JOCAR: A Jointly Optimal Caching and Routing Framework for Cooperative Edge Caching Networks

Yuris Mulya Saputra, Dinh Thai Hoang, Diep N. Nguyen, and Eryk Dutkiewicz
School of Electrical and Data Engineering, University of Technology Sydney, Australia

***Abstract*—We propose a jointly optimal caching and routing framework (JOCAR) for a cooperative mobile edge caching network. This novel network architecture enables mobile edge servers/nodes (MENs) to collaborate in not only caching but also routing contents to users, in order to simultaneously minimize the total content-access delay for all mobile users and reduce the traffic on the backhaul network. To that end, we first formulate an access-delay minimization problem by jointly optimizing the content caching and routing decisions while accounting for various network configurations. Solving this problem requires us to deal with a nested dual optimization due to the strong mutual dependence between content caching and routing decisions. To tackle it, we first transform the nested dual problem to an equivalent mixed-integer nonlinear programming (MINLP) problem. Then, we design a branch-and-bound based algorithm with the interior-point method to find the optimal policy for the MINLP problem. Extensive simulations show that JOCAR can reduce the total average delay and increase the cache hit rate for the whole network by more than 40% and by four times, respectively, compared with other conventional policies.**

Keywords- Mobile edge caching, joint caching and routing, branch-and-bound, mutual dependency, latency.

I. Introduction

According to the Cisco's latest forecast, a massive number of smart devices together with a huge demand for emerging services in the networks will drive the data traffic explosion in the near future [1]. Mobile edge caching (MEC) has been introduced as a promising solution to address the ever-increasing mobile data traffic. This MEC can distribute popular contents as well as computing resources closer to mobile users by deploying servers at the "edge" of the network, referred to as mobile edge node/servers (MENs). As a result, the MEC networks can remarkably reduce the service delay and mitigate network congestion on the backhaul link.

Nonetheless, due to special characteristic of MEC networks, e.g., small coverage and limited storage capacity, cooperative caching has been used to distribute the popular contents through the cooperation among the nodes. For example, the authors in [2] introduced a hierarchical cooperative caching model taking the request hit probability and caching capacities of MENs into account to find an optimal cooperative caching strategy. In [3], a shared caching model was proposed to reduce the content delivery delay for mobile users from different network providers. In another work, a cooperative two-level caching cluster model in order to minimize the total bandwidth cost for the system was considered in [4]. As an extension, [5] then took the network topology into account while optimizing the content placement strategy.

Although some cooperative caching schemes can reduce service delay, they only focus on content placement strategies. To minimize more service delay, some researches have studied the joint caching and delivery (referred to as routing) strategies through *vertical* cooperation (e.g., between MENs and the base station, between the base station and the CS). In [6], the authors developed a cooperative caching and delivering policy for heterogeneous cellular networks (HetNets) with femtocells and D2D communications. However, in this work, caching placement and delivery are separately optimized. The authors in [7] studied a joint content caching and request routing in a hybrid network in which stored contents can be accessed through vertical paths (from each MEN to the CS). Nevertheless, both works did not leverage the horizontal routing among MENs. In fact, in addition to vertical cooperation, the *horizontal* cooperation among MENs in delivering contents to users has a great potential in reducing service delay. The reason is that MENs are often deployed in a close proximity area [3], [8] and the links among MENs are much faster than those from MENs to the base station (BS).

Given the above, we develop a jointly optimal cooperative caching and routing (JOCAR) framework that allows MENs to collaborate in caching and routing/delivering contents to users. To that end, we first introduce a novel MEC network model in which MENs can directly communicate with some other MENs, in addition to CS via the BS. Then, given the content demand distributions, diverse data sizes, various MENs' storage capacities, and network topology, the proposed framework needs to jointly address two questions (1) how to place contents at the MENs and (2) how to choose the best routes to deliver contents.

However, the aforementioned joint optimization is intractable due to the strong mutual dependence between content placement and routing decisions. Specifically, where to cache contents will impact how to deliver contents and how to deliver contents will have an influence on how/where to cache them. As a result, this leads to a nested dual optimization problem that is proved to be NP-hard. To tackle it, we propose a novel transformation method to transform the nested dual problem to an equivalent MINLP optimization problem. Then, we develop a branch-and-bound algorithm with the interior-point method to find the optimal solution of the original nested dual problem. Extensive simulations demonstrate that the proposed framework can reduce the total average delay and increase the cache hit rate for the whole network by more than 40% and up to 3-4 times, respectively, compared with other conventional caching policies. The major contributions of the paper can be summarized as follows:

- Design the JOCAR framework that enables horizontal cooperations among MENs to reduce the total delay for the MEC network and backhaul traffic load.

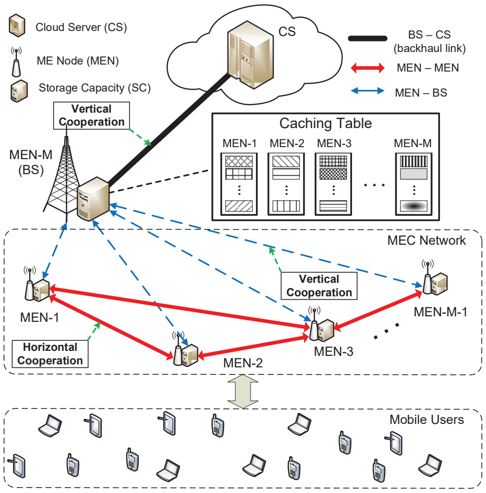


Fig. 1: Proposed cooperative MEC network model

- Formulate the jointly content caching and routing problem as the nested dual problem with mutual dependency.
- Propose a novel method to transform the nested dual problem into the equivalent MINLP problem.
- Develop the improved branch-and-bound based algorithm with interior-point method to find the optimal jointly caching and routing policy for the nested dual problem.

II. Cooperative MEC Network Model

Fig. 1 describes our proposed cooperative MEC network model of the JOCAR framework. Each MEN can connect to its nearby MENs (with overlapping and non-overlapping areas) including the BS using wireless (e.g., Wi-Fi) or wired interfaces through horizontal cooperation. Mobile users can move from one MEN's coverage to another MEN's area (including overlapping coverages of MENs). Each MEN and the BS are equipped with a finite storage to cache high-demand contents. To find the best place to download a requested content, the BS provides to each MEN a caching and routing table. This table is constructed by solving the nested dual optimization problem formulated in Section III. Based on this table, an MEN, upon receiving a content request from its mobile user, will send the content to the user immediately if the content is cached locally. If the content is not cached locally but at other MENs that are directly connected to the requesting MEN, the MEN will download the content from the node which has the lowest delivery delay (instructed in the table) before delivering to the user. Otherwise, the MEN, through vertical cooperation, will download the content from the CS or one MEN that is not directly connected to the requesting MEN (via the BS).

Let $\mathcal{M} = \{1, \dots, m, \dots, M\}$ denote the set of MENs including the BS (denoted by M as a special MEN). c_m and U_m denote storage capacity and the number of mobile users in the coverage area of MEN-m, respectively. B denotes the bandwidth between the BS and the CS. We also define $b_m^n, \forall m, n \in \mathcal{M}$ and $m \neq n$, and $b_m^u, \forall m \in \mathcal{M}$ as the available bandwidth between MEN-m and MEN-n, and the allocated bandwidth of a user u to its MEN-m, respectively. Furthermore, we define $\mathcal{K} = \{1, \dots, k, \dots, K\}$ as the set of contents. The contents may have diverse data sizes denoted by $\mathcal{S} = \{s_1, \dots, s_k, \dots, s_K\}$. The users' demands at each MEN-m are captured by its frequency-of-access (FoA) for content k as f_m^k.

III. Problem Formulation and Transformation of Joint Caching and Routing

In this section, we first formulate the joint cooperative caching and routing problem, and then introduce the novel method to transform the problem into the equivalent MINLP for optimal caching and routing policy decision.

A. Problem Formulation

We define $\mathbf{x} \overset{\text{def}}{=} [\mathbf{x}_1, \dots, \mathbf{x}_m, \dots, \mathbf{x}_M]^T$, where $\mathbf{x}_m = [x_m^1, \dots, x_m^k, \dots, x_m^K]$ and $x_m^k \in \{0, 1\}$, as the set of binary decision variables at the MENs. $x_m^k = 1$ means that content k is cached at MEN-m, and $x_m^k = 0$ otherwise. When a user u at MEN-m requests content k, the following three cases can occur. For **CASE 1**, MEN-m contains the requested content (i.e., $x_m^k = 1$). Then, the delay to download the content will be $x_m^k d_\alpha^m$ where $d_\alpha^m = \frac{s_k}{b_m^u}$. For **CASE 2**, MEN-$m$ does not have the requested content (i.e., $x_m^k = 0$). However, at least one MEN from the set of MENs (denoted by the set $\mathcal{N}_m^k$) that are directly connected to MEN-m has this content k, i.e., $x_{n*}^k \overset{\text{def}}{=} \prod_{n \in \mathcal{N}_m^k} (1 - x_n^k) = 0$. Then, MEN-$m$ will download content k from one node which has the lowest delivery time, and thus the delay to download the content in this case will be $(1 - x_m^k)(1 - x_{n*}^k)(d_\alpha^m + d_\beta^m)$, where $d_\beta^m \overset{\text{def}}{=} \min_n \Big(\big[x_n^k + (1 - x_n^k) G \big] \frac{s_k}{b_m^n} \Big)$ and G is a very large constant number. The use of value G is to ensure that MENs without containing the requested content k are practically ignored. Furthermore, in the case $m = M$ (i.e., the BS), only **CASE 1** and **CASE 2** can happen as the BS is directly connected to the CS. For **CASE 3**, MEN-m and all of its directly connected nodes do not contain the requested content, i.e., $x_m^k = 0$ and $x_{n*}^k = 1$. Then, there are two possibilities. First, there is no MEN storing the content. Then, the content will be downloaded from the CS via the BS, and the delay will be $(1 - x_m^k) x_{n*}^k d_\delta^m$, where $d_\delta^m = (d_\alpha^m + \frac{s_k}{b_m^n} + \frac{s_k}{B})$. Second, if there is at least one MEN that is not directly connected to MEN-m, say MEN-n (with some abusing of notation), storing this content, MEN-m will download the content from either the CS via the BS or from MEN-n via the BS or any intermediate MEN (whichever has lower delivery delay). In practice, the bandwidth between the BS and CS is usually much higher than among MENs, and thus the content will be downloaded from the CS in the second case. In this way, the delay to download the content in **CASE 3** will be d_δ^m.

Then, the jointly optimal cooperative caching and routing optimization problem $(\mathbf{Q}_1)$ can be formulated as follows:

$$(\mathbf{Q}_1) \quad \min_{\mathbf{x}} F(\mathbf{x}), \tag{2}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} x_m^k s_k \leq c_m, \forall m \in \mathcal{M}, \tag{3}$$

$$x_m^k, x_n^k \in \{0, 1\}, \forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \tag{4}$$

$$F(\mathbf{x}) = \sum_{k=1}^{K}\sum_{m=1}^{M} f_m^k \Bigg[\sum_{u=1}^{U_m} \Bigg(\underbrace{x_m^k d_\alpha^m}_{\textbf{CASE 1}} + \underbrace{(1-x_m^k)(1-x_{n^*}^k)\Bigg(d_\alpha^m + \min_n \Big([x_n^k + (1-x_n^k)G]\frac{s_k}{b_m^n}\Big)\Bigg)}_{\textbf{CASE 2}} + \underbrace{(1-x_m^k)x_{n^*}^k d_\delta^m}_{\textbf{CASE 3}} \Bigg) \Bigg]$$
$$= \sum_{k=1}^{K}\sum_{m=1}^{M} f_m^k \Bigg[\sum_{u=1}^{U_m} \Bigg(x_m^k d_\alpha^m + (1-x_m^k)\Big[(1-x_{n^*}^k)(d_\alpha^m + d_\beta^m) + x_{n^*}^k d_\delta^m\Big]\Bigg)\Bigg] \tag{1}$$

where the objective function $F(\mathbf{x})$ in Eq. (1) represents the total average delay for the whole network. The constraints in Eq. (3) guarantee that the total size of all cached contents does not exceed the storage capacity of each MEN. Moreover, the constraints in Eq. (4) show that caching decision variables (i.e., x_m^k and x_n^k in Eq. 1) are binary. Based on ($\mathbf{Q}_1$), our optimization problem is a *nested dual binary nonlinear programming* with a binary decision variable vector used in both minimization functions. In particular, this optimization has two levels: (1) outer level (OL) which contains the main objective function $F(\mathbf{x})$ and (2) inner level (IL) which contains the optimal routing decision problem $F^*(\mathbf{x}) = d_\beta^m$. In Lemma 1, we show that problem ($\mathbf{Q}_1$) is an NP-hard problem.

Lemma 1. *The nested dual optimization problem with mutual dependency ($\mathbf{Q}_1$) is NP-hard.*

Proof. Refer to the proof of Lemma 1 in [9]. □

B. Problem Transformation

In ($\mathbf{Q}_1$), we need to simultaneously optimize the content caching and routing decisions. However, where to store contents will be influenced by how we determine to route/deliver the contents (to minimize the total average delay). Similarly, the decisions to route/deliver the contents will be impacted by the decisions where we cache the contents. Thus, caching and routing the contents have mutual-dependent decisions.

To resolve this problem, we first propose a novel method to transform the nested dual optimization problem into the equivalent MINLP optimization problem. In particular, we use $O_m^k(\mathbf{x})$ along with an additional set of binary variables $\mathbf{y}$ and a set of non-integer variables $\mathbf{z}$ to replace the IL problem $F^*(\mathbf{x})$. In this way, if we denote $F(\mathbf{x},\mathbf{y},\mathbf{z}) = \sum_{k=1}^{K}\sum_{m=1}^{M} f_m^k \Big[\sum_{u=1}^{U_m} \Big(x_m^k d_\alpha^m + (1-x_m^k)\Big[(1-x_{n^*}^k)(d_\alpha^m + z_m^k) + x_{n^*}^k d_\delta^m\Big]\Big)\Big]$, then new equivalent optimization problem ($\mathbf{Q}_2$) can be expressed as follows:

$$(\mathbf{Q}_2) \quad \min_{\{\mathbf{x},\mathbf{y},\mathbf{z}\}} F(\mathbf{x},\mathbf{y},\mathbf{z}), \tag{5}$$

$$\text{s.t. (3)-(4) and}$$
$$z_m^k \geq O_m^k(x_n^k) - Gy_n^k, \forall m,n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \tag{6}$$
$$\sum_{n \in \mathcal{N}_m^k} y_n^k = N_m^k - 1, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \tag{7}$$
$$y_n^k \in \{0,1\}, \forall n \in \mathcal{M}, \forall k \in \mathcal{K}, \tag{8}$$
$$z_m^k \in \mathbb{R}_0^+, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \tag{9}$$

where N_m^k is the cardinality of the set $\mathcal{N}_m^k$, $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_M]^T$ with $\mathbf{y}_n = [y_n^1, \ldots, y_n^k, \ldots, y_n^K], y_n^k \in \{0,1\}$, and $m \neq n$, while $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_m, \ldots, \mathbf{z}_M]^T$ with $\mathbf{z}_m = [z_m^1, \ldots, z_m^k, \ldots, z_m^K]$, where $z_m^k \in \mathbb{R}_0^+$. The constraints in Eq. (6) represent the selection of one MEN-n which has the lowest delivery time. Furthermore, the constraints in Eq. (7) and Eq. (8) aim to guarantee that only one variable y_n^k is set to be "0", while the rest of variables are set to be "1". Specifically, $N_m^k - 1$ indicates that we exclude one node which has $y_n^k = 0$ (i.e., the selected MEN-n to deliver the content k) when N_m^k number of directly connected MENs storing content k for MEN-m are considered. The equivalent transformation is formally stated in Theorem 1.

Theorem 1. *The nested dual optimization problem ($\mathbf{Q}_1$) is equivalent to MINLP optimization problem ($\mathbf{Q}_2$).*

Proof. See Appendix A. □

IV. Improved Branch-and-Bound Algorithm with Interior-Point Method

To solve the aforementioned MINLP problem, we first adopt the branch-and-bound algorithm (BBA) which can reduce the time complexity by leveraging the characteristics of binary variables to find the optimal solution of ($\mathbf{Q}_2$). Then, to address the non-linearity and continuous relaxation of ($\mathbf{Q}_2$), we utilize the interior-point method (IPM) which has polynomial time complexity. Overall, both BBA and IPM can be integrated to address binary and continuous variables of ($\mathbf{Q}_2$) efficiently.

Given the number of contents, i.e., K, and the number of nodes, i.e., M, the number of variables of vector $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are $J_\mathbf{x} = J_\mathbf{y} = J_\mathbf{z} = I \times N$. The BBA first relaxes all binary variables x_m^k, x_n^k, and y_n^k of ($\mathbf{Q}_2$) into continuous variables at the root problem ($\mathbf{RQ}$). In the ($\mathbf{RQ}$), the relaxed binary variables are bounded by $0 \leq x_m^k, x_n^k, y_n^k \leq 1$. If we denote $F_{\mathbf{RQ}}(\mathbf{x},\mathbf{y},\mathbf{z}) = \sum_{k=1}^{K}\sum_{m=1}^{M} f_m^k \Big[\sum_{u=1}^{U_m} \Big(\Upsilon\big(x_m^k, x_n^k\big) + \Psi\big(x_m^k, x_n^k\big) z_m^k \Big)\Big]$, where $\Upsilon\big(x_m^k, x_n^k\big) \overset{\text{def}}{=} x_m^k d_\alpha^m + (1-x_m^k)(1-x_{n^*}^k)d_\alpha^m + (1-x_m^k)x_{n^*}^k d_\delta^m$, and $\Psi\big(x_m^k, x_n^k\big) \overset{\text{def}}{=} (1-x_m^k)(1-x_{n^*}^k)$ are the simplified forms from some parts of the objective function in ($\mathbf{Q}_2$), then the ($\mathbf{RQ}$) can be expressed as follows:

$$(\mathbf{RQ}) \quad \min_{\{\mathbf{x},\mathbf{y},\mathbf{z}\}} F_{\mathbf{RQ}}(\mathbf{x},\mathbf{y},\mathbf{z}), \tag{10}$$

$$\text{s.t. } \Theta\big(x_m^k\big) \leq 0, \forall m \in \mathcal{M}, \tag{11}$$
$$O_m^k\big(x_n^k\big) - Gy_n^k - z_m^k \leq 0, \forall m,n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \tag{12}$$
$$\Gamma_m\big(y_n^k\big) = 0, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \tag{13}$$
$$0 \leq x_m^k, x_n^k, y_n^k \leq 1, z_m^k \in \mathbb{R}_0^+, \forall m,n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \tag{14}$$

where $\Theta\big(x_m^k\big) \overset{\text{def}}{=} \sum_{k=1}^{K} x_m^k s_k - c_m$ and $\Gamma_m\big(y_n^k\big) \overset{\text{def}}{=} \sum_{n \in \mathcal{N}_m^k} y_n^k - (N_m^k - 1)$ are the simplified expressions of the constraints in Eq. (3) and Eq. (7), respectively. If a feasible solution is obtained, i.e., all decision variables x_m^k, x_n^k, and y_n^k are binary values, the algorithm will terminate. Otherwise, it will break the $(\mathbf{RQ})$ into subproblems $(\mathbf{SQ})$s (i.e., branch problems). In this way, the $(\mathbf{SQ})$s fix one of the fractional decision variables (i.e., x_m^k, x_n^k, or y_n^k) to be "0" at the left branch and "1" at the right branch. We denote the fixed decision variable to be $\theta \in \{x_m^k, x_n^k, y_n^k\}$ with $F_{\mathbf{SQ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = F_{\mathbf{RQ}}(\mathbf{x}, \mathbf{y}, \mathbf{z})$.

In the BBA, the $(\mathbf{SQ})$ is pruned if one of the following conditions is met: (1) $\phi^c > \beta_U$, (2) $\phi^c \geq \phi$, or (3) $\phi^c < \phi$, where ϕ^c, ϕ, and β_U are the current total average delay, the incumbent total average delay, and the upper bound of the total average delay, respectively. Then, the optimal solution $\hat{x}_m^k$, $\hat{x}_n^k$, $\hat{y}_n^k$, and $\hat{z}_m^k$ with $\forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}$ are obtained if the total average delay satisfies $\hat{\phi} = \phi \leq \phi^c$ for all x_m^k, x_n^k, y_n^k, and z_m^k in the problem. To guarantee that the optimal solution exists, we set an optimality tolerance ζ as a non-negative value. Specifically, $\hat{x}_m^k$, $\hat{x}_n^k$, $\hat{y}_n^k$, and $\hat{z}_m^k$ are $\zeta-$optimal when the total average delay $\hat{\phi}$ is tightened within the bounds (i.e., $\beta_L \leq \hat{\phi} \leq \beta_U$) and the difference between the upper and lower bound is less than the optimality tolerance (i.e., $\beta_U - \beta_L \leq \zeta$).

While the BBA handles the feasibility of the $(\mathbf{SQ})$s and the optimality of $(\mathbf{Q}_2)$, the interior-point method (IPM) is used to solve the nonlinear continuous relaxation of the $(\mathbf{SQ})$s. In particular, we derive an interior-point subproblem $(\mathbf{IQ})$ from the $(\mathbf{SQ})$ as the approximation problem such that

$$
\begin{aligned}
&F_{\mathbf{IQ}_\gamma}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \xi_\mathbf{1}, \xi_\mathbf{2}, \xi_\mathbf{3}, \xi_\mathbf{4}) = \\
&\min_{\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}} F_{\mathbf{SQ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \gamma \Bigg[\sum_{i_1=1}^{M} \log\, \xi_1^{i_1} + \sum_{i_2=1}^{N} \log\, \xi_2^{i_2} \\
&+ \sum_{i_3=1}^{I_\mathbf{x}+I_\mathbf{y}} \log\, \xi_3^{i_3} + \sum_{i_4=1}^{I_\mathbf{x}+I_\mathbf{y}} \log\, \xi_4^{i_4} \Bigg],
\end{aligned} \tag{15}
$$

and thus

$$
(\mathbf{IQ}) \quad \min_{\substack{\mathbf{x}, \mathbf{y}, \mathbf{z}, \\ \{\xi_\mathbf{1}, \xi_\mathbf{2}, \xi_\mathbf{3}, \xi_\mathbf{4}\}}} F_{\mathbf{IQ}_\gamma}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \xi_\mathbf{1}, \xi_\mathbf{2}, \xi_\mathbf{3}, \xi_\mathbf{4}), \tag{16}
$$

$$
\begin{aligned}
\text{s.t.} \quad \Theta\big(x_m^k\big) + \xi_1^{i_1} = 0, \forall m, i_1 \in \mathcal{M}, \quad & (17) \\
O_m^k\big(x_n^k\big) - G y_n^k - z_m^k + \xi_2^{i_2} = 0, \forall i_2 \in [1, N], \quad & \\
\forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \quad & (18) \\
\Gamma_m\big(y_n^k\big) = 0, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \quad & (19) \\
V - \xi_3^{i_3} = 0, V + \xi_4^{i_4} = 1, \forall i_3, i_4 \in [1, I_\mathbf{x} + I_\mathbf{y}], \quad & \\
\forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \quad & (20) \\
z_m^k \in \mathbb{R}_0^+, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \quad & (21)
\end{aligned}
$$

where $N \overset{\text{def}}{=} \sum_{k=1}^{K} \sum_{m=1}^{M} N_m^k$, $V = \{x_m^k, x_n^k, y_n^k\}$, $\gamma > 0$ is the barrier parameter, $\xi_\mathbf{1}$, $\xi_\mathbf{2} \in \mathbb{R}_0^+$ are slack variables in the inequality constraints (11) and (13), respectively, and $\xi_\mathbf{3}$, $\xi_\mathbf{4} \in \mathbb{R}_0^+$ are slack variables for lower and upper bounds of V in the inequality constraints (14), respectively. The improved BBA-IPM (iBBA-IPM) algorithm is shown in Algorithm 1.

Algorithm 1 iBBA-IPM

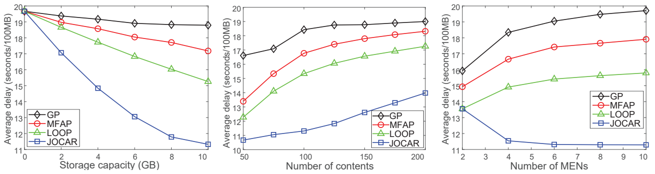
```
$\mathcal{R}_a$: the set of active problems $r_a$, $r_c$: the current problem
Set $V \in [0, 1]$, $z_m^k \in \mathbb{R}_0^+$, $\forall m, n \in \mathcal{M}$, $m \neq n$, $\forall k \in \mathcal{K}$
$r_c \leftarrow r_a \leftarrow 1$, $\beta_L \leftarrow -\infty$, $\beta_U \leftarrow +\infty$, $\phi \leftarrow \beta_U$
Set $\tau$ and $\zeta$ /*Integrality gap & optimality tolerance*/
Solve $(\mathbf{RQ})$ for $r_c$
if $\forall x_m^k, x_n^k, y_n^k \in \{0, 1\}$ then
    Store $\hat{x}_m^k, \hat{x}_n^k, \hat{y}_n^k, \hat{z}_m^k$, and $\hat{\phi}$, $\forall m, n \in \mathcal{M}$, $\forall k \in \mathcal{K}$
    return  /*Prune all active problems*/
end if
while $\mathcal{R}_a \neq \varnothing$ and $\beta_U - \beta_L > \zeta$ do
    $r_c \leftarrow r_a$ /*Choose an $(\mathbf{SQ})$ from depth-first-search*/
    $r_a \leftarrow r_a - 1$ /*Remove the $(\mathbf{SQ})$ from the set*/
    Solve subproblem $(\mathbf{IQ})$ for $r_c$
    if $(\mathbf{IQ})$ for $r_c$ is infeasible then
        Prune $r_c$, exit
    else
        Set current $V$ and $z_m^k$, $\forall m, n \in \mathcal{M}$, $\forall k \in \mathcal{K}$
        Calculate $\sigma \leftarrow |V - \text{round}(V)|$
        if $\sigma < \tau$ then
            if $\phi^c < \phi$ then
                Store current $V$ and $z_m^k$, $\forall m, n \in \mathcal{M}$, $\forall k \in \mathcal{K}$
                Set $\phi \leftarrow \phi^c$ and $\beta_U \leftarrow \phi^c$
            end if
            Prune $r_c$, exit
        else
            Choose $\theta \in V$
            $r_a \leftarrow r_a + 2$, $\beta_L \leftarrow \phi^c$ /*Add 2 $(\mathbf{SQ})$ subproblems to
            the set*/
            Update the constraints with $\theta \leftarrow 0$, $\theta \leftarrow 1$, exit
        end if
    end if
    Store $\hat{x}_m^k, \hat{x}_n^k, \hat{y}_n^k, \hat{z}_m^k$, and $\hat{\phi}$, $\forall m, n \in \mathcal{M}$, $\forall k \in \mathcal{K}$
end while
```

V. Simulation Results

In this section, we conduct extensive simulations to evaluate the performance of the proposed JOCAR. In particular, we vary important parameters and compare the performance of JOCAR with those of other optimal caching policies including (1) Greedy Policy (GP) where MENs try to cache as many contents as possible [10], (2) Most frequency of Access Policy (MFAP) where MENs try to cache contents with high FoA [11], and Locally Optimal Policy (LOOP) where MENs try to cache contents to minimize their delay locally [12]. In all simulations, the content size is generated using a uniform distribution between 100 and 300 MB, while the FoA follows a Zipf distribution [3], [8] with shape parameter 0.1. The bandwidth between a mobile user and its connected MEN and between two MENs are set to be 10 and 45 Mbps, respectively. Additionally, bandwidth between an MEN and the BS and between the BS and the CS are set to be 10Mbps and 60Mbps, respectively. Note that the bandwidth between two MENs is usually higher than that between an MEN and the BS because the MENs are often deployed in a close proximity area where wired and/or fast wireless links are in place [3], [8].

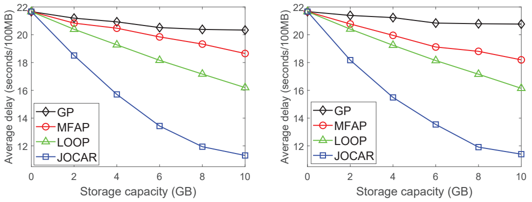
A. Total Average Delay

Fig. 2(a) shows the trend of the total average delay in the MEC network when the storage capacity increases from 0 to 10 GB. It can be observed that JOCAR dramatically



(a) (b) (c)

Fig. 2: Average delay as (a) the storage capacity increases, (b) the number of contents increases, and (c) the number of MENs increases



(a) (b)

Fig. 3: Average delay when 2 MENs are used: (a) MEN-1 and (b) MEN-2

reduces the average delay by as much as 40%, 55%, and 74%, compared with those of LOOP, MFAP, and GP, respectively. The delay reduction by JOCAR is greater as the MEN storage capacity increases. This trend aligns with the average delay for each MEN as observed in Fig. 3. The reason is that the MENs can collaborate using direct horizontal cooperations to improve caching efficiency for the whole network. Then, Fig. 2(b) shows the total average delay when the number of contents increases from 50 to 200 contents. We fix the storage capacity for all MENs (including the BS) at 5GB. As expected, for a given storage capacity, if the number of available contents increases, the average delays obtained by all policies will increase. This is because more contents must be downloaded from the CS as the MEN's storage capacity decreases. Nevertheless, the average delay obtained by JOCAR is still much lower than those of all other policies. In Fig. 2(c), we increase the number of MENs from 2 to 10, and fix the storage capacity of all the nodes and the number of contents at 10 GB and 200 contents, respectively. Interestingly, as the number of MENs increases, the average delays obtained by GP, MFAP, and LOOP increase gradually. However, the average delay obtained by JOCAR first dramatically decreases when the number of MENs increases from 2 to 4, then slightly reduces and remains stable after 8. The reason is that for GP, MFAP, and LOOP, they do not consider the collaboration, and thus as the number of nodes increases, the delay at each MEN will contribute to the total delay of the network, yielding to an increasing trend. Nevertheless, JOCAR can utilize the efficiency of collaboration among the MENs, and thus as the number of MENs increases, more connections will be created, thereby greater leveraging the collaboration among MENs.

B. Cache Hit Rate Probability

In this section, we consider two types of the cache hit rate (1) *hit-itself* and (2) *hit-others*. The former represents the cache hit rate at each individual MEN, while the latter captures the cache hit rate by any MEN in the network. Specifically, h_m denotes the hit-itself rate of MEN-m. Then, the overall average cache hit rate in the MEC network (of all MENs) of the GP, MFAP, and LOOP can be calculated by $h_t \overset{\text{def}}{=} \frac{1}{M}\Big[\sum_{m=1}^{M-1} h_m + M h_M\Big]$, and the overall average cache hit rate in the MEC network for the JOCAR is $h_t^* \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M}\Big(h_m + \sum_{n \in \mathcal{N}_m^k, m \neq n} h_n\Big), \forall k \in \mathcal{K}$, where h_M and $\sum_{n \in \mathcal{N}_m^k, m \neq n} h_n$ are the hit-others of the conventional policies (i.e., GP, MFAP, and LOOP) and the JOCAR, respectively.

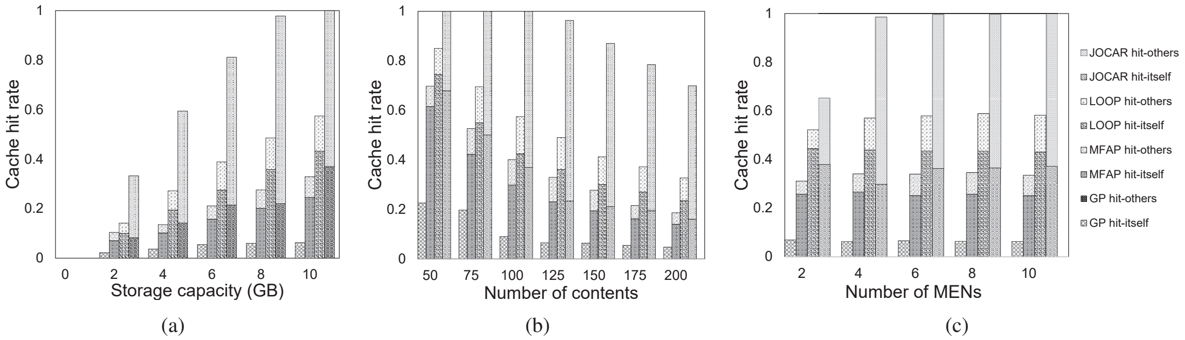
As observed in Fig. 4, the GP only shows the hit-itself because the GP is based on the sizes of contents only, and thus the MENs and the BS have the same set of cached contents. In addition, the average cache hit rates of MFAP and LOOP are greater than that of the GP because they have less duplicate contents on the MENs. Particularly, the MENs may have different users' demands, and thus they may cache dissimilar contents. For the JOCAR, due to the collaboration, the total cache hit rate (including hit-itself and hit-others) obtained by JOCAR is as much as 4 times higher than those of other policies. This figure also clearly shows impacts of the collaboration among MENs. In particular, although the hit-itself may not always be the best (because MENs may sacrifice to cache contents for other nodes to minimize the overall delay for the network), but the total cache hit rate for the network obtained by JOCAR always achieves the highest value. It is also worth noting that when the total cache hit rate in the network increases, the traffic load on the backhaul network will be reduced.

VI. Conclusion

In this paper, we have introduced JOCAR, an effective jointly optimal cooperative caching and routing framework considering horizontal cooperation among MENs. JOCAR aims to minimize the total delay for the MEC network and reduce the network traffic on the backhaul network. To find the optimal caching and routing policy for the nested dual optimization of JOCAR, we have introduced the novel transformation method and proposed the improved branch-and-bound algorithm with the interior-point method. Through the simulation results, we have demonstrated that the proposed solution can significantly outperform all other caching methods in terms of the total average delay and cache hit rate for the whole network.

Appendix A
Proof of Theorem 1

First, recall $F^*(\mathbf{x})$ from $(\mathbf{Q}_1)$ as follows: $\min_n\Big([x_n^k + (1 - x_n^k)G]\frac{s_k}{b_m^n}\Big) \overset{\text{def}}{=} \min_n\Big(O_m^k(x_1^k), \ldots, O_m^k(x_n^k), \ldots, O_m^k(x_{N_m^k}^k)\Big)$, where $O_m^k(x_n^k) \overset{\text{def}}{=} [x_n^k + (1 - x_n^k)G]\frac{s_k}{b_m^n}$. Then, we rewrite Eq. (2) as follows $\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} \sum_{k=1}^{K}\sum_{m=1}^{M} f_m^k\Big[\sum_{u=1}^{U_m}\Big(\Upsilon(\mathbf{x}) + \Psi(\mathbf{x})\min_n\Big(O_m^k(x_1^k), \ldots, O_m^k(x_n^k), \ldots, O_m^k(x_{N_m^k}^k)\Big)\Big)\Big]$, where $\mathbf{x} = (x_m^k, x_n^k)$, $\forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}$. Consider



(a)

(b)

(c)

Fig. 4: Average cache hit rate when (a) the storage capacity increases, (b) the number of contents increases, and (c) the number of MENs increases

$\mathbf{x}^* = (\hat{x}^k_m, \hat{x}^k_n), \forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}$ as the optimal solution of $F(\mathbf{x})$. Without loss of generality, $n = j^k_m$ will be selected to be the MEN with minimum delivering decision to download content k for MEN-m if the following condition is satisfied: $O^k_m(\hat{x}^k_{n=j^k_m}) \leq O^k_m(\hat{x}^k_{n\neq j^k_m}), \forall m, n, j^k_m \in \mathcal{M}, m \neq n, j^k_m \neq m, \forall k \in \mathcal{K}$, and thus

$$F(\mathbf{x}^*) = \sum_{k=1}^{K}\sum_{m=1}^{M} f^k_m \left[\sum_{u=1}^{U_m}\Big(\Upsilon\big(\mathbf{x}^*\big) + \Psi\big(\mathbf{x}^*\big)O^k_m(\hat{x}^k_{n=j^k_m})\Big)\right]. \tag{22}$$

Next, from $(\mathbf{Q}_2)$, we also have $\min\limits_{\{\mathbf{x},\mathbf{y},\mathbf{z}\}} F(\mathbf{x},\mathbf{y},\mathbf{z}) = \min\limits_{\{\mathbf{x},\mathbf{y},\mathbf{z}\}} \sum_{k=1}^{K}\sum_{m=1}^{M} f^k_m \Big[\sum_{u=1}^{U_m}\Big(\Upsilon\big(\mathbf{x}\big) + \Psi\big(\mathbf{x}\big)z^k_m\Big)\Big]$, s.t. (6)-(9). Given $\mathbf{x}^*$ from the Eq. (22), we can rewrite:

$$\begin{aligned}
&\min_{\{\mathbf{x}^*,\mathbf{y},\mathbf{z}\}} F(\mathbf{x}^*,\mathbf{y},\mathbf{z}) = \\
&\min_{\{\mathbf{x}^*,\mathbf{y},\mathbf{z}\}} \sum_{k=1}^{K}\sum_{m=1}^{M} f^k_m \left[\sum_{u=1}^{U_m}\Big(\Upsilon\big(\mathbf{x}^*\big) + \Psi\big(\mathbf{x}^*\big)z^k_m\Big)\right], \\
&\text{s.t.} \begin{cases}
z^k_m \geq O^k_m(\hat{x}^k_n) - Gy^k_n, \forall m, n \in \mathcal{M}, m \neq n, \forall k \in \mathcal{K}, \\
\sum\limits_{n\in\mathcal{N}^k_m} y^k_n = N^k_m - 1, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}, \\
y^k_n \in \{0,1\}, \forall n \in \mathcal{M}, \forall k \in \mathcal{K}, \\
z^k_m \in \mathbb{R}^+_0, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}.
\end{cases}
\end{aligned} \tag{23}$$

Then, take all feasible values of y^k_n in Eq. (23) into account. From $\sum_{n\in\mathcal{N}^k_m} y^k_n = N^k_m - 1$, there are only one MEN-n with $y^k_n = 0$ and $N^k_m - 1$ MENs with $y^k_n = 1$. Consider that $n = \nu^k_m$, where $\forall \nu^k_m \in \mathcal{N}^k_m, \forall m \in \mathcal{M}, \forall k \in \mathcal{K}$, is the MEN which has $y^k_{n=\nu^k_m} = 0$, and thus $z^k_m \geq O^k_m(\hat{x}^k_{n=\nu^k_m})$ for $n = \nu^k_m$ and $z^k_m \geq O^k_m(\hat{x}^k_{n\neq\nu^k_m}) - G$ for $n \neq \nu^k_m$. Since G is a very big value, we can remove the rest of the constraints when $n \neq \nu^k_m$. Consequently, we apply $z^k_m \geq O^k_m(\hat{x}^k_{n=\nu^k_m}), \forall m \in \mathcal{M}, \forall k \in \mathcal{K}$ only for each possible optimization as follows: $\min\limits_{\{\mathbf{x}^*,\mathbf{z}\}} F(\mathbf{x}^*,\mathbf{z}) = \min\limits_{\{\mathbf{x}^*,\mathbf{z}\}} \sum_{k=1}^{K}\sum_{m=1}^{M} f^k_m \Big[\sum_{u=1}^{U_m}\Big(\Upsilon\big(\mathbf{x}^*\big) + \Psi\big(\mathbf{x}^*\big)z^k_m\Big)\Big]$, s.t. $z^k_m \geq O^k_m(\hat{x}^k_{n=\nu^k_m}), \forall m \in \mathcal{M}, \forall k \in \mathcal{K}$.

If $\mathbf{z}^* = \hat{z}^k_m$, $\forall m \in \mathcal{M}, \forall k \in \mathcal{K}$ is also the optimal solution, then the possible optimal $F(\mathbf{x}^*,\mathbf{z}^*)$ for each ν^k_m is $F(\mathbf{x}^*,\mathbf{z}^*) = \sum_{k=1}^{K}\sum_{m=1}^{M} f^k_m \Big[\sum_{u=1}^{U_m}\Big(\Upsilon\big(\mathbf{x}^*\big) + \Psi\big(\mathbf{x}^*\big)\hat{z}^k_m\Big)\Big]$, where $\hat{z}^k_m = O^k_m(\hat{x}^k_{n=\nu^k_m}), \forall m \in \mathcal{M}, \forall k \in \mathcal{K}$. Thus, we obtain:

$$F(\mathbf{x}^*) = \sum_{k=1}^{K}\sum_{m=1}^{M} f^k_m \left[\sum_{u=1}^{U_m}\Big(\Upsilon\big(\mathbf{x}^*\big) + \Psi\big(\mathbf{x}^*\big)O^k_m(\hat{x}^k_{n=\nu^k_m})\Big)\right]. \tag{24}$$

We can see that the selected MEN ν^k_m in $(\mathbf{Q_2})$ is the same as the selected MEN j^k_m in $(\mathbf{Q_1})$, $\forall m \in \mathcal{M}$ and $\forall k \in \mathcal{K}$. As a result, if $\mathbf{x}^*$ is the optimal solution of $(\mathbf{Q_1})$, it is also the optimal solution of $(\mathbf{Q_2})$ and vice versa.

References

[1] Cisco Mobile Visual Networking Index. Available Online: https://newsroom.cisco.com/press-release-content?articleId=1819296. Last Accessed on April 2019.

[2] J. Wen, *et al.*, "Cache-enabled heterogeneous cellular networks: optimal tier-level content placement," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5939-5952, Sept. 2017.

[3] K. Poularakis, *et al.*, "Caching and operator cooperation policies for layered video content delivery," in *IEEE INFOCOM*, Apr. 2016, pp. 1-9.

[4] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *IEEE INFOCOM*, Mar. 2010, pp. 1-9.

[5] K. Shanmugam, *et al.*, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402-8413, Dec. 2013.

[6] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382-1393, May 2017.

[7] M. Dehghan, *et al.*, "On the complexity of optimal routing and content caching in heterogeneous networks," in *IEEE INFOCOM*, Apr. 2015, pp. 936-944.

[8] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *ACM MOBIHOC*, Jun. 2015, pp. 127-136.

[9] Y. M. Saputra, *et al.*, "A novel mobile edge network architecture with joint caching-delivering and horizontal cooperation," arXiv:1810.10700 [cs.NI], Oct. 2018.

[10] S. Zhang, *et al.*, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791-1805, Aug. 2018.

[11] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 250-264, Jan. 2017.

[12] J. Liu, *et al.*, "Cache placement in fog-RANs: from centralized to distributed algorithms," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7039-7051, Nov. 2017.