# CLOUD RADIO ACCESS NETWORK WITH OPTIMIZED BASE-STATION CACHING

*Binbin Dai and Wei Yu*

Department of Electrical and Computer Engineering
University of Toronto, Toronto ON, Canada M5S 3G4
Emails: {bdai, weiyu}@comm.utoronto.ca

*Ya-Feng Liu*

LSEC, ICMSEC, AMSS
Chinese Academy of Sciences, Beijing 100190, China
Email: yafliu@lsec.cc.ac.cn

## ABSTRACT

The performance of cloud radio access networks (C-RAN) is limited by the finite capacities of the backhaul links connecting the cloud with the base-stations (BSs). A promising approach to improving the performance of C-RAN is to augment the backhaul through BS caching, where the BSs pre-store some of the popular contents. In this paper, we first derive a multicast backhaul rate expression based on a joint cache-channel coding scheme, and show that, as compared to the uniform cache allocation, it is better to allocate larger cache sizes to the weaker BSs. Then, by leveraging the sample approximation method and the alternating direction method of multipliers, we develop an efficient algorithm to optimize the cache allocation by maximizing the BS expected file downloading rate from the cloud. Numerical results show considerable performance improvement of the optimized cache allocation scheme over heuristic schemes.

***Index Terms***— Caching, C-RAN, multicast backhaul

## 1. INTRODUCTION

Cloud radio access network (C-RAN) has been recognized as one of the enabling technologies for the next generation wireless networks [1, 2, 3]. In C-RAN, the base-stations (BSs) are connected to a centralized processor (CP) through the backhaul links. A main advantage of C-RAN is that it enables cooperation among the BSs for inter-cell interference cancellation. The cooperation gain brought by C-RAN is however limited by the capacities of the backhaul links [4, 5]. In particular, this paper focuses on the data-sharing strategy for the downlink C-RAN, where the user messages are shared with multiple BSs for cooperative transmission. In this case, the BS cooperation cluster size is limited by the backhaul.

This paper proposes the use of caching at the BSs for alleviating backhaul traffic in the downlink C-RAN. Caching at the network edge has attracted extensive research interests. For example, the pioneering work of [6] uses network coding to simultaneously deliver multiple files through a commmon noiseless channel to multiple receivers, each caching different parts of the files. This paper studies a different scenario in which the *same* content is required at multiple BSs, but the backhaul is wireless with *different* channel conditions to different BSs. In particular, we consider a practical C-RAN model where each BS only caches a fraction of each file and requests the rest from the CP via noisy wireless backhaul channel. Under a total cache size constraint, we investigate the optimal cache size allocation strategy at each BS and the transmission strategy at the cloud such that the file requests can be delivered most efficiently from the cloud to the BSs.

As related works, wireless multicast backhaul for C-RAN has been considered in [7, 8]. However, [7] does not consider BS caching, while [8] considers network coded multicasting with fixed cache sizes. This paper differs from the above works in optimizing the cache allocation among the BSs to further improve the efficiency of wireless multicast backhaul.

This paper is motivated by the information theoretical study of [9], which shows that it is advantageous to allocate different cache sizes to different BSs depending on the channel conditions. In addition, [9] proposes a joint cache-channel coding scheme that optimally utilizes the caches at the BSs. We take the findings in [9] one step closer to the practice and consider the cache allocation problem under a realistic downlink C-RAN setup with a single BS cluster. This paper first derives a new multicast backhaul rate expression with BS caching, then formulates the cache allocation problem which maximizes the expected file delivering rate from the cloud to the BSs under a total cache size constraint. By taking advantage of the sample approximation method and the alternating direction method of multipliers (ADMM), we propose an efficient cache allocation scheme that together with transmit optimization considerably outperforms the uniform cache allocation scheme and the heuristic scheme of allocating cache in proportion to the long term channel statistics.

## 2. CACHING IN C-RAN BACKHAUL

Consider a downlink C-RAN model in which all the BSs are connected to a cloud center through shared wireless backhaul. The cloud employs a *data-sharing* strategy which delivers each user's intended message to a predefined cluster of BSs and the BS cluster subsequently serves the user through cooperative beamforming. The performance of C-RAN is largely limited by the capacities of the backhaul [4, 5]. In order to mitigate the demand for high backhaul capacities needed for large cooperation gain, we assume that each BS is equipped with a local cache and can pre-store content during off-peak traffic time to reduce the peak time backhaul traffic.

For simplicity, we consider a network consisting of a single cluster of cooperative BSs. A file of size $F$ needs to be delivered to all the BSs in order to enable cooperation. Each BS $l \in \mathcal{L} := \{1, 2, \ldots, L\}$ is equipped with a local storage unit of size $C_l$ that can cache some contents of the file, where $\sum_{l \in \mathcal{L}} C_l \leq C$. In this paper, the backhaul connecting the cloud with the BSs is assumed to be a shared wireless medium. We address the question of how to distribute the total cache size $C$ among the BSs and design transmission strategies at the cloud such that the cloud can deliver the file to the BSs in the most effective way.

We assume a block-fading model for the wireless backhaul channel from the cloud to the BSs. The cloud transmitter has $M$ transmit antennas, while all the BSs are equipped with a single antenna. Let $\mathbf{h}_l \in \mathbb{C}^{M \times 1}$ denote the channel vector between the cloud transmitter and BS $l$ within a coherent block. The received signal at BS $l$ can be written as

$$y_l = \mathbf{h}_l^H \mathbf{x} + z_l \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{M \times 1}$ is the transmit signal of the cloud transmitter, $y_l \in \mathbb{C}$ is the received signal at BS $l$, and $z_l \sim \mathcal{CN}\left(0, \sigma^2\right)$ is the background noise at BS $l$. We optimize the transmit covariance matrix for each channel realization.

## 3. EFFECTIVE DOWNLOAD RATE WITH CACHING

### 3.1. Broadcast Channel with Uncoded Caching

The downlink C-RAN wireless backhaul network with a single BS cluster can be modeled as a broadcast channel (BC) with common message, the capacity of which is given as

$$R_0 \leq \min_l \left\{ I\left(\mathbf{x}; y_l\right) \right\}, \tag{2}$$

where $R_0$ denotes the multicast rate, $I\left(\mathbf{x}; y_l\right)$ is the mutual information between the transmit signal $\mathbf{x}$ at the cloud and the received signal $y_l$ at BS $l$. It can be seen from (2) that the common information rate is limited by the worst channel.

With caching at the BSs and assuming that BS $l$ fills up its cache by naively caching the $C_l/F$ fraction of the file, the cloud then only needs to deliver the rest $1 - C_l/F$ fraction of the file to each BS. However, since the BSs are served through multicasting, the cloud has to send the maximum of the rest of the request file, i.e., $\max_l \{1 - C_l/F\}$, to make sure that the BS with least cache size can get the entire file. Therefore, the effective file downloading rate is

$$D_0 = \frac{\min_l \left\{ I\left(\mathbf{x}; y_l\right) \right\}}{\max_l \left\{ 1 - C_l/F \right\}}. \tag{3}$$

Under this naive caching strategy, it is optimal to allocate the cache size uniformly, i.e., $C_l = C/L, \forall\, l \in \mathcal{L}$.

### 3.2. BC with Joint Cache-Channel Coding

The above downloading rate can be improved if a joint cache-channel coding scheme is employed, in which BSs use the cached bits to facilitate the decoding of the received signal. From an information theoretic analysis, we have:

**Lemma 3.1 ([9])** *In a broadcast channel with common message, if each receiver $l$ caches $m_l$ bits, the common message rate $R_c$ bits per channel use is achievable if and only if*

$$R_c \leq I\left(\mathbf{x}; y_l\right) + m_l, \forall\, l \in \mathcal{L}. \tag{4}$$

Now if the BS $l$ caches $C_l/F$ fraction of the file, it only needs to have its channel to be able to support the rest $1 - C_l/F$ fraction of the file. By (4), this condition needs to be satisfied for each BS individually. This gives the effective downloading rate with caching as

$$D_c = \min_l \left\{ \frac{I\left(\mathbf{x}; y_l\right)}{1 - C_l/F} \right\}. \tag{5}$$

Clearly, the effective downloading rate in (5) is strictly larger than the one in (3) except when all $I\left(\mathbf{x}; y_l\right)$ are equal. Instead of allocating the cache size $C_l$ uniformly, (5) shows that it is advantageous to allocate more cache to the BS with weaker channel to achieve an overall higher multicast rate. In the next section, we formulate an optimization problem to find the optimal cache allocation among the BSs.

## 4. OPTIMIZING BS CACHE ALLOCATION

Based on (5) and the channel model (1) and by further including the optimization of the transmit covariance matrix, we now formulate the optimal multicast downloading rate problem with BS caching as

$$D_c^* = \max_{\mathbf{W} \in \mathbb{W}} \min_l \left\{ \frac{\log\left(1 + \frac{\mathrm{Tr}(\mathbf{H}_l \mathbf{W})}{\sigma^2}\right)}{1 - C_l/F} \right\}, \tag{6}$$

where $\mathbf{H}_l = \mathbf{h}_l \mathbf{h}_l^H$, $\mathbf{W}$ is the covariance matrix of the transmit signal $\mathbf{x}$, and $\mathbb{W} = \{\mathbf{W} \succeq \mathbf{0} \mid \mathrm{Tr}\left(\mathbf{W}\right) \leq P\}$ with $P$ being the transmit power budget.

Note that the optimal multicast rate in (6) is a function of both the channel condition and the cache allocation. In practice, the transmit covariance matrix is adapted to each channel realization, but the cache allocation is determined at the cache deployment phase so can only adapt to the channel statistics. We therefore take expectation of $D_c^*$ in (6) over the channel distribution, and aim to find an optimal cache allocation that maximizes the long-term expected effective downloading rate. The overall optimization problem is now formulated as:

$$\underset{\{C_l\}}{\text{maximize}} \quad \mathbb{E}_{\{\mathbf{H}_l\}}\left[D_c^*\right] \tag{7a}$$

$$\text{subject to} \quad \sum_{l \in \mathcal{L}} C_l \leq C,\ 0 \leq C_l \leq F,\ l \in \mathcal{L}. \tag{7b}$$

In the full version of this paper [10], we also consider the expected file downloading time as the objective function.

Finding a closed-form expression for the objective function in (7a) is in general challenging. This paper proposes to replace the objective function of the above with its sample approximation [11] and to reformulate the problem as:

$$\underset{\{C_l,\,\mathbf{W}^n\}}{\text{maximize}} \quad \frac{1}{N}\sum_{n=1}^{N}\min_{l}\left\{\frac{\log\left(1+\frac{\text{Tr}(\mathbf{H}_l^n\mathbf{W}^n)}{\sigma^2}\right)}{1-C_l/F}\right\} \tag{8a}$$

$$\text{subject to} \quad \sum_l C_l \leq C,\ 0\leq C_l\leq F,\ l\in\mathcal{L}, \tag{8b}$$

$$\text{Tr}\left(\mathbf{W}^n\right)\leq P,\ \mathbf{W}^n\succeq\mathbf{0},\ n\in\mathcal{N}, \tag{8c}$$

where $\mathcal{N} := \{1,2,\ldots,N\}$, $\{\mathbf{H}_l^n\}_{n\in\mathcal{N}}$ are the samples drawn according to the distribution of $\mathbf{H}_l$, and $\mathbf{W}^n$ is the covariance matrix adapted to the samples $\{\mathbf{H}_l^n\}_{l\in\mathcal{L}}$.

Problem (8) is still not easy to solve mainly due to the following two reasons. First, the objective function of problem (8) is nonsmooth and nonconvex, although all of its constraints are convex. Second, the sample size $N$ generally needs to be sufficiently large such that the sample average is a good approximation to the original expected rate [11], leading to a high complexity for solving problem (8) directly. In the next section, we first reformulate problem (8) as a smooth problem, then linearize the nonconvex term, and finally leverage the ADMM to decouple the problem into $N$ low-complexity convex subproblems.

## 5. ADMM WITH SAMPLE APPROXIMATION

Dropping the constant $1/N$ and introducing the auxiliary variables $\{\xi^n\}$, we first reformulate problem (8) as

$$\underset{\{C_l,\,\mathbf{W}^n,\,\xi^n\}}{\text{maximize}} \quad \sum_{n=1}^{N}\xi^n \tag{9a}$$

$$\text{subject to} \quad \log\left(1+\frac{\text{Tr}\left(\mathbf{H}_l^n\mathbf{W}^n\right)}{\sigma^2}\right)\geq\xi^n(1-C_l/F),\quad l\in\mathcal{L},\ n\in\mathcal{N}, \tag{9b}$$

$$\text{(8b) and (8c).}$$

The above problem (9) is smooth but still nonconvex due to the term $\xi^n(1-C_l/F)$. To deal with the nonconvex term, we approximate it by its first-order Taylor expansion at some appropriate point $(\bar{\xi}^n,\bar{C}_l)$, i.e.,

$$\begin{aligned}\xi^n(1-C_l/F) \approx\ & \bar{\xi}^n(1-\bar{C}_l/F)+\left[1-\bar{C}_l/F,\ -\bar{\xi}^n/F\right]\\ & \left[\xi^n-\bar{\xi}^n,\ C_l-\bar{C}_l\right]^T\\ =\ & \bar{\xi}^n\left(1-C_l/F\right)+\left(1-\bar{C}_l/F\right)\left(\xi^n-\bar{\xi}^n\right).\end{aligned}$$

An iterative first-order approximation leads to the Algorithm 1 for solving problem (8) shown on the next page. More specifically, let $\{\xi^n(t),\,C_l(t)\}$ be the iterates at the $t$-th iteration, the algorithm solves

$$\underset{\{C_l,\,\mathbf{W}^n,\,\xi^n\}}{\text{maximize}} \quad \sum_{n=1}^{N}\xi^n \tag{10a}$$

$$\begin{aligned}\text{subject to}\ \log\left(1+\frac{\text{Tr}\left(\mathbf{H}_l^n\mathbf{W}^n\right)}{\sigma^2}\right) &\geq \xi^n(t)\left(1-C_l/F\right)\\ &+\left(1-C_l(t)/F\right)\left(\xi^n-\xi^n(t)\right),\ l\in\mathcal{L},\ n\in\mathcal{N},\end{aligned} \tag{10b}$$

$$|\xi^n-\xi^n(t)|\leq r(t),\ n\in\mathcal{N}, \tag{10c}$$

$$\text{(8b) and (8c),}$$

where (10c) is the trust region constraint (i.e., within the region the linear approximation in (10b) is trusted to be of good quality) and $r(t)$ is the radius of the trust region at the $t$-th iteration; then the algorithm updates the parameters for the next iteration as

$$\xi^n(t+1)=\min_{l\in\mathcal{L}}\left\{\frac{\log\left(1+\frac{\text{Tr}(\mathbf{H}_l^n\mathbf{W}^{n*}(t))}{\sigma^2}\right)}{1-C_l^*(t)/F}\right\},\ n\in\mathcal{N}, \tag{11}$$

$$C_l(t+1)=C_l^*(t),\ l\in\mathcal{L}, \tag{12}$$

where $\mathbf{W}^{n*}(t)$ and $C_l^*(t)$ are solutions to problem (10). For the initial point, we can set $C_l(1)$ to be $C/L$ for all $l\in\mathcal{L}$, and decouple problem (9) into $N$ convex optimization subproblems to solve for $\xi^n(1)$ for all $n\in\mathcal{N}$.

It remains to solve problem (10). Note that problem (10) is a convex problem but with a (potentially) large number of variables (due to the large sample size). We propose to use the ADMM [12] to solve problem (10), which decouples the high-dimensional problem into many small-dimensional subproblems. In particular, we introduce the so-called consensus constraints $C_l^n=C_l,\ \forall l,\ n$ and reformulate problem (10) as

$$\underset{\substack{\{\xi^n,\,\mathbf{W}^n,\\ C_l^n,\,C_l\}}}{\text{maximize}} \quad \sum_{n=1}^{N}\xi^n \tag{13a}$$

$$\begin{aligned}\text{subject to}\ \log\left(1+\frac{\text{Tr}\left(\mathbf{H}_l^n\mathbf{W}^n\right)}{\sigma^2}\right) &\geq \xi^n(t)\left(1-C_l^n/F\right)\\ &+\left(1-C_l(t)/F\right)\left(\xi^n-\xi^n(t)\right),\ l\in\mathcal{L},\ n\in\mathcal{N},\end{aligned} \tag{13b}$$

$$C_l^n=C_l,\ l\in\mathcal{L},\ n\in\mathcal{N}, \tag{13c}$$

$$|\xi^n-\xi^n(t)|\leq r(t),\ n\in\mathcal{N}, \tag{13d}$$

$$\text{(8b) and (8c).}$$

The partial augmented Lagrangian of problem (13) is

$$\begin{aligned}\mathcal{L}_\rho\left(\xi^n,\mathbf{W}^n,C_l^n,C_l;\lambda_l^n\right)=-\sum_{n=1}^{N}\xi^n+&\\ \sum_{l\in\mathcal{L}}\sum_{n\in\mathcal{N}}\left[\lambda_l^n\left(C_l^n-C_l\right)+\frac{\rho}{2}\left(C_l^n-C_l\right)^2\right],&\end{aligned} \tag{14}$$

**Algorithm 1** Proposed Algorithm for Problem (8)

```
Initialization: Initialize $C_l(1) = C/L,\ l \in \mathcal{L}$, and $\xi^n(1)$ as the solution to problem (8) with $C_l = C_l(1)$; set $t = 1$;
Repeat:
    1. Use the ADMM to solve problem (10);
    2. Update $\{\xi^n(t+1), C_l(t+1)\}$ according to (11) and (12), respectively;
    3. Set $t = t + 1$;
Until convergence
```

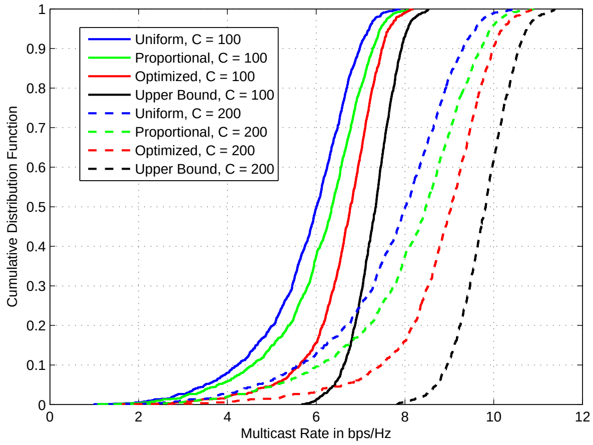

**Fig. 1**. CDF of multicast rates under different caching schemes.

where $\lambda_l^n$ is the Lagrange multiplier corresponding to the constraint $C_l = C_l^n$ and $\rho > 0$ is the penalty parameter. To solve the original problem, the ADMM sequentially updates the primal variables via minimizing the augmented Lagrangian, followed by an update of the dual variable. In our case, at each iteration, the ADMM first minimizes (14) subject to (13b), (13d), and (8c) over the variables $\{\xi^n, \mathbf{W}^n, C_l^n\}$ ; then minimizes (14) subject to (8b) over $\{C_l\}$ ; and finally updates the Lagrange multiplier. All subproblems at each iteration of the ADMM here are easy to solve. In particular, the subproblem of minimizing (14) subject to (13b), (13d), and (8c) over the variables $\{\xi^n, \mathbf{W}^n, C_l^n\}$ automatically decouples into $N$ low-dimensional convex problems.

## 6. SIMULATION RESULTS

We consider a downlink C-RAN model with $L = 5$ BSs located at distances $(398, 278, 473, 286, 267)$ meters from the cloud. The cloud transmitter is equipped with $M = 10$ transmit antennas. We generate 1000 sets of channel realizations according to $\mathbf{h}_l = \mathbf{K}_l^{1/2}\mathbf{v}_l$, where $\mathbf{K}_l$ is fixed and generated in the same fashion as in [13] with path-loss component modeled as $128.1 + 37.6\log_{10}(d)$ dB and $d$ is the distance between the cloud and the BS; $\mathbf{v}_l$ is a Gaussian random vector with each element independently and identically distributed as $\mathcal{CN}(0, 1)$. The first $N = 100$ sets of channel realizations are used in problem (8) to optimize the cache allocation while the rest 900 are used to evaluate the multicast rates under the optimized cache allocation. The transmit power at the cloud is set to be 40 watts and the background noise level is set to be $-150$ dBm/Hz. The file size is set as $F = 100$.

**Table 1**. Cache allocations for $(\text{BS}_1, \ldots, \text{BS}_5)$ at distances $(398, 278, 473, 286, 267)$ meters from the cloud transmitter with file size $F = 100$.

| | $C = 100$ | $C = 200$ |
|---|---|---|
| Uniform | $(20, 20, 20, 20, 20)$ | $(40, 40, 40, 40, 40)$ |
| Proportional | $(23, 17, 26, 18, 16)$ | $(42, 38, 45, 38, 37)$ |
| Optimized | $(25, 10, 45, 13, 7)$ | $(44, 33, 58, 35, 30)$ |

We compare our proposed cache allocation scheme with the following two benchmarks: the uniform cache allocation scheme and the proportional allocation scheme which allocates cache such that $\log\left(1 + \frac{P\text{Tr}(\mathbf{K}_l)}{L\sigma^2}\right) / (1 - C_l/F)$ for all $l$ are equalized (if possible). We also simulate an (impractical) scheme of dynamically and optimally allocating cache based on each channel realization, which provides a (generally not achievable) upper bound of the optimal multicast rate in problem (8). Note that the cache allocation problem based on each channel realization is a convex optimization problem.

In Table 1, we list the cache size allocated to each BS under two different settings of the total cache size $C = 100$ and 200 for the proposed optimized scheme as compared to the two benchmarks. Fig. 1 plots the cumulative distribution functions (CDF) of the effective multicast rates under the cache allocation in Table 1. As we can see from Fig. 1, the proposed optimized caching scheme exhibits considerable performance improvement as compared to the other two naive baseline schemes (i.e., the uniform and proportional allocation schemes). The improvement is due to that the BSs farther away from the cloud are more aggressively allocated larger amount of cache under the optimized scheme.

## 7. CONCLUSION

This paper studies the optimal BS cache allocation problem in the downlink C-RAN with edge caching. We first derive the optimal file downloading rate with given BS cache size, then formulate the cache optimization problem of maximizing the expected downloading rate over channel fading realizations subject to the total cache size constraint. By leveraging the sample approximation method and the ADMM, we propose an efficient cache allocation algorithm. Simulation results show that the optimized cache allocation scheme significantly outperforms the heuristic caching schemes.

## 8. REFERENCES

[1] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[2] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[3] T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, Eds., *Cloud Radio Access Networks: Principles, Technologies, and Applications*, Cambridge University Press, 2017.

[4] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited-backhaul capacity," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1–10, Dec. 2009.

[5] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access, Special Issue on Recent Advances in Cloud Radio Access Networks*, vol. 2, pp. 1326–1339, 2014.

[6] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[7] 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, Aug. 2017.

[8] S.-H. Park, O. Simeone, W. Lee, and S. Shamai (Shitz), "Coded multicast fronthauling and edge caching for multi-connectivity transmission in fog radio access networks," 2017. [Online]. Available: http://arxiv.org/abs/1705.04070.

[9] S. S. Bidokhti, M. A. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," 2016. [Online]. Available: http://arxiv.org/abs/1605.02317.

[10] B. Dai, Y.-F. Liu, and W. Yu, "Optimized base-station cache allocation for cloud radio access network with multicast backhaul," *submitted for possible publication*, 2018.

[11] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, 2nd edition, 2011.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] A. Lozano, "Long-term transmit beamforming for wireless multicasting," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2007, vol. 3, pp. 417–420.