CONTENT DELIVERY DESIGN FOR CACHE-AIDED CLOUD RADIO ACCESS NETWORK TO ACHIEVE LOW LATENCY

Xiongwei Wu, P. C. Ching

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong S.A.R. of China

ABSTRACT

In this paper, we examine the content delivery design for a cache-aided cloud radio access network (CA-CRAN), where users are served by multiple base stations (BSs) that are connected to cloud processor via fronthaul link. We propose a unified framework for cooperative delivery, which aims to minimize the total latency in the network. With fairness among users and physical-layer transmission, beamformers and content assignment are jointly optimized to fully exploit the benefits of caching. To address the resulting mixed binary nonconvex problem, a successive convex approximation (SCA)-based algorithm is derived with low complexity. Through simulations, the proposed design reduces latency significantly compared with existing work.

Index Terms— Beamforming, Cache, Content assignment, Latency

1. INTRODUCTION

Rapid growth of mobile data traffic, which is often dominated by some popular multimedia contents, has introduced a heavy burden for the limited backhaul link and degraded the quality of service (QoS) of users, particularly causing long delays. As a result, caching some popular content at BSs during the off-peak time is a promising technique to alleviate the traffic load and improve the QoS [1]. In this way, frequently-requested contents are able to be transferred to users from the local base station (BS) directly without producing extra latency and burden on the backhaul link.

Cache-aided cloud radio access network (CA-CRAN) is an important candidate for the cache-aided system. Benefiting from the cloud processor, it enables cooperative BS transmission and centralized interference management by jointly allocating physical layer resources, i.e., caching content, power and so on. Thus, CA-CRAN can boost the throughput, allow for low latency, and reduce the system cost [2]. So far, some studies on the design of the CA-CRAN have been conducted, and mainly consider two perspectives: (i) content placement and (ii) content delivery. For example, the network-wide cost and system throughput are investigated in [3, 4] and [5].

Nevertheless, the core issue of latency has not been fully addressed. Most related works focus on the content placement to achieve low latency. In [6], the content placement strategies are designed by reducing the content size in network traffic, without considering the physical-layer transmission such as beamforming. Although in [7], it designs caching policies by coupling the physical-layer transmission, it only minimizes the average delay of all users and ignores fairness. For content delivery design, many studies only focus on the information theoretical model [2, 8]. These works intend to develop coding strategies and capture the achievable latency simply with respect to an ideal interference-free system. Some other references, such as [9, 10], just deal with delivery design for a priori content assignment. However, the content assignment is a non-trivial task, especially when the collaborative BS beamforming is considered.

In this paper, we emphasize on the content delivery design for CA-CRAN. Motivated by the practical scenarios mentioned, we propose a unified framework for cooperative content delivery, which aims to minimize the total latency with consideration of fairness among users. To alleviate fronthaul link burden and encourage cooperative transmission, we jointly process the design of beamformers and content assignment. The problem is formulated as a mixed binary nonconvex programming. To combat this difficulty, a successive convex approximation (SCA)-based algorithm with exponential penalty is derived with low complexity. The effectiveness of the proposed algorithm is demonstrated by simulations.

Notations: Define $\mathbf{X}^H, \mathbf{X}^{-1}, \text{tr}\{\mathbf{X}\}, \text{Re}\{\mathbf{X}\}$ as the Hermitian, inverse, trace, and real value part of matrix $\mathbf{X}$. The trace of $\mathbf{A}\mathbf{B}^H$ is denoted as $\langle \mathbf{A}, \mathbf{B} \rangle$. $[\mathbf{A}; \mathbf{B}]$ represents vertical concatenation of matrix $\mathbf{A}, \mathbf{B}$.

2. NETWORK MODEL AND ASSUMPTIONS

As shown in Fig. 1, we consider the downlink transmission of a $J \times K$ CA-CRAN, where K users are cooperatively served by a cluster of J densely deployed BSs through the wireless channels, referred to as edge link. Each BS $i \in \mathcal{J}$ is connected to the central processor (CP) through a wired fronthaul link with limited capacity C_F. In this system, each user $k \in \mathcal{K}$ requests a file from the library of F files, where each file $f \in \mathcal{F}$ is assumed to have equal-sized S bits. We assume that the CP has access to the entire library. We label the files by the order of popularity, and the probability $P(f)$ of file f being selected is given by Zipf distribution $p(f) = cf^{-\gamma}$ where $\gamma \geq 0$ is a given popularity exponent, and c is set for normalization [3, 9]. Each BS has a local cache with a storage of μFS bits, where $\mu \in [0, 1]$ is the fractional caching capacity. A cache-aided system usually operates in two phases, i.e., content placement phase and content delivery phase. In the content placement phase, BSs cache a fraction of total content at the off-peaking time. Each file is split into L subfiles $(f, 1), (f, 2), \cdots, (f, L)$, which are mutually exclusive [9, 11]. Define set $\mathcal{L} = \{1, 2, \cdots, L\}$. Accordingly, we use the binary variable $c^i_{f,l}$ to indicate the subfile (f, l) is cached in i-th BS once $c^i_{f,l} = 1$, otherwise 0.

We only focus on content delivery design, with the knowledge of cached files in all BSs, i.e., $c^i_{f,l}$ being known a priori. The delivery phase is defined as follows. Initially, each user requests arbitrary file f_k . The set $\mathcal{F}_{\text{req}} = \{f_1, f_2, \cdots, f_K\}$ denotes the requested files of all users, which is a subset of $\mathcal{F}$. Users requesting the same files are grouped together, as the model in [3]. For simplicity, we assume that

only one user is in each group, so we have $f_{k_1} \neq f_{k_2}$ for $\forall k_1 \neq k_2$.

To minimize the network latency, the content assignment and beamformers are optimized in CP. The content assignment is characterized by determining which subfile (f,l) should be transferred through BSs. If the subfile (f,l) is not cached in the i-th BS, it should be accessed via the fronthaul link. The assignment is completed by setting the binary variables $e^i_{f,l} = 1$ if subfile (f,l) is transferred by the i-th BS, otherwise 0. Define the variable $d^i_{f,l}$ as $d^i_{f,l} = (1 - c^i_{f,l})e^i_{f,l}$ to indicate that subfile (f,l) is accessed via fronthaul link at the i-th BS when $d^i_{f,l} = 1$. We can observe that when $c^i_{f,l} = 1$, subfile (f,l) can directly be sent from the i-th BS without fronthaul latency. Thus, the size of the content transferred via the fronthanl link to the i-th BS is given by $S^i_F = \sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} d^i_{f,l}S_L$, where each subfile is assumed to have the same size of S_L bits. The signal transferred from i-th BS is given by

$$\mathbf{x}_i = \sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} \mathbf{V}^i_{f,l}\mathbf{s}_{f,l}, \tag{1}$$

where $\mathbf{V}_i{}^{f,l} \in \mathbb{C}^{M\times d}$ is the precoding matrix for the signal $\mathbf{s}_{f,l} \in \mathbb{C}^d$ that encodes the subfile (f,l). It is distributed as $\mathbf{s}_{f,l} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. Denote the network-wide beamformers that precode subfile (f,l) from all BSs as $\mathbf{V}_{f,l} = [\mathbf{V}^1_{f,l}; \mathbf{V}^2_{f,l}; \cdots; \mathbf{V}^J_{f,l}]$. Each subfile is independently coded, and user can reconstruct the file by receiving all subfiles [12]. Note that if subfile (f,l) is not served by i-th BS, the corresponding beamformer $\mathbf{V}^i_{f,l}$ should be $\mathbf{0}$. The received signal at the k-th user is given by

$$\mathbf{y}_k = \sum_{l\in\mathcal{L}} \mathbf{H}_k\mathbf{V}_{f_k,l}\mathbf{s}_{f_k,l} + \sum_{f\neq f_k}\sum_{l\in\mathcal{L}} \mathbf{H}_k\mathbf{V}_{f,l}\mathbf{s}_{f,l} + \mathbf{z}_k, \tag{2}$$

where $\mathbf{H}_k = [\mathbf{H}_{k1}, \mathbf{H}_{k2}, \cdots, \mathbf{H}_{kJ}]$, each $\mathbf{H}_{ki} \in \mathbb{C}^{N\times M}$ denotes the channel matrix between the i-th BS and the k-th user, $\mathbf{z}_k$ denotes the additive complex Gaussian noise with distribution $\mathbf{z_k} \sim \mathcal{CN}(\mathbf{0},\sigma^2_k\mathbf{I})$. For notation simplicity, define the signal matrix $\mathbf{S}_k$ and covariance matrix $\mathbf{J}_k$ as

$$\mathbf{S}_k = [\mathbf{H}_k\mathbf{V}_{f_k,1}, \mathbf{H}_k\mathbf{V}_{f_k,2}, \cdots, \mathbf{H}_k\mathbf{V}_{f_k,L}], \tag{3}$$

$$\mathbf{J}_k = \sum_{f\in\mathcal{F}_{\text{req}}\setminus\{f_k\}}\sum_{l\in\mathcal{L}} \mathbf{H}_k\mathbf{V}_{f,l}\mathbf{V}^H_{f,l}\mathbf{H}^H_k + \sigma^2_k\mathbf{I}. \tag{4}$$

Assuming that the receiver regards the interference as noise, the achievable rate of the k-th user to decode the total file f_k can be given by $R_k = B\phi(\mathbf{S}_k,\mathbf{J}_k)$, where the function $\phi(\mathbf{S}_k,\mathbf{J}_k) = \log\det\left(\mathbf{I} + \mathbf{S}^H_k\mathbf{J}^{-1}_k\mathbf{S}_k\right)$ [13] and B is system bandwidth. Denote the data rates of all users as the set $\left\{\mathcal{R}|R_k, k\in\mathcal{K}\right\}$.

Let us define the latency in the total network T_{total}, which shows the number of symbols or channel uses that are needed to accomplish the requested files transmission [10, 14]. We assume that the information delivery is half-duplex [15]; thus, the system operates in a serial manner. That is to say, the CP first communicates with BSs and is then followed by the wireless transmission stage. Considering the fairness [21], we evaluate the edge latency as

$$T_E = \frac{S}{\min_{k\in\mathcal{K}} R_k}, \tag{5}$$

where minimization is over the rates of all users. For the non-cached content transferred via fronthaul link, the latency in this process is given by

$$T_F = \frac{\max_{i\in\mathcal{J}} S^i_F}{C_F}, \tag{6}$$

so that the total latency of the network T_{total} is defined as

$$T_{\text{total}} = T_E + T_F. \tag{7}$$

Similar assumptions have been used in previous studies [7,10,15]. In our work, we emphasize the latency caused by the data transmission in the network, instead of the latency due to the geometry propagation, network coding, and other factors.

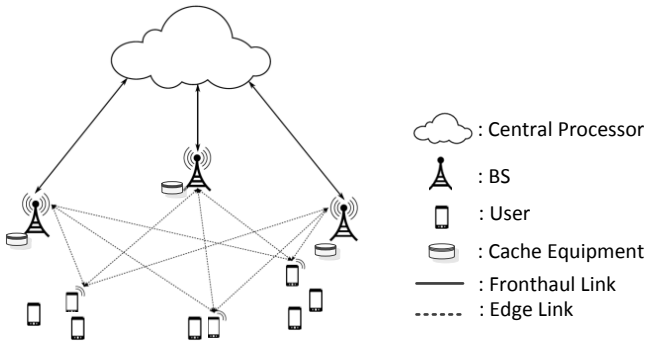

Fig. 1. An example of CA-CRAN downlink

3. PROBLEM FORMULATION

In this section, we propose a unified framework for content delivery design, which aims to minimize the total latency T_{total} in the network via BSs cooperative transmission. We perform the joint content assignment and beamforming at the CP. In the considered CA-CRAN, all CSI, local content in BSs, and knowledge of requested files are available at the CP. Define set $\mathcal{E} = \left\{e^i_{f,l} | f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}, i\in\mathcal{J}\right\}$ and set $\mathcal{V} = \left\{\mathbf{V}^i_{f,l} | f\in\mathcal{F}_{\text{req}}, l\in\mathcal{L}, i\in\mathcal{J}\right\}$; thus, the problem is stated as

$$\mathcal{P}_0 : \min_{T_E,T_F,\mathcal{E},\mathcal{V}} \; T_E + T_F \tag{8a}$$

$$\text{s.t. } T_F \geq \frac{S^i_F}{C_F}, i\in\mathcal{J} \tag{8b}$$

$$T_E \geq \frac{S}{R_k}, k\in\mathcal{K} \tag{8c}$$

$$\sum_{f\in\mathcal{F}_{\text{req}}}\sum_{l\in\mathcal{L}} \text{tr}\left\{\mathbf{V}_{f,l}\mathbf{V}^H_{f,l}\right\} \leq P_0 \tag{8d}$$

$$\sum_{i\in\mathcal{J}} e^i_{f,l} \geq 1, \; \forall\, e^i_{f,l} \in \mathcal{E} \tag{8e}$$

$$\left(1 - e^i_{f,l}\right)\mathbf{V}^i_{f,l} = \mathbf{0}, \; \forall\, e^i_{f,l}\in\mathcal{E}, \mathbf{V}^i_{f,l}\in\mathcal{V} \tag{8f}$$

$$e^i_{f,l} \in \{0,1\}, \; \forall\, e^i_{f,l}\in\mathcal{E}. \tag{8g}$$

where constraints (8b), (8c) are equivalent to the latency defined previously, once the optimality is attained. The constraint (8d) indicates that the total transmit power for all BSs is limited by P_0. Constraints (8e)–(8g) mainly account for the determination of assignment $e^i_{f,l}$. In particular, to alleviate fronthaul link burden and encourage cooperative transmission, each file is likely to be accessed via multiple BSs. This is achieved by constraint (8e). The constraint (8f) implies that if the i-th BS is not assigned to transfer subfile (f,l), i.e. $e^i_{f,l} = 0$, the related beamformer $\mathbf{V}^i_{f,l}$ must be $\mathbf{0}$.

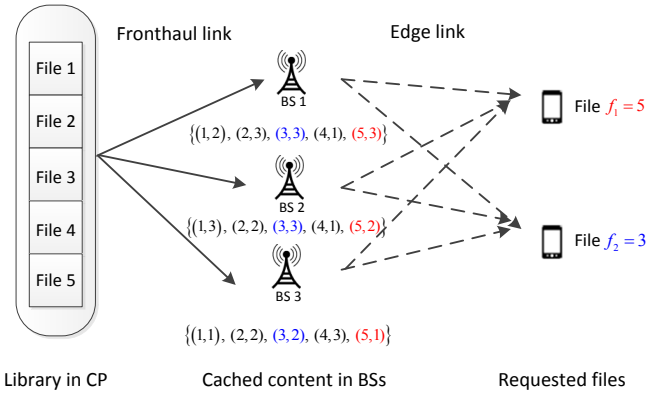

Fig. 2. An example of proposed design.

To explain the proposed design in greater detail, Fig. 2 illustrates a simple example. Consider a CP with a library storing 5 most popular files. Each file is split into $L = 3$ subfiles. During off-peak times, it performs the content placement phase such that BSs prefetch some popular contents that users will request potentially. As a result, requested file 3 is divided into subfiles $(3,1), (3,2), (3,3)$, where subfile $(3,2)$ is available in BS 3 and subfile $(3,3)$ is available in BS 1 and 2. In such case, subfiles $(3,2), (3,3)$ can be directly accessed from BSs without producing the fronthual latency. The assignment of BSs serving the transmission of subfile (3,3) will be carefully selected by the proposed design, i.e., with the consideration of the channel gain. As for subfile $(3,1)$, it should be firstly fetched from the CP to certain BSs via fronthual link. With consideration of fairness, the one with the least fronthual traffic load is likely to be assigned for transferring subfile $(3,1)$. In terms of requested file 5, benefited from BS cooperation, all related subfiles can be accessed through local BSs. Hence the leverage of local cache can significantly reduce the latency and alleviate fronthaul traffic.

Proposition 1: If the file f is requested by users and the subfile (f,l) is available in i-th BS, i.e., $c^i_{f,l} = 1$, we can set variable $e^i_{f,l} = 1$ for problem $\mathcal{P}_0$ without loss of optimality.

Proposition 1 also reveals insights into the proposed design. When allocating a BS that has cached the requested content already to the cluster of BSs serving this content, it always produces no extra fronthaul latency but potentially a decrease of edge latency. The proof is similar to what is presented in [3].

4. PROPOSED ALGORITHM

Note that problem (8) is a nonconvex problem with the binary assignment mixed. While using the exhaustive search to find all possibilities of binary variables, the computational complexity grows exponentially with $\mathcal{O}(2^{MNL-L_a})$, where L_a is the number of content copies in BSs. A greedy approach can be derived based on the idea in [16] with computational complexity $\mathcal{O}((MNL-L_a)^2)$, which is still excessively high because, in each stage, it needs to solve a nonconvex problem. To address this difficulty, an SCA-based algorithm with exponential penalty is derived with low complexity.

First, equivalently transform the constraint (8f) as

$$\left\|\mathbf{V}^i_{f,l}\right\|^2_F \le \left(e^i_{f,l}\right)^\alpha P_0, \tag{9}$$

with $\alpha \ge 1$ because $e^i_{f,l}$ is a binary term. After we relax the binary variable $e^i_{f,l}$ into a continuous one, noted as problem $\mathcal{P}^c_0$, $\alpha \ge 1$ is served as a penalty to encourage toward a binary solution in such continuous relaxation [17]. Nevertheless, (9) is still nonconvex. Observe that $(e^i_{f,l})^\alpha$ is lower bounded by the first-order Taylor expansion $\theta^{(r)}(e^i_{f,l})$ within $0 \le e^i_{f,l} \le 1$. In particular, we have

$$\theta^{(r)}\left(e^i_{f,l}\right) = (1-\alpha)(e^{i}_{f,l}{}^{(r)})^\alpha + \alpha(e^{i}_{f,l}{}^{(r)})^{(\alpha-1)} e^i_{f,l}, \tag{10}$$

where the local point $0 \le e^{i}_{f,l}{}^{(r)} \le 1$.

Lemma 1: Consider the function $\phi(\mathbf{A},\mathbf{B})$, where matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$ and positive definite matrix $\mathbf{B} \in \mathbb{C}^{n\times n}$. The following quadratic function is minorant of $\phi(\mathbf{A},\mathbf{B})$ at $(\overline{\mathbf{A}},\overline{\mathbf{B}})$:

$$\overline{\phi}(\mathbf{A},\mathbf{B}) = q + 2\,\mathrm{Re}\left\{\langle \mathbf{Q}_1, \mathbf{A}\rangle\right\} - \left\langle \mathbf{Q}_2, \mathbf{A}\mathbf{A}^H + \mathbf{B}\right\rangle, \tag{11}$$

where constant $q = \phi(\overline{\mathbf{A}},\overline{\mathbf{B}}) - \langle \overline{\mathbf{A}}, \overline{\mathbf{B}}^{-1}\overline{\mathbf{A}}\rangle$, $\mathbf{Q}_1 = \overline{\mathbf{B}}^{-1}\overline{\mathbf{A}}$ and $\mathbf{Q}_2 = \overline{\mathbf{B}}^{-1} - (\overline{\mathbf{B}} + \overline{\mathbf{A}}\,\overline{\mathbf{A}}^H)^{-1} \succeq \mathbf{0}$. It shows that $\overline{\phi}(\mathbf{A},\mathbf{B})$ is a lower bound of $\phi(\mathbf{A},\mathbf{B})$, and tight at $(\overline{\mathbf{A}},\overline{\mathbf{B}})$.

The proof of Lemma 1 is provided in [19]. Applying Lemma 1, for any feasible $\mathbf{V}^{i}_{f,l}{}^{(r)}$ for problem (8), we have such inequality $\overline{\phi}^{(r)}(\mathbf{S}_k,\mathbf{J}_k) \le \phi(\mathbf{S}_k,\mathbf{J}_k)$. Thus, $B\overline{\phi}^{(r)}(\mathbf{S}_k,\mathbf{J}_k)$ is minorant of the rate R_k at $(\mathbf{S}^{(r)}_k,\mathbf{J}^{(r)}_k)$, and minorant $\overline{\phi}^{(r)}$ is a concave quadratic function versus beamformer $\mathbf{V}_{f,l}$. Ultimately, we tackle problem $\mathcal{P}_0$ by successively solving the following relaxation problem:

$$\mathcal{P}^{(r)}_1 : \min_{T_E,T_F,\mathcal{E},\mathcal{R},\mathcal{V}} \quad T_E + T_F \tag{12a}$$

$$\text{s.t. } R_k \le B\overline{\phi}^{(r)}(\mathbf{S}_k,\mathbf{J}_k), k \in \mathcal{K} \tag{12b}$$

$$\left\|\mathbf{V}^i_{f,l}\right\|^2_F \le \theta^{(r)}\left(e^i_{f,l}\right) P_0,\ \forall\, e^i_{f,l} \in \mathcal{E}, \mathbf{V}^i_{f,l} \in \mathcal{V} \tag{12c}$$

$$0 \le e^i_{f,l} \le 1,\ \forall\, e^i_{f,l} \in \mathcal{E} \tag{12d}$$

$$\text{(8b), (8c), (8d), (8e),} \tag{12e}$$

which is a convex problem and can be efficiently solved via CVX [18]. We start with any points $(\mathcal{E}^{(0)},\mathcal{V}^{(0)})$ feasible to $\mathcal{P}_0$, and iteratively solve problem (12) until variables $(\mathcal{E}^{(r)},\mathcal{V}^{(r)})$ convergence. Solving such a continuous relaxation problem (12) may also result in some non-binary assignment variables $\mathcal{E}^*$ even with the exponential penalization. Thus, we propose to set all assignment variables to 0 when $(e^i_{f,l})^* < \epsilon$, otherwise 1. To make the solution feasible and encourage a sparse structure of beamformers, the threshold ϵ is carefully selected with value $\min_{f,l}\{\max_i (e^i_{f,l})^*\}$. The derived approach is summarized as Algorithm 1.

For problem (12), any feasible solution is also feasible for problem $\mathcal{P}^c_0$, but the reverse usually does not hold. One can prove that the optimal value of (12) normally serves as a locally tight upper bound of problem $\mathcal{P}^c_0$ by following propositions 1 and 2 in [19]. Subsequently in Algorithm 1, a sequence of points $\{\mathcal{E}^r,\mathcal{V}^r\}$ are generated with decreasing objective value for problem $\mathcal{P}^c_0$ and eventually, convergence to a local point of $\mathcal{P}^c_0$ [19, 20].

5. SIMULATION RESULTS AND CONCLUSION

In this section, we provide numerical simulation to show the performance of the proposed design in practical scenarios. Consider a CA-CRAN with 3 cells and each has a hexagonal shape, with the edge length of 300 m. A total of 3 BSs are in the cell and each BS

Algorithm 1 Proposed method for content assignment and beamforming joint design

1: **Initialize** $r = 0$, and set $(\mathcal{E}^{(0)}, \mathcal{V}^{(0)})$ feasible to $\mathcal{P}_0$
2: **repeat**
3: Solve the problem $\mathcal{P}_1^{(r)}$ for an optimal solution $(\mathcal{E}^*, \mathcal{V}^*)$
4: $\mathcal{E}^{(r+1)} \leftarrow \mathcal{E}^*$
5: $\mathcal{V}^{(r+1)} \leftarrow \mathcal{V}^*$
6: $r \leftarrow r + 1$
7: **until** the stopping criterion is satisfied, and output $(\mathcal{E}^*, \mathcal{V}^*)$
8: Set $(e_{f,l}^i)^* = 0$ when $(e_{f,l}^i)^* < \epsilon$, otherwise 1.
9: Run steps 2–7 again with the fixed $\mathcal{E}^*$, and output $\mathcal{V}^*$.

is located at the center of each cell, owning antennas $M = 12$ with the gain of 5 dBi. Assume BSs are equipped with the same caching storage. Each user equipment has 3 antennas. The number of transmit data streams $d = 1$. The exponent parameter for path loss is set as 2. Log-normal shadowing parameter is assumed to be 7 dB. The small-scale fading is Rayleigh fading, with covariance 1. Channel bandwidth is set as 1 MHz and noise power spectral density is -174 dBm/Hz. There are 100 files in library, following Zipf distribution with parameter $\gamma = 0.5$. Each file is 1GB and divided into 5 subfiles. We use the randomized fractional cache distinct pre-fetching approach as a baseline for content placement, which is provided in [9]. The total latency of the network is exploited by averaging results of 100 independent simulation trials.

First, we evaluate the performance of the proposed design. Consider that 3 users are active, the capacity of fronthaul link C_F is 1 Mbps, and fractional caching capacity of each BS $\mu = 0.4$. Set penalty $\alpha = 5$. In Fig. 3, the network latency is shown in relation to the total transmit power. We can see the results of the following benchmarks: (i) "Greedy Assignment" (GA) indicates the case where problem (8) is solved by greedy method, inspired by the idea in [16] with high computation complexity, which serves as near optimal result; (ii) "Random Assignment" (RA) shows the case where the content assignment is randomly selected without any design; (iii)"Non-cooperation " (NC) indicates the case where each user is only served by one BS; (iv) "Traffic-Aware " (TA) is obtained by minimizing the total data size of the fronthaul traffic. Consequently, the proposed method and GA achieve better results than benchmarks (ii)–(iv), due to the joint content assignment and BS cooperation. In particular, the proposed design reduces latency significantly compared with RA, which reveals that the proper content assignment will bring along benefits. It can be observed that when the transmit power is less than 37 dBm, NC scheme suffers from a longer delay than RA. This indicates the case where the total latency is dominant by the wireless transmission, the non-cooperative method will degrade the performance dramatically. Although the proposed design obtains slightly lower latency in contrast with TA, the gap will increase successively versus transmit power. This is because the physical transmission is considered in our design. Furthermore, the average simulation time required for the different schemes is shown in Table 1. It can be seen that compared with benchmarks (ii)–(iv), the proposed algorithm reduces latency substantially while the increase of time cost is not significant. When the transmit power is larger than 40 dBm, the proposed algorithm achieves better performance while consuming less simulation time as compared with GA, indicating that a lower computational complexity. In summary, these results demonstrate the superior performance of the proposed design.

Table 1. Average Simulation Time for Different Schemes

Schemes	GA	**PD**	RA	NC	TA
Time (s)	3636.83	**227.84**	48.08	78.58	63.67

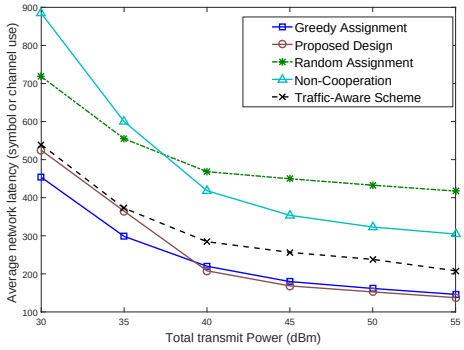

Fig. 3. Average latency vs. the transmit power.

In the following, we consider a larger system with 6 users and set transmit power as 40 dBm. Fig. 4 investigates the impact of the fronthaul link capacity. The fractional caching storage is 0.4. We can observe that with penalty $\alpha = 5$, a lower latency can be obtained because the exponent penalty is likely to result towards a binary solution. Furthermore, as the capacity of the fronthaul link increases, T_F reduces significantly while the edge latency T_E sees almost no change. Thus, the total latency of the network is dominant by the edge latency T_E for unlimited fronthaul capacity.

Fig. 5 exploits the impact of catching capability of BS μ on the network latency. The fronthaul link capacity is 3 Mbps. Interestingly, as the fractional catching storage $0.2 \leq \mu \leq 0.6$, the fronthaul latency T_F reduces rapidly while the edge latency T_E fluctuates around a certain level. Such thresholds are important for system design, because after which the benefit with more storage is limited. Accordingly the proposed design can help balance the system cost.

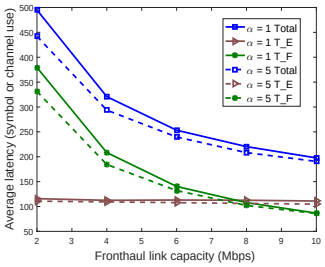

Fig. 4. Average latency vs. fronthaul link capacity.

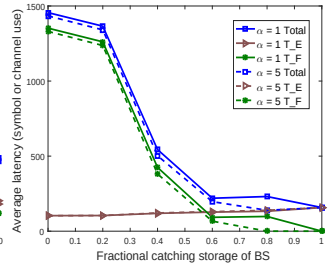

Fig. 5. Average latency vs. caching storage of BSs.

In this paper, we proposed a unified delivery design for CA-CRAN to achieve low latency with fairness among users. To fully exploit the benefits of caching resources, we jointly process content assignment and beamforming. To address the mixed binary nonconvex problem, an SCA-based algorithm is derived with low complexity. Simulations results demonstrate the superior performance of the proposed method over other existing schemes.

6. REFERENCES

[1] R. Huo, F. R. Yu, T. Huang, R. Xie, J. Liu, V. C. Leung, and Y. Liu, "Software defined networking, caching, and computing for green wireless networks," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 185–193, 2016.

[2] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, 2017.

[3] 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, 2016.

[4] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proc. IEEE INFOCOM*. Apr–May, 2014, pp. 1078–1086.

[5] 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, 2017.

[6] 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, 2016.

[7] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, 2015.

[8] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in Fog radio access networks," in *Proc. IEEE Intern. Symp. on Inf. Theory (ISIT)*. July, 2016, pp. 2029–2033.

[9] 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, 2016.

[10] S. H. Park, O. Simeone, W. Lee, and S. Shamai, "Coded multicast fronthauling and edge caching for multi-connectivity transmission in Fog radio access networks," CoRR, vol. abs/1705.04070, 2017.[Online]. Available:https://arxiv.org/abs/1705.04070

[11] X. Wang, J. Wang, and Y. Xu, "Data dissemination in wireless sensor networks with network coding," *EURASIP J. Wireless Commun. Netw.*, vol. 2010, no. 1, p. 465915, 2010.

[12] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, 2016.

[13] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[14] J. Koh, O. Simeone, R. Tandon, and J. Kang, "Cloud-aided edge caching with wireless multicast fronthauling in Fog radio access networks," in *Proc. IEEE Wireless Comm. Network. Conf. (WCNC)*. March, 2017, pp. 1–6.

[15] T. Ding, X. Yuan, and S. C. Liew, "Network-coded fronthaul transmission for cache-aided C-RAN," in *Proc. IEEE Intern. Symp. on Inf. Theory (ISIT)*. June, 2017, pp. 1182–1186.

[16] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.* , vol. 13, no. 5, pp. 2809–2823, 2014.

[17] O. Tervo, L.-N. Tran, H. Pennanen, S. Chatzinotas, M. Juntti, and B. Ottersten, "Energy-efficient coordinated multi-cell multigroup multicast beamforming with antenna selection," *Proc. IEEE Int. Conf. on Commun Workshops. (ICC Workshops)*, May, 2017, pp. 1209–1214.

[18] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008.

[19] H. H. M. Tam, H. D. Tuan, and D. T. Ngo, "Successive convex quadratic programming for quality-of-service management in full-duplex MU-MIMO multicell networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2340–2353, 2016.

[20] H. H. M. Tam, H. D. Tuan, A. A. Nasir, T. Q. Duong, and H. V. Poor, "MIMO energy harvesting in Full-duplex multiuser networks," *IEEE Trans. Wireless Commun.* , vol. 16, no. 5, pp. 3282–3297, 2017.

[21] H. Shi, R. V. Prasad, E. Onur, and I. G. M. M. Niemegeers, "Fairness in wireless networks: Issues, measures and challenges," *IEEE Commun. Surveys Tuts.* , vol. 16, no. 1, pp. 5–24, 2014.