LATENCY DRIVEN FRONTHAUL BANDWIDTH ALLOCATION AND COOPERATIVE BEAMFORMING FOR CACHE-ENABLED CLOUD-BASED SMALL CELL NETWORKS

Xiongwei Wu, Xiuhua Li†,**, Qiang Li‡, Victor C. M. Leung** and P. C. Ching**

*Dept. of Elec. Eng., The Chinese University of Hong Kong, Hong Kong SAR, China
†School of Big Data & Software Engineering, Chongqing University, Chongqing, China
‡School of Info. & Comm. Eng., University of Electronic Science and Technology of China, Chengdu, China
**Dept. of Elec. & Comp. Eng., The University of British Columbia, Vancouver, Canada

ABSTRACT

This paper considers content delivery of the cache-enabled small cell networks (C-SCNs), where users with the same request form a multicast group and are served by a cluster of small-cell base stations (SBSs) under the coordination of the central processor. The performance of such a coordination is severely limited by the fronthaul link, which may be saturated and degrade quality of service (QoS). To improve user QoS, we propose a latency driven scheme by jointly optimizing fronthaul bandwidth allocation, multicast beamforming, and BS clustering. Accordingly, with min-max fairness among multicast groups, a latency minimization problem is formulated under the constraints of fronthaul bandwidth and transmission power. The resultant problem is a mixed-integer nonlinear program, which is NP-hard. To address such a complex problem, a quadratic penalty-based algorithm is proposed by using a reformulation of binary constraint. Meanwhile, we present the necessary condition for an optimal solution, which shows that fronthaul bandwidth allocation is inherently adaptive to cached contents and patterns of BS cooperation. Finally, simulation results demonstrate that the proposed scheme can effectively reduce latency under different caching strategies.

***Index Terms*—** Fronthaul Bandwidth Allocation, Multicast Beamforming, BS Clustering, C-SCNs

1. INTRODUCTION

By connecting multiple small-cell base stations (SBSs) to the central processor (CP) via fronthaul, the cloud-based small cell networks (C-SCNs) enable centralized optimization for signal processing and resource allocation across multiple SBSs through the coordination of the CP [1]. However, the performance of such coordination relies highly on the fronthaul link. The exponential growth of mobile data traffic may cause the capacity-limited fronthaul link saturated, and thus degrade quality of service (QoS). To deal with these challenges, wireless caching has been proposed as a promising approach in recent years.

As reported in [2], mobile data traffic is generally dominated by a few popular contents. For instance, the top 10% of videos in YouTube approximately occupy 80% of total views [2]. In the cache-enabled C-SCNs, each SBS can fetch those frequently requested contents during off-peak sessions. When users make requests, the cached contents can be locally transmitted without duplicated transmissions in the fronthaul link. In particular, the cache-enabled C-SCNs need to address two fundamental problems, i.e., content placement and content delivery. Content placement decides which files are allowed to be cached in SBSs for a long timescale, while content delivery determines how the contents are transmitted in a short timescale given their caching status in local SBSs.

Preliminary caching studies mainly focused on performance metric optimization, such as network traffic, power consumption, and system cost [3–5]. Nevertheless, latency, which is an important metric for users' QoS, has not been well investigated in cache-enabled wireless networks. The authors in [6, 7] investigated caching strategies to achieve low latency. These works only focused on basic cell caching, and ignored BS cooperation. Although in [8], the authors investigated cooperative cell caching to reduce delay. However, physical-layer transmissions were not considered, such as multicast beamforming. To efficiently reduce download latency, the joint design of cloud-to-BS layer and physical-layer transmissions are needed. The authors in [9] examined beamforming design under a fixed pattern for BS cooperation. In our previous work [10], we proposed a cooperative transmission scheme to achieve low latency by jointly optimizing fronthaul traffic assignment and multicast beamforming. To our best knowledge, the vast majority of preliminary studies generally considered fixed fronthaul bandwidth (resource) for each multicast group in the cooperative transmission. As we mentioned previously, the capacity-limited fronthaul presents one of major challenges for cooperative transmission. Thus, fronthaul resource should also be reasonably scheduled for cooperative beamforming so as to provide high QoS for users.

In this paper, we focus on content delivery of the cache-enable C-SCNs. Users with the same content request form a multicast group and are served by a cluster of SBSs through multicast beamforming. We aim to reduce the latency of content delivery through edge and fronthaul links and guarantee fairness among multicast groups. To reduce edge latency, more SBSs should be involved in cooperative transmissions so as to provide higher spatial gain. To reduce fronthaul latency, more fronthaul bandwidth should be allocated to multicast groups with fewer cached contents in local SBSs. However, the SBSs should be reasonably clustered according to their cache resource and fronthaul bandwidth. For instance, when assigning SBSs without cache content to serve users, it may lead to extremely high fronthaul traffic load and bandwidth consumption. Therefore, we need to consider a joint design of fronthaul bandwidth allocation, multicast beamforming, and BS clustering. In particular, a latency minimization problem is formulated under transmission power and fronthaul bandwidth allocation constraints. The resulting problem is a mixed-integer nonlinear program (MINLP), which is hard to obtain an optimal solution. As a compromise, a quadratic penalty-based al-

gorithm is proposed to iteratively compute an approximate solution with convergence guarantee. Moreover, necessary condition for an optimal solution is also derived. Finally, the effectiveness of the proposed scheme is demonstrated by simulations.

2. SYSTEM MODEL

As depicted in Fig. 1, we consider the downlink transmission of the cloud-based small cell networks (C-SCNs), where a total of B SBSs are connected to the CP through wireless fronthaul links with finite capacity [12]. Meanwhile, K users are cooperatively served by a cluster of SBSs on the wireless channels, which is referred to as edge links. Each SBS has M antennas. We assume that the CP has access to the whole content library, which stores F files. For notational convenience, each file is assumed to have S bits. We index the files by the rank order of content popularity. The probability of file f being selected by users obeys the Zipf distribution $p(f) = cf^{-\gamma}$, where f denotes the rank order, $\gamma \geq 0$ is the skewness factor, and c is a normalization constant [3]. Let $\mathcal{B} = \{1, \ldots, B\}, \mathcal{F} = \{1, \ldots, F\}$, and $\mathcal{K} = \{1, \ldots, K\}$ be the sets of SBSs, files in the library, and users, respectively.

A cache-enabled system usually operates in two phases, i.e., content placement and content delivery. Specifically, for content placement, we use cache allocation matrix $\mathbf{M} = [m_{f,b}]$ to indicate that a certain fraction $0 \leq m_{f,b} \leq 1$ of file b is cached in SBS b. Consequently, we have $\sum_{f\in\mathcal{F}} m_{f,b} S \leq S_b$, where S_b is the caching storage in SBS b. Notably, the wireless channel usually varies much faster than content popularity distribution. Therefore, the cached contents usually remain unchanged in a long timescale, that is, $\mathbf{M}$ is prefixed and can be scheduled by many effective caching strategies [13]. In this paper, we focus on content delivery, which is detailed as follows.

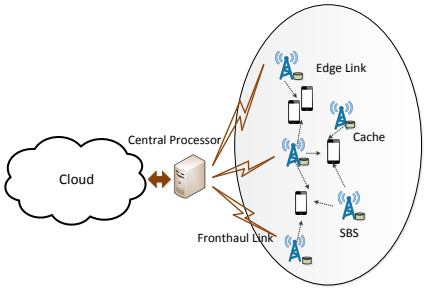

Fig. 1. An illustration of cache-enabled C-SCNs.

We assume system operates in a block manner, during which the wireless channel is quasi-static but varies from one to another [14]. In each transmission block, each user k requests a certain file f_k. The set $\mathcal{F}_{\text{req}} \subset \mathcal{F}$ denotes the indexes of all requested files with cardinality $F_{\text{req}} = |\mathcal{F}_{\text{req}}|$. Then, users are grouped according to their requests. Specifically, users in group f are denoted as $\mathcal{G}_f$ and only request file f. Accordingly, we define a BS clustering matrix $\mathbf{E} = [e_{f,b}] \in \{0,1\}^{B\times F_{\text{req}}}$, where the element $e_{f,b} = 1$ indicates that SBS b is selected to serve the multicast group f, otherwise 0. The signal transmitted from SBS b is $\mathbf{x}_b = \sum_{f\in\mathcal{F}_{\text{req}}} \mathbf{v}_{f,b} x_f$, where the signal $x_f \in \mathbb{C}$ independently encodes the information symbols for multicast group f, with distribution $x_f \sim \mathcal{N}(0,1)$, for $\forall f \in \mathcal{F}$; and $\mathbf{v}_{f,b} \in \mathbb{C}^M$ denotes the transmit beamformer for file b. Let $\mathbf{v}_f = \left[\mathbf{v}_{f,1}^H, \mathbf{v}_{f,2}^H, \cdots, \mathbf{v}_{f,B}^H\right]^H$ be the aggregate beamformer from all SBSs for delivering file f. Note that if SBS b is not selected to deliver file f, the corresponding transmit beamformer $\mathbf{v}_{f,b}$ should be $\mathbf{0}$. At the k-th user, the received signal is given by

$$y_k = \underbrace{\textstyle\sum_{b\in\mathcal{B}} \mathbf{h}_{kb}\mathbf{v}_{f_k,b}x_{f_k}}_{\text{desired signal}} + \underbrace{\textstyle\sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{b\in\mathcal{B}} \mathbf{h}_{kb}\mathbf{v}_{f,b}x_f}_{\text{inter-group interference}} + z_k,$$

where $\mathbf{h}_{kb} \in \mathbb{C}^M$ denotes the channel matrix between SBS b and user k; and z_k denotes the additive complex Gaussian noise with distribution $z_k \sim \mathcal{CN}(0, \sigma_k^2)$. We define $D_k = |\mathbf{h}_k\mathbf{v}_{f_k}|^2$, and $J_k = \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}} |\mathbf{h}_k\mathbf{v}_f|^2 + \sigma_k^2$, where $\mathbf{h}_k = [\mathbf{h}_{k1}, \mathbf{h}_{k2}, \cdots, \mathbf{h}_{kB}], \forall k$. By treating the interference as noise, the achievable data rate for multicast group f is given by $R_f = \min_{k\in\mathcal{G}_f} B_0\phi(D_k, J_k)$, where function $\phi(D_k, J_k) = \log(1 + D_k/J_k)$ and B_0 is the bandwidth.

When the requested files are not entirely cached in local SBSs, the missing contents should be fetched from the CP via fronthaul. To schedule fronthaul spectrum, we utilize frequency division multiplexing access technique. We consider that the total fronthaul capacity is C_F bps. Besides, fronthaul spectrum is orthogonal to that of the edge link, and thus no interference is heard between the fronthaul and edge links. Let $\mathbf{T} = [t_{f,b}]$ be bandwidth allocation matrix. Accordingly, a fraction $t_{f,b}$ of the fronthaul bandwidth is allocated for SBS b to serve multicast group f; and the associated fronthaul rate $R_{f,b}^{\text{fh}}$ is given by $R_{f,b}^{\text{fh}} = t_{f,b}C_F$.

We mainly deal with the latency arising from content delivery through the fronthaul and edge links, which is defined as the number of symbols or channel uses needed to complete content transmission [9, 10]. By considering BS clustering and multicast beamforming, the edge latency is given by $T_E = S/\min_{f\in\mathcal{F}_{\text{req}}} R_f$, and the fronthaul latency is given by $T_F = \max_{f\in\mathcal{F}_{\text{req}}, b\in\mathcal{B}} e_{f,b}m'_{f,b}/R_{f,b}^{\text{fh}}$, where $m'_{f,b} = (1 - m_{f,b})S$. We assume that message delivery is half-duplex, and the system operates in a serial mode. Hence, the total latency is given by

$$T_{\text{total}}(\mathcal{V}, \mathbf{E}, \mathbf{T}) = \max_{f\in\mathcal{F}_{\text{req}}} \max_{k\in\mathcal{G}_f} \frac{S}{B_0\phi(D_k, J_k)} + \max_{f\in\mathcal{F}_{\text{req}}, b\in\mathcal{B}} \frac{e_{f,b}m'_{f,b}}{t_{f,b}C_F + \tau_0}, \quad (1)$$

where $\tau_0 > 0$ is extremely small to avoid null denominator, considering some files may be fully cached in SBSs and $t_{f,b}$ being 0; and $\mathcal{V} = \{\mathbf{v}_f, \forall f \in \mathcal{F}_{\text{req}}\}$.

3. PROBLEM FORMULATION & PROPOSED DESIGN

Our goal is to minimize network latency for content delivery in the fronthaul and edge links. Notably, edge latency depends on multicast beamforming $\mathcal{V}$ and BS clustering $\mathbf{E}$, whereas fronthaul latency depends on BS clustering $\mathbf{E}$ and fronthaul bandwidth allocation $\mathbf{T}$. Thus, it is necessary to jointly optimize variables $\{\mathcal{V}, \mathbf{E}, \mathbf{T}\}$, which yields the following optimization problem,

$$\mathcal{P}_0: \quad \min_{\mathcal{V},\mathbf{E},\mathbf{T}} \quad T_{\text{total}}(\mathcal{V}, \mathbf{E}, \mathbf{T}) \quad (2a)$$

$$\text{s.t.} \ \textstyle\sum_{f\in\mathcal{F}_{\text{req}}} \|\mathbf{v}_{f,b}\|_2^2 \leq P_b, \ b \in \mathcal{B}, \quad (2b)$$

$$\|\mathbf{v}_{f,b}\|_2^2 \leq e_{f,b}P_b, \ \forall f, b \quad (2c)$$

$$\mathbf{E} = [e_{f,b}] \in \{0,1\}^{B\times F_{\text{req}}} \quad (2d)$$

$$\sum_{b\in\mathcal{B}} e_{f,b} \geq 1, \forall f, \tag{2e}$$
$$\sum_{f\in\mathcal{F}_{\text{req}}, b\in\mathcal{B}} t_{f,b} = 1, 0 \leq t_{f,b} \leq 1,\ \forall f, b, \tag{2f}$$

where constraint (2b) indicates the maximum transmit power P_b for SBS b. Constraint (2c) shows the coupling between beamformer design and BS clustering, i.e., when $e_{f,b} = 0$, the associated beamformer $\mathbf{v}_{f,b} = \mathbf{0}$. Constraint (2e) is used to avoid a standstill service and guarantee BS cooperation. Fronthaul bandwidth allocation for all multicast groups is provided by constraint (2f).

Problem $\mathcal{P}_0$ is MINLP and generally NP-hard. The challenges are due to binary constraints (2d) and non-convexity of objective function. We present the following necessary condition for an optimal solution to problem $\mathcal{P}_0$.

Proposition 1: *Let $\{\mathcal{V}^*, \mathbf{E}^*, \mathbf{T}^*\}$ be an optimal solution to $\mathcal{P}_0$. It holds true that*

$$t^*_{f,b} = \frac{\left\|\|\mathbf{v}^*_{f,b}\|_2\right\|_0 m'_{f,b}}{\|\text{vec}(\mathbf{S}^*)\|_1},\ \forall f \in \mathcal{F}_{\text{req}}, b \in \mathcal{B}, \tag{3}$$

where operator $\|\cdot\|_0$ denotes the number of non-zero elements in a vector, the load matrix $\mathbf{S}^* = [s^*_{f,b}]$, and $s^*_{f,b} = e^*_{f,b} m'_{f,b}$.

The proof is omitted due to page limit. Proposition 1 shows that the optimal bandwidth allocation is inherently adaptive to the cached resources $m_{f,b}$ in local SBSs and also depends on the design of beamformer and BS clustering. In addition, this Proposition can also be utilized for algorithm design.

To efficiently solve problem $\mathcal{P}_0$, a quadratic penalty-based algorithm is proposed. First of all, we reformulate problem $\mathcal{P}_0$ by introducing several slack variables t_E, t_F, r_k and the semidefinite relaxation (SDR) technique. By using the epigraph reformulation, problem $\mathcal{P}_0$ is rewritten as

$$\min_{\mathcal{V},\mathbf{E},\mathbf{T},t_E,t_F,r_k} \quad t_E + t_F \tag{4a}$$
$$\text{s.t.} \quad t_E \geq S/r_k, r_k \geq 0,\ \forall k, \tag{4b}$$
$$r_k \leq B_0 \log(D_k + J_k) - B_0 \log(J_k),\ \forall k, \tag{4c}$$
$$t_F \geq \frac{e_{f,b} m'_{f,b}}{t_{f,b} C_F + \tau_0},\ \forall f, b, \tag{4d}$$
$$(2b) - (2f), \tag{4e}$$

Note that constraint (4c) is nonconvex. We define $\mathcal{W}$ as $\{\mathbf{W}_f = \mathbf{v}_f \mathbf{v}_f^H, \forall f \in \mathcal{F}_{\text{req}}\}$. Thus, we have $D_k = \mathbf{h}_k \mathbf{W}_{f_k} \mathbf{h}_k^H$, $J_k = \chi_{k,1}(\mathcal{W})$, and $\chi_{k,1}(\mathcal{W}) = \sum_{f\neq f_k} \mathbf{h}_k \mathbf{W}_f \mathbf{h}_k^H + \sigma_k^2$. Let $\{\mathbf{L}_b, \forall b \in \mathcal{B}\}$ be a set of selection matrices, where $\mathbf{L}_b$ is a diagonal matrix and $\mathbf{L}_b = \text{diag}\left(\left[\mathbf{0}_{(b-1)M}, \mathbf{I}_M, \mathbf{0}_{(B-b)M}\right]\right)$. We further define functions $g_{k,1}(\mathcal{W}, r_k) = r_k - B_0 \log(D_k + J_k)$, and $g_{k,2}(\mathcal{W}) = -B_0 \log(J_k)$. Hence, constraint (4c) can be rewritten as

$$g_{k,1}(\mathcal{W}, r_k) - g_{k,2}(\mathcal{W}) \leq 0,\ \forall k, \tag{5}$$

where the left-hand side is in the form of difference-of-convex (DC) functions. To combat the discontinuity, constraint (2d) can be rewritten as

$$(e_{f,b} - 1) e_{f,b} = 0,\ \forall f, b, \tag{6}$$
$$0 \leq e_{f,b} \leq 1,\ \forall f, b, \tag{7}$$

where the equilibrium constraint (6) is still non-convex. We can construct a ***quadratic penalty*** function $h_1(\mathbf{E}) = \sum_{b\in\mathcal{B}, f\in\mathcal{F}_{\text{req}}} (e^2_{f,b} - e_{f,b})$ to penalize the violation of the equilibrium constraint (6).

We denote sets $\mathbf{r} = \{r_k, \forall k \in \mathcal{K}\}$ and $\Theta = \{\mathcal{V}, \mathbf{E}, \mathbf{T}, \mathbf{r}, t_E, t_F\}$. As a result, by removing the rank constraint $\text{rank}(\mathbf{W}_f) = 1$, problem (4) can be relaxed as

$$\mathcal{R}_0: \quad \min_{\Theta} \ t_E + t_F - \lambda h_1(\mathbf{E}) \tag{8a}$$
$$\text{s.t.} \ \sum_{f\in\mathcal{F}_{\text{req}}} \text{tr}\{\mathbf{L}_b \mathbf{W}_f\} \leq P_b,\ \forall b, \tag{8b}$$
$$\text{tr}\{\mathbf{L}_b \mathbf{W}_f\} \leq e_{f,b} P_b,\ \forall f, b, \tag{8c}$$
$$\mathbf{W}_f \succeq \mathbf{0},\ \forall f, \tag{8d}$$
$$(2e), (2f), (4b), (4d), (5), (7), \tag{8e}$$

which is a continuous programming with penalty parameter $\lambda > 0$. Notably, $e_{f,b} - e^2_{f,b} \geq 0$, and this relation holds for any feasible point $e_{f,b}$ within constraint (7). In particular, the optimal solution $\mathbf{E}^*$ meets the condition $e^*_{f,b}(1 - e^*_{f,b}) = 0$, and hence $h(\mathbf{E}^*) = 0$. Accordingly, one can always penalize $e_{f,b} - e^2_{f,b} = 0$ by gradually lifting the penalty parameter λ, which results in a binary solution.

Problem $\mathcal{R}_0$ is still nonconvex. We observe that the objective function of $\mathcal{R}_0$ and constraint (5) have the structure of DC. To take advantage of this property, we propose a penalty-based algorithm by applying the convex and concave procedure (CCCP) and inexact block coordinate update (BCU) to efficiently solve problem $\mathcal{R}_0$.

We first focus on the optimization for block variable $\{\mathcal{W}, \mathbf{E}\}$ and fix $\mathbf{T} = \overline{\mathbf{T}}$. Specifically, functions $h_1(\mathbf{E})$ and $g_{k,2}(\overline{\mathcal{U}}, \mathcal{W})$ can be lower bounded by their first-order Taylor expansions at certain local points $\mathcal{W}^{(i)} = \{\mathbf{W}_f^{(i)}, \forall f\}$ and $\mathbf{E}^{(i)} = \left[e_{f,b}^{(i)}\right]$, respectively. Thus, we have $\widetilde{h}_1(\mathbf{E}|\mathbf{E}^{(i)}) = \sum_{f,b} -{e_{f,b}^{(i)}}^2 + (2e_{f,b}^{(i)} - 1) e_{f,b}$, and

$$\widetilde{g}_{k,2}(\mathcal{W}|\mathcal{W}^{(i)}) = g_{k,2}(\mathcal{W}^{(i)}) - B_0 (\overline{J}_k^{(i)})^{-1} \left(\overline{J}_k - \overline{J}_k^{(i)}\right), \tag{9}$$

where $\overline{J}_k = \chi_{k,1}(\mathcal{W})$ and $\overline{J}_k^{(i)} = \chi_{k,1}(\mathcal{W}^{(i)})$. Subsequently, in the i-th iteration, by applying the CCCP technique, problem $\mathcal{R}_0$ is solved by the following inner approximation problem.

$$\widetilde{\mathcal{R}}_1(\overline{\mathbf{T}}, \mathcal{W}^{(i)}, \mathbf{E}^{(i)}): \min_{\mathcal{W},\mathbf{E},t_E,t_F,r_k} \ t_E + t_F - \lambda \widetilde{h}_1(\mathbf{E}|\mathbf{E}^{(i)})$$
$$\text{s.t.} \ g_{k,1}(\mathcal{W}, r_k) - \widetilde{g}_{k,2}(\mathcal{W}|\mathcal{W}^{(i)}) \leq 0,\ \forall k,$$
$$(2e), (2f), (4b), (4d), (7), (8b) - (8d),$$

which can be efficiently solved by interior point methods using standard solver, such as CVX [15]. The optimal solution is denoted as $\mathcal{W}^{(i+1)}$ and $\mathbf{E}^{(i+1)}$.

On the other hand, when $\{\mathcal{W}, \mathbf{E}\}$ are fixed as $\overline{\mathbf{E}} = \left[\overline{e}_{f,b}\right]$, $\overline{\mathcal{W}} = \{\overline{\mathbf{W}}_f, \forall f\}$, the other block variable $\mathcal{T}$ can be updated by Proposition 1 with the following closed-form expression,

$$t_{f,b}^{(i+1)} = \overline{e}_{f,b} m'_{f,b} / \|\text{vec}(\mathbf{S})\|_1,\ \forall f \in \mathcal{F}_{\text{req}}, b \in \mathcal{B}, \tag{10}$$

where matrix $\mathbf{S} = [s_{f,b}]$ and $s_{f,b} = \overline{e}_{f,b} m'_{f,b}$.

Finally, the whole process for the inexact BCU-CCCP design is presented in Algorithm 1. To find a suitable starting point, the penalty parameter $\lambda \geq 0$ is generally initialized as a small value. Since problem $\mathcal{R}_0$ relaxes the rank-one constraint on $\mathbf{W}_f$, the resultant $\mathbf{W}_f$ may have high rank. In such a case, we can simply apply the Gaussian randomization to obtain an approximate solution [16]. When $\eta > 1$, the parameter λ is lifted gradually to enforce $h_1(\mathbf{E}) = 0$. To avoid numerical problems, the penalty parameter usually does not increase until it reaches a sufficiently large value $\lambda_{\max}$. Notably, the optimal value of $\widetilde{\mathcal{R}}_1$ serves as an upper bound of that of problem $\mathcal{R}_0$ due to the inner approximation. When

Algorithm 1 Inexact BCU-CCCP Design

1: **Initialize** $i = 0$, $\mathcal{W}^{(0)}, \mathbf{E}^{(0)}, \mathbf{T}^{(0)}, \lambda > 0, \eta > 1, \lambda_{\max}$
2: **repeat**
3: Solve problem $\widetilde{\mathcal{R}}_1(\mathbf{T}^{(i)}, \mathcal{W}^{(i)}, \mathbf{E}^{(i)})$ to obtain an optimal solution $\mathcal{W}^{(i+1)}, \mathbf{E}^{(i+1)}$
4: Update $\mathcal{T}^{(i+1)}$ by closed-form expression (10)
5: Update $\lambda \leftarrow \min\{\lambda\eta, \lambda_{\max}\}$
6: $i \leftarrow i + 1$
7: **until** some stopping criterion is satisfied

$\lambda = \lambda_{\max}$, the sequence $\{\mathbf{T}^{(i)}, \mathcal{W}^{(i)}, \mathbf{E}^{(i)}\}$ generated by Algorithm 1 provides a monotonically non-increasing objective value $\{\mathcal{R}_0^{(i)}\}$, which can converge [17].

4. PERFORMANCE EVALUATION & CONCLUSION

Numerical experiments are provided to evaluate the performance of the proposed scheme. We consider that the cache-enabled SCNs covers a square area $[1\text{km}, 1\text{km}] \times [1\text{km}, 1\text{km}]$, in which B SBSs and K active users are randomly and uniformly distributed in this region. The wireless channel between each user and each SBS is modeled as large-scale fading and small-scale fading. Specifically, the pass-loss (dB) is $36.8 + 36.7\log(d)$, where d is the distance in meter; the log-normal shadowing parameter is 7 dB and antenna gain is 5 dBi; the small-scale fading is Rayleigh fading with unit variance. The bandwidth for edge link is 10 MHz. The noise power is assumed to be -102 dBm. Consider the following default scenario: 3 SBSs and 8 active users, each SBS has 5 antennas and the same fractional caching capacity (S_b/FS) 20%, the maximum transmit power is 30 dBm, the fronthaul capacity is 10 Mbps, if not specified otherwise. In each simulation trial, users make requests from the library, which stores 100 files and the skewness factor of the Zipf distribution is 0.5. The size of each file is assumed to be 100 MB. All of the results are obtained by averaging over 100 independent simulation trials.

As a comparison, two caching strategies are considered. (i) Uniform Caching (UC): each SBS randomly and uniformly caches a certain fraction m of each file in the library, i.e., $m_{f,b} = m, \forall f, b$. (ii) Probabilistic Caching (ProbC): each SBS caches several contents according to the Zipf distribution until it reaches the maximum storage. To evaluate the potentials of joint fronthaul bandwidth allocation, BS clustering and beamforming, we also implement the Coordinate Beamforming (CB) scheme: all users are served by SBSs through full cooperation, i.e. $e_{f,b} = 1, \forall f, b$, and fronthaul bandwidth allocation is fixed and given by condition (3) without loss of optimality.

First, we illustrate the convergence behavior of the inexact BCU-CCCP design in Fig. 2. For comparison, we set $e_{f,b}^{(0)} = 0.1, \forall f, b$; the inital penalty parameter is 1, which is increased by $\eta = 5$ in each iteration; and $\lambda_{\max} = 3000$. The ProbC strategy is adopted. Three independent trials are considered, and the beamformer is randomly generated in each trial. We observe that the objective values increase within the first five iterations, due to the lift of the penalty parameter. When λ reaches the maximum value, the objective values decrease gradually and almost converge to the same value. This observation demonstrates that the proposed algorithm converges fast and is robust to the initial beamformer.

In Fig. 3, we investigate the impact of the fractional caching capacity on the average network latency. As it can be observed that the achieved latencies gradually decrease as the storage in each SBS increases. This is because when more contents are cached in local SBSs, these contents can be directly delivered via SBSs without fronthaul transmission. The proposed scheme outperforms than the CB scheme under two caching strategies. Specifically, the latencies achieved by the inexact BCU-CCCP design are less than 47.3% and 33.7% of that of CB under ProbC and UC strategy over the whole horizontal axis, respectively. In addition, the latencies achieved under ProbC strategy are less than those under UC strategy, because the former one is based on the distribution of content popularity. Fig. 4 presents the impact of fronthaul capacity on the average network latency. Under the condition that the fronthaul capacity is limited ($C_F \leq 10$ Mbps), the average latencies decrease dramatically. This fact indicates that the fronthaul latency dominates system performance. In particular, when $C_F = 5$ Mbps, the proposed scheme achieves 49.5% and 36.4% less latencies than that of CB scheme under two caching strategies. This finding implies that a reasonable fronthaul allocation can help to substantially reduce latency. Notably, the latency only witnesses a slight decrease after the fronthaul capacity is larger than 10 Mbps. This observation indicates that the edge latency starts to dominate system performance. All of the above results demonstrate superior performance of the proposed scheme.

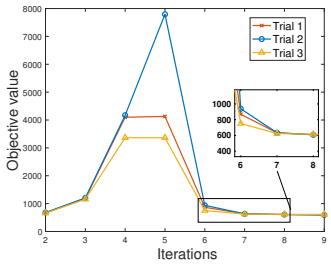

Fig. 2. Convergence behavior of inexact BCU-CCCP design

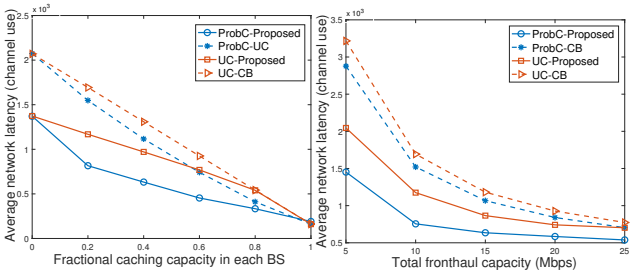

Fig. 3. Average network latency versus the fractional caching capacity in each SBS.

Fig. 4. Average network latency versus the total fronthaul capacity.

To conclude, in this paper, we have developed a latency driven transmission scheme in cache-enabled C-SCNs by exploiting fronthaul bandwidth allocation, multicast beamforming and BS clustering. We have formulated the problem to minimize delivery latency under the constraints of fronthaul bandwidth and transmission power. To solve the resulting mixed-integer nonlinear program, a penalty-based algorithm has been proposed. Moreover, we have presented the necessary condition for an optimal solution. Simulation results have revealed superior performance of the proposed scheme.

5. REFERENCES

[1] X. Li, X. Wang, Z. Sheng, H. Zhou, and V. C. Leung, "Resource allocation for cache-enabled cloud-based small cell networks," *Computer Communications*, vol. 127, pp. 20–29, Sept. 2018.

[2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, Oct. 2007, pp. 1–14.

[3] J. Liao, K.-K. Wong, Y. Zhang, Z. Zheng, and K. Yang, "Coding, multicast, and cooperation for cache-enabled heterogeneous small cell networks," *IEEE Trans Wireless Commun*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.

[4] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans Wireless Commun*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[5] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans Wireless Commun*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[6] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *INFOCOM, 2012 Proceedings IEEE*. Sept. 2012, pp. 1107–1115.

[7] H. Hsu and K.-C. Chen, "A resource allocation perspective on caching to achieve low latency," *IEEE Commun Lett*, vol. 20, no. 1, pp. 145–148, Jan. 2016.

[8] X. Li, X. Wang, S. Xiao, and V. C. Leung, "Delay performance analysis of cooperative cell caching in future mobile networks," in *Communications (ICC), 2015 IEEE International Conference on*. June 2015, pp. 5652–5657.

[9] S.-H. Park, O. Simeone, and S. Shamai, "Joint cloud and edge processing for latency minimization in Fog radio access networks," in *Signal Processing Advances in Wireless Communications (SPAWC), 2016 IEEE 17th International Workshop on*. July 2016, pp. 1–5.

[10] X. Wu and P. Ching, "Content delivery design for cache-aided cloud radio access network to achieve low latency," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2018, pp. 3749–3753.

[11] C. H. Jun Shi and B. Hu, "Delay optimal beamformer design for cache-enabled wireless backhaul networks," in *Communications (ICC), 2018 IEEE International Conference on*. July 2018, pp. 1–6.

[12] B. Hu, C. Hua, J. Zhang, C. Chen, and X. Guan, "Joint fronthaul multicast beamforming and user-centric clustering in downlink C-RANs," *IEEE Trans Wireless Commun*, vol. 16, no. 8, pp. 5395–5409, Jane 2017.

[13] S.-H. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for Fog radio access networks," *IEEE Trans Wireless Commun*, vol. 15, no. 11, pp. 7621–7632, Sept. 2016.

[14] R. Sun, Y. Wang, N. Cheng, H. Zhou, and X. Shen, "QoE driven BS clustering and multicast beamforming in cache-enabled C-RANs," in *2018 IEEE International Conference on Communications (ICC)*. May, 2018, pp. 1–6.

[15] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," Dec. 2017.

[16] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process Mag*, vol. 27, no. 3, pp. 20–34, 2010.

[17] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, June 2016.