1) For fixed ρ^*, round the caching variables Y^* to integer variables $X^{'}$, such that $G(\rho^*, X^{'}) \geq G(\rho^*, Y^*)$;
2) For fixed $X^{'}$, round the association variables ρ^* to integer variables $\beta^{'}$ such that $G(\beta^{'}, X^{'}) \geq G(\rho^*, X^{'})$.

Algorithm 1: Pipage Rounding for Association

```
Input: $\mathcal{K}, \mathcal{J}$ and $\rho$
while $\rho$ has non-integral components do
    Let $\mathcal{H}_\rho = (\mathcal{K}, \mathcal{J}; E_\rho)$, where $E_\rho$ is the set of edges
      with fractional values.
    Let $R$ be a cycle or a maximal path of $\mathcal{H}_\rho$. $R$ can
      be represented by two matchings, $M_1$ and $M_2$.
    $\epsilon_1 =$
      $\min\{\min_{(k,j)\in M_1} \rho_{k,j}, \min_{(k,j)\in M_2}(1-\rho_{k,j})\}$
    $\epsilon_2 =$
      $\min\{\min_{(k,j)\in M_1}(1-\rho_{k,j}), \min_{(k,j)\in M_2} \rho_{k,j}\}$
    if $G(\rho(-\epsilon_1, R)) > G(\rho(\epsilon_2, R))$ then
        $\rho \leftarrow \rho(-\epsilon_1, R)$
    else
        $\rho \leftarrow \rho(\epsilon_2, R)$
    end
end
return $\rho$
```

1) Rounding for caching variables: The main idea is that given a fractional caching variable $y_{vi} \in (0,1)$, $v \in \mathcal{J} \cup \mathcal{N}$, $i \in \mathcal{C}$ there must exist another fractional $y_{vi'} \in (0,1)$, for the same v and $i^{'} \in \mathcal{C}$, because of the integer cache capacity constraint, i.e., $\sum_{i\in\mathcal{C}} y_{vi} = c_v$, $v \in \mathcal{J} \cup \mathcal{N}$. Observe that objective function G, restricted to only these two entries, is convex. As such, it is maximized at the extrema of the set of values that the pair $(y_{vi}, y_{vi'})$ may take, presuming all other entries are constant. This implies that we can transfer equal mass between these two entries such that at least one of them becomes 0 or 1. Transferring equal mass ensures constraint (5) is satisfied and pairwise convexity ensures that the objective value is non-decreased. For more detailed descriptions one can refer to [9].

2) Rounding for association variables: Association variables need to satisfy an additional load-balancing constraint (3), which necessitates transferring an equal mass among more than two entries in each iteration. We summarize the pipage rounding algorithm for association variables in Alg. 1. It consists of steps at each of which a current fractional solution ρ is transformed into a new solution ρ' with a smaller number of non-integral components. At each step, we consider a bipartite graph $\mathcal{H}_\rho = (\mathcal{K}, \mathcal{J}; E_\rho)$ where $\mathcal{K}$ is the set of UEs, $\mathcal{J}$ is the set of BSs. and E_ρ is the set of edges satisfying the condition that $(k,j) \in E_\rho$ if and only if ρ_{kj} is non-integral. If $\mathcal{H}_\rho$ contains cycles, we let R be such a cycle. If $\mathcal{H}_\rho$ is a forest, we let R be a maximal path [3] of $\mathcal{H}_\rho$. Since $\mathcal{H}_\rho$ is bipartite, in both cases R can be represented as the union of two matchings M_1 and M_2. A new solution $\rho(\epsilon, R)$ is produced as follows: $\rho_{k,j}(\epsilon, R) = \rho_{k,j} + \epsilon$ if $(k,j) \in M_1$ and $\rho_{k,j}(\epsilon, R) = \rho_{k,j} - \epsilon$

[3] We say a path is maximal if you cannot add any new nodes to it to make it longer.

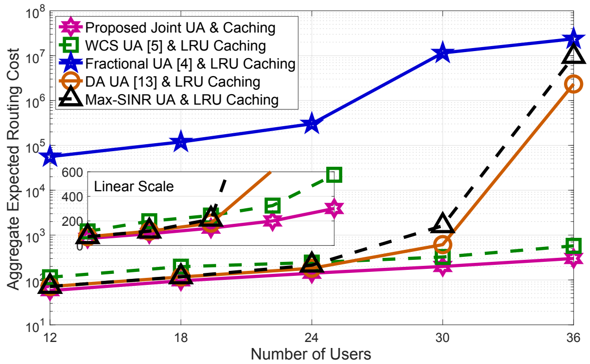

Fig. 2. Comparison of the aggregate expected routing cost with different numbers of UEs. The main figure is plotted in log-scale, while the zoomed sub-figure is presented in linear-scale.

if $(k, j) \in M_2$. Define

$$\epsilon_1 = \min\left\{ \min_{(k,j)\in M_1} \rho_{k,j},\ \min_{(k,j)\in M_2} (1-\rho_{k,j}) \right\};$$

$$\epsilon_2 = \min\left\{ \min_{(k,j)\in M_1} (1-\rho_{k,j}),\ \min_{(k,j)\in M_2} \rho_{k,j} \right\}.$$

Let $\rho_1 = \rho(-\epsilon_1, R)$, $\rho_2 = \rho(\epsilon_2, R)$. We let the new solution $\rho' = \rho_1$ if $G(\rho_1, X^{'}) > G(\rho_2, X^{'})$ and $\rho' = \rho_2$ otherwise.

The crucial property (ϵ-convexity) of G that makes this procedure work is the following: let function $\phi(\epsilon, \rho, R) = G(\rho(\epsilon, R), X^{'})$, then ϕ is convex in ϵ. In this case, the maximum value of ϕ over $[-\epsilon_1, \epsilon_2]$ is attained at one of the endpoints. Thus, in each step, we produce a new solution ρ' with a non-decreased objective value. Also, ρ' is still in D_2^A (see [12] for detailed proofs) and the number of non-integral components is reduced at least by 1. At the end of the iterations, we obtain an integral $\beta' \in D_1^A$ satisfying (18).

Combining Lemma 1 and (18), we have the following result that the process of convex approximation and pipage rounding produces an integer solution whose corresponding association and caching gain is within a constant factor from the optimal association and caching gain:

Theorem 1. *Let $(\beta^{'}, X^{'})$ be the rounded solution and (β^*, X^*) be the optimal solution for the original problem. We have:*

$$G(\beta^{'}, X^{'}) \geq (1 - 1/e)G(\beta^*, X^*)$$

As in each step, the number of fractional components is reduced by at least 1, the time complexity of the rounding process (including rounding for both caching and association variables) is $O(|\mathcal{V}| \times |\mathcal{C}| + |\mathcal{K}| \times |\mathcal{J}|)$. Thus, the algorithm that consists of convex approximation and pipage rounding can also be finished in polynomial time.

IV. Numerical Results

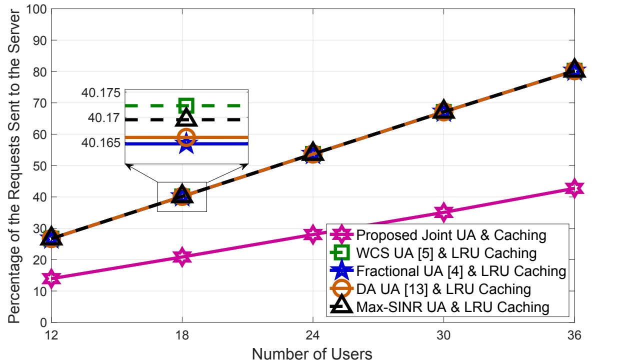

Fig. 3. Percentage of the requests served by the server versus the number of UEs.

In our simulations, we consider the mobile network topology as shown in Fig. 1. There is one server for all the content items. The mobile network includes five gateways, three MBSs operating at 1.8 GHz, and six SBSs operating at 28 GHz. MBSs and SBSs are deployed at fixed locations, while UEs are randomly located following a Poisson point process, all in a 1×1 km^2 square area. We set the load balancing constraints $D_j = 6$, $j \in \mathcal{J}_M$ and $D_j = 3$, $j \in \mathcal{J}_S$. Sub-6 GHz channels, mmWave channels, beamforming vectors, and antenna configuration are set following the simulation settings in [5]. As the majority of mobile traffic is usually caused by requests for a small subset of the content items, we focus on the caching and delivery of these most popular content items. Accordingly, we set the catalog size $|\mathcal{C}| = 20$. The initial cache sizes of gateways, MBSs, and SBSs are 2, 2, and 1 respectively. The overall request rate of each UE is generated uniformly at random (u.a.r.) from $[0, 1]$ to reflect the different activity levels of the UEs. For each UE, the request rates of different content items are generated according to Zipf distribution with an exponent 0.8. We compare our proposed algorithm with several baselines which combine Least-Recently-Used (LRU) policy for caching with several different user association (UA) algorithms:

- `Max-SINR UA & LRU`: Adopts Max-SINR for user association and LRU policy for caching.
- `DA UA & LRU`: BSs and UEs run a Deferred Acceptance (DA) matching game [13] for user association with preference lists generated based on wireless link costs. LRU is adopted as the caching scheme.
- `WCS UA & LRU`: Adopts the max-min fairness version of the worst connection swapping (WCS) algorithm [5] for user associations and LRU policy for caching.
- `Fractional UA & LRU`: Adopts the fractional user-cell association strategy [4] with LRU policy for caching.

We run simulations using MATLAB and solve the convex optimization problem (15) by CVX toolbox. Schemes with LRU caching are simulated for 1000 time units, and all statistical results stated are averaged over all observations during the simulation time, several independent runs of random locations of UEs and random channel realizations for wireless links.

We first evaluate the impact of the number of UEs in the network. As the maximum number of UEs that can be served is fixed and determined by the load constraints of BSs, given by $\sum_{j\in\mathcal{J}} D_j = 36$, by varying the number of UEs, we can evaluate the performance of the algorithms for under-loaded and fully-loaded networks. Fig. 2 evaluates the aggregate expected routing cost in the network, given by (9), with different numbers

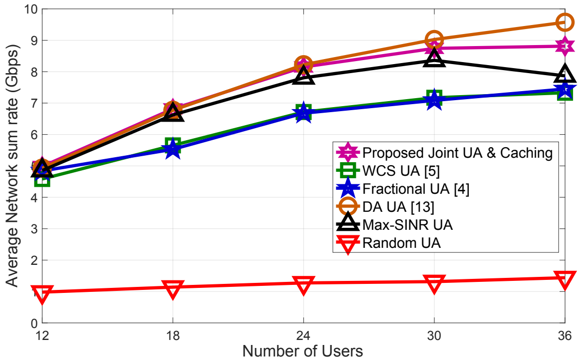

Fig. 4. Comparison of the sum rate derived by different user association schemes at the wireless access network. We also include the results of random user association in this figure.

of UEs. It shows that the proposed algorithm always leads to the lowest routing cost. As the number of UEs increases, baseline algorithms including `Fractional UA & LRU`, `DA UA & LRU`, and `Max-SINR UA & LRU` lead to high routing costs. `WCS & LRU` is the best baseline when the network is fully-loaded ($|\mathcal{K}| = 36$), however, it still produces a routing cost twice as much as the cost produced by the proposed algorithm.

Fig. 3 evaluates the ratio of requests that are served by the server as a function of the number of UEs. This shows the ability of each algorithm to serve the requests within the edge/mobile network. We can see that when the network is lightly-loaded ($|\mathcal{K}| = 12$), all algorithms are able to serve the majority of the requests at the edge. However, as the network becomes fully-loaded ($|\mathcal{K}| = 36$), all baseline algorithms let nearly 80% of the requests be served by server, compared to the proposed algorithm that lets only 42% of the requests served by server. Since the server are usually far away from the gateways and the links close to the server are more congested, the costs of serving requests at the server are high, which also explains why in Fig. 2, the aggregate expected routing costs of the baseline algorithms are higher.

In Fig. 4, we evaluate the performance in terms of the sum rate of UEs in the wireless access network, which is a widely-used metric to compare user association schemes. We can observe that, though the proposed algorithm focuses on joint optimization of user association and caching, its network sum rate is higher than most user association schemes and very close to that of `DA UA`. Fig. 5 compares the ratio of requests that are served by the server as a function of cache size. In this simulation, we consider a fully-loaded network ($|\mathcal{K}| = 36$) and vary the cache size of BSs and gateways. This shows again the ability of the proposed algorithm to efficiently serve the requests at the edge.

V. Conclusion

In this paper, the minimization of aggregate routing cost of all requests is investigated via the joint optimization of user association and caching in both edge and backhaul networks. Besides allowing caching at base stations, we also consider cache-enabled gateways which can help further improve the network efficiency (e.g. reduce the routing costs). Compared to existing works that sub-optimally divide the joint problem into subproblems, we solve it jointly via a polynomial time-complexity approximation algorithm. This algorithm also guarantees that the produced solution is within a constant factor from the optimal. By extensive simulations on a 5G network topology, we show the efficiency of the proposed algorithm in minimizing the aggregate routing cost, achieving high data sum rate and serving requests at the edge.

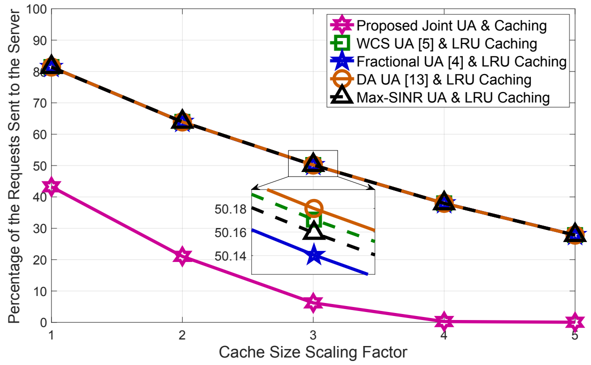

Fig. 5. Percentage of requests that go to the server versus the cache size scaling factor with 36 UEs. The initial cache size of gateways and MBSs is 2 and the cache size of SBSs is 1. To change the cache sizes, we multiply them by scaling factors 1, 2, 3, 4, and 5.

References

[1] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[2] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *MobiSys*, 2013, pp. 319–332.

[3] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, 2013.

[4] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive mimo wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1835–1850, 2016.

[5] A. Alizadeh and M. Vu, "Load balancing user association in millimeter wave mimo networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 2932–2945, 2019.

[6] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *2016 IEEE GLOBECOM*. IEEE, 2016, pp. 1–6.

[7] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3521–3525.

[8] C. Chen, T. Zhang, Y. Liu, G. Y. Li, and Z. Zeng, "Joint user association and caching placement for cache-enabled uavs in cellular networks," in *IEEE INFOCOM Workshops*. IEEE, 2019, pp. 1–6.

[9] S. Ioannidis and E. Yeh, "Adaptive caching networks with optimality guarantees," *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 737–750, 2018.

[10] ——, "Jointly optimal routing and caching for arbitrary network topologies," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1258–1275, 2018.

[11] M. X. Goemans and D. P. Williamson, "New 34-approximation algorithms for the maximum satisfiability problem," *SIAM Journal on Discrete Mathematics*, vol. 7, no. 4, pp. 656–666, 1994.

[12] A. A. Ageev and M. I. Sviridenko, "Pipage rounding: A new method of constructing algorithms with proven performance guarantee," *Journal of Combinatorial Optimization*, vol. 8, no. 3, pp. 307–328, 2004.

[13] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.