Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN
IEEE TWC, Vol. 15, No. 9, September 2016

Sai Subramanyam Thoota

SPC Lab, Department of ECE
Indian Institute of Science

December 31, 2016

Table of contents

Introduction

- Motivation
  - Wireless services are experiencing a shift from connection-centric communications (phone calls, e-mails etc.) to content-centric communications (video streaming, mobile TV etc.).
  - Content diversity: Same copy of content may be needed by multiple mobile users. Multicasting and caching are two enabling techniques to exploit such content diversity.
- Contributions
  - Joint design of content-centric base station (BS) clustering and multicast beamforming to improve the network performance and to reduce the backhaul cost.
  - Formulation of a mixed-integer nonlinear programming (MINLP) problem to minimize the total network cost subject to QoS constraint for each multicast group.
  - Reformulation to a sparse multicast beamforming (SBF) problem, which is transformed into the difference of convex programming problems and solved using convex-concave (CCP) procedure algorithms.
  - Evaluation of the effects of heuristic caching strategies.

System Model

- Downlink transmission of a cache-enabled cloud RAN with multiple-antenna BSs and single-antenna mobile users.
- Each BS has a local cache with finite storage size and is connected to the central processor (CP) via a limited capacity backhaul link.
- Users requesting the same content are grouped together to form a multicast group. Each user can request a maximum of only one content at a time.
- Channel is assumed to be constant for each transmission frame.
- Notation
  - N - number of base stations.
  - K - number of mobile users.
  - L - number of transmit antennas in each BS.
  - $\mathcal{N} = \{1, 2, \ldots, N\}$ - set of BSs.
  - M - number of multicast groups.
  - F - total number of contents in the CP.
  - $\mathcal{G}_m$ - set of users in multicast group m.
  - $\mathcal{Q}_m$ - set of BSs serving the multicast group m.

System Model contd.

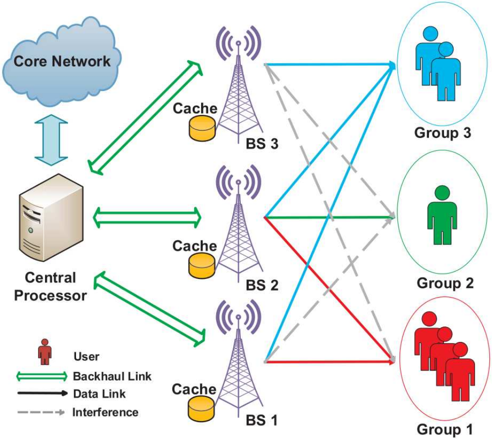
- Notation
  - $\mathbf{S} \in \{0,1\}^{M \times N}$ - binary BS clustering matrix.
  - $\mathbf{w}_m = \left[\mathbf{w}_{m,1}^H, \mathbf{w}_{m,2}^H, \ldots, \mathbf{w}_{m,N}^H\right]^H \in \mathbb{C}^{NL \times 1}$ - aggregate network-wide beamforming vector of group m from all BSs.
  - $\mathbf{h}_k \in \mathbb{C}^{NL \times 1}$ - network-wide channel vector from all the BSs to user k.
  - x_m - data symbol of the content requested by group m. $\mathbb{E}\left[|x_m|^2\right] = 1$.
  - $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ - additive white Gaussian noise at user k.
- Received signal at user k from group $\mathcal{G}_m$ can be written as

$$y_k = \mathbf{h}_k^H \mathbf{w}_m x_m + \sum_{j \neq m}^{M} \mathbf{h}_k^H \mathbf{w}_j x_j + n_k, \qquad \forall k \in \mathcal{G}_m. \tag{1}$$

- Received SINR for user $k \in \mathcal{G}_m$ is,

$$\mathrm{SINR}_k = \frac{\left|\mathbf{h}_k^H \mathbf{w}_m\right|^2}{\sum_{j \neq m}^{M} \left|\mathbf{h}_k^H \mathbf{w}_j\right|^2 + \sigma_k^2} \tag{2}$$

System Model contd.



Cache Model

- Notation
  - $\mathcal{F} = \{1, 2, \ldots, F\}$ - database of F contents, each with normalized size of 1.
  - $Y_n \triangleq$ Local storage size of BS n. $Y_n < F$.
  - $\mathbf{C} \in \{0, 1\}^{F \times N}$ - binary cache placement matrix, where $c_{f,n} = 1$ indicates that the f-th content is cached in the n-th BS.
- Due to limited cache size, $\sum_{f=1}^{F} c_{f,n} \leq Y_n$.
- Cache placement happens in a much larger timescale than scheduling and transmission.

Cost Model

- Total network cost consists o both the backhaul cost and the transmission power cost.
- Backhaul cost:

$$C_B = \sum_{m=1}^{M} \sum_{n=1}^{N} s_{m,n}(1 - c_{f_m,n}) R_m. \tag{3}$$

  where R_m is the content-delivery rate, f_m is the content requested by the multicast group m.
- Transmission power cost:

$$C_P = \sum_{m=1}^{M} \sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|^2. \tag{4}$$

- Total network cost:

$$C_N = C_B + \eta C_P \tag{5}$$

  where $\eta > 0$ is a weighting parameter between backhaul cost and transmission power.

Problem Formulation

- Objective: To optimize the BS clustering and multicast beamforming at each transmission frame so as to minimize the total network cost.
- This is formulated as

$$
\begin{aligned}
\mathcal{P}_0 : \min_{\{\mathbf{w}_{m,n}\},\{s_{m,n}\}} \quad & \sum_{m=1}^{M}\sum_{n=1}^{N} s_{m,n}\left(1 - c_{f_m,n}\right) R_m \\
& + \eta \sum_{m=1}^{M}\sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|^2 && \text{(6a)} \\
\text{s.t} \quad & \mathrm{SINR}_k \geq \gamma_m, \qquad \forall k \in \mathcal{G}_m, \forall m && \text{(6b)} \\
& s_{m,n} \in 0, 1, \qquad \forall m, n && \text{(6c)} \\
& \left(1 - s_{m,n}\right) \mathbf{w}_{m,n} = \mathbf{0}, \forall m, n. && \text{(6d)}
\end{aligned}
$$

Problem Formulation contd.

- Problem $\mathcal{P}_0$ is combinatorial in nature. 2^{MN} possible BS clustering matrices $\{\mathbf{S}\}$.
- For each BS clustering matrix $\mathbf{S}$, the backhaul cost C_B is constant, and the problem $\mathcal{P}_0$ reduces to

$$\mathcal{P}\left(\mathcal{Z}_{\mathbf{S}}\right) : \min_{\{\mathbf{w}_{m,n}\}} \sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|^2 \tag{7a}$$

$$\text{s.t} \quad (6b) \tag{7b}$$

$$\mathbf{w}_{m,n} = \mathbf{0}, \qquad \forall (m,n) \in \mathcal{Z}_{\mathbf{S}}. \tag{7c}$$

where $\mathcal{Z}_{\mathbf{S}} = \{(m,n) | s_{m,n} = 0\}$ is the set of inactive BS-content associations.

- Problem $\mathcal{P}(\mathcal{Z}_{\mathbf{S}})$ is a quadratically constrained quadratic programming (QCQP) problem.
- Once problem $\mathcal{P}(\mathcal{Z}_{\mathbf{S}})$ is solved for all possible BS clustering matrices $\mathbf{S}$'s, the one with the minimum objective is selected to be the global optimum.

Problem Formulation contd.

Proposition

If the content f_m requested by a multicast group m has been cached in BS n, i.e., $c_{f_m,n} = 1$, then without loss of optimality, one can set $s_{m,n} = 1$ in problem $\mathcal{P}_0$.

Proof.

- Assume that $c_{f_{m^*},n^*} = 1$.
- Consider an arbitrary BS clustering matrix $\mathbf{S}'$ with $s_{m^*,n^*} = 0$.
- Define a new BS clustering matrix $\mathbf{S}''$ with $s_{m^*,n^*} = 1$ and the remaining contents same as that of $\mathbf{S}'$.
- Since the feasible set of the original problem (with $\mathbf{S}'$) is a subset of the new optimization problem (with $\mathbf{S}''$), the cost $C'_P \geq C''_P$.

□

- Based on the above proposition, the authors have developed a cache aware greedy BS clustering algorithm, the details of which are omitted in the paper due to page limit.

Problem Formulation contd.

- BS cluster matrix $\mathbf{S}$ can be specified with the knowledge of the beamformers $\mathbf{w}_{m,n}$'s.
- When $\mathbf{w}_{m,n} = \mathbf{0}$,

$$s_{m,n} = \begin{cases} 0, & \text{if} \quad c_{f_m,n} = 0, \\ 0 \quad \text{or} \quad 1, & \text{if} \quad c_{f_m,n} = 1. \end{cases} \tag{8}$$

- When $\mathbf{w}_{m,n} = \mathbf{1}$, $s_{m,n} = 1$ from constraint (6d).
- Thus, $s_{m,n}$ can be replaced by the ℓ_0-norm of $\|\mathbf{w}_{m,n}\|^2$.
- $\mathcal{P}_0$ can be transformed into the following equivalent problem:

$$\begin{aligned} \mathcal{P}_{SBF} : \min_{\{\mathbf{w}_{m,n}\}} & \sum_{m=1}^{M} \sum_{n=1}^{N} \left\| \|\mathbf{w}_{m,n}\|_2^2 \right\|_0 (1 - c_{f_m,n}) R_m \\ & + \eta \sum_{m=1}^{M} \sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|_2^2 \qquad (9) \\ \text{s.t} \quad & \text{(6b).} \qquad (10) \end{aligned}$$

- Problem $\mathcal{P}_{SBF}$ is challenging due to the nonconvex QoS constraints and the nonconvex discontinuous ℓ_0-norm in the objective.

Basics of DC Programming and CCP

- Optimization problems of functions with each represented as a difference of two convex functions.
- General form of DC programming problems is written as:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & g_0(x) - h_0(x) \\ \text{s.t.} \quad & g_i(x) - h_i(x) \leq 0, \qquad i = 1, 2, \ldots, m. \end{aligned} \tag{11}$$

  where g_i and h_i for $i = 0, 1, \ldots, m$, are all convex functions.
- Main idea of CCP is to replace the concave part in the DC functions by their first order Taylor expansions, and solve a sequence of convex problems successively.
- CCP falls in the category of MM algorithms for a particular choice of the majorization function.
- CCP can also be derived from the DC algorithm (DCA), a primal-dual subdifferential method, where the objective can be a difference of proper lower semi-continuous convex functions.

Smoothed ℓ_0-norm Approximation

- The discontinuous ℓ_0-norm in the problem $\mathcal{P}_{SBF}$ is approximated with a continuous smooth function, denoted as $f(x)$.
- Three frequently used smooth concave functions:
  - logarithmic function.
  - exponential function.
  - arctangent function.

$$f_\theta(x) = \begin{cases} \frac{\log\left(\frac{x}{\theta}+1\right)}{\log\left(\frac{1}{\theta}+1\right)}, & \text{for log-function} \\ 1 - exp\left(-\frac{x}{\theta}\right), & \text{for exp-function} \\ \arctan\left(\frac{x}{\theta}\right), & \text{for arctan-function} \end{cases} \tag{12}$$

where $\theta > 0$ is a parameter controlling the smoothness of approximation.

Smoothed ℓ_0-norm Approximation contd.

- The problem $\mathcal{P}_{SBF}$ is approximated as:

$$\mathcal{P}_1 : \min_{\{\mathbf{w}_{m,n}\}} \sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{m,n} f_\theta \left(\|\mathbf{w}_{m,n}\|_2^2 \right) + \eta \sum_{m=1}^{M}\sum_{n=1}^{N} \|\mathbf{w}_{m,n}\|_2^2 \quad (13)$$
$$\text{s.t} \quad \text{(6b)}.$$

where $\alpha_{m,n} \triangleq (1 - c_{f_m,n})R_m, \forall m, n.$

- Note that the smooth function $f_\theta \left(\|\mathbf{w}_{m,n}\|_2^2 \right)$ is concave in $\|\mathbf{w}_{m,n}\|_2^2$, but not concave in $\mathbf{w}_{m,n}$.

SDR-based CCP Algorithm

- Semidefinite relaxation (SDR) approach to convert $\mathcal{P}_1$ into a DC program.
- Define two sets of matrices $\left\{\mathbf{W}_m \in \mathbb{C}^{NL\times NL}\right\}_{m=1}^{M}$ and $\left\{\mathbf{H}_k \in \mathbb{C}^{NL\times NL}\right\}_{k=1}^{K}$ as

$$\mathbf{W}_m = \mathbf{w}_m\mathbf{w}_m^H, \forall m \quad \text{and} \quad \mathbf{H}_k = \mathbf{h}_k\mathbf{h}_k^H, \forall k. \tag{14}$$

- Define a set of selection matrices $\{\mathbf{J}_n\}_{n=1}^{N}$, as

$$\mathbf{J}_n = \operatorname{diag}\left(\left[\mathbf{0}_{(n-1)L}^H, \mathbf{1}_L^H, \mathbf{0}_{(N-n)L}^H\right]\right), \quad \forall n. \tag{15}$$

- By removing the rank constraint $\operatorname{rank}\{\mathbf{W}_m\} = 1$, problem $\mathcal{P}_1$ can be relaxed as

$$\mathcal{P}_2: \min_{\{\mathbf{W}_m\}} \sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{m,n} f_\theta\left(\operatorname{Tr}\left(\mathbf{W}_m\mathbf{J}_n\right)\right) + \eta\sum_{m=1}^{M}\operatorname{Tr}\left(\mathbf{W}_m\right) \tag{16a}$$

$$\text{s.t} \quad \frac{\operatorname{Tr}\left(\mathbf{W}_m\mathbf{H}_k\right)}{\sum_{j\neq m}^{M}\operatorname{Tr}\left(\mathbf{W}_j\mathbf{H}_k\right) + \sigma_k^2} \geq \gamma_m, \quad \forall k \in \mathcal{G}_m, \forall m \tag{16b}$$

$$\mathbf{W}_m \succeq \mathbf{0}, \quad \forall m \tag{16c}$$

SDR-based CCP Algorithm contd.

- $\mathcal{P}_2$ is DC program with DC objective and convex constraints, which can be solved using CCP.
- If the solution to problem $\mathcal{P}_2$ is of rank 1, then eigen value decomposition is used to obtain the network-wide beamformer $\mathbf{w}_m^*$.
- Else, randomization and scaling methods are used to generate a suboptimal solution.
- By adopting SDR method, the number of variables is roughly squared (from MNL to $M(NL)^2$), which is not computationally efficient.

Generalized CCP Algorithm

- Nonconvex SINR constraints in $\mathcal{P}_1$ can be written as

$$\gamma_m \left(\sum_{j \neq m} \left| \mathbf{h}_k^H \mathbf{w}_j \right|^2 + \sigma_k^2 \right) - \left| \mathbf{h}_k^H \mathbf{w}_m \right|^2 \leq 0, \quad \forall k \in \mathcal{G}_m. \tag{17}$$

- By introducing auxiliary variables $\{t_{m,n} \in \mathbb{R}\}_{n=1,\ldots,N}^{m=1,\ldots,M}$, problem $\mathcal{P}_1$ can be transformed into the following problem as

$$\mathcal{P}_3 : \min \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha_{m,n} f_\theta \left(t_{m,n}\right) + \eta \sum_{m=1}^{M} \sum_{n=1}^{N} t_{m,n} \tag{18a}$$

$$\text{s.t} \quad \|\mathbf{w}_{m,n}\|_2^2 - t_{m,n} \leq 0, \quad \forall m, n, \tag{18b}$$

$$\gamma_m \left(\sum_{j \neq m} \left| \mathbf{h}_k^H \mathbf{w}_j \right|^2 + \sigma_k^2 \right) - \left| \mathbf{h}_k^H \mathbf{w}_m \right|^2 \leq 0, \quad \forall k \in \mathcal{G}_m. \tag{18c}$$

- Problem $\mathcal{P}_3$ is a general DC program with DC objective and constraints, which can be solved using CCP algorithm.

Discussions

- Initial feasible starting point is found by solving the following power minimization problem with full BS cooperation.

$$\mathcal{P}_{\text{INI}} : \min_{\{\mathbf{W}_m\}} \sum_{m=1}^{M} \text{Tr}\left(\mathbf{W}_m\right) \tag{19a}$$

$$\text{s.t} \quad \frac{\text{Tr}\left(\mathbf{W}_m \mathbf{H}_k\right)}{\sum_{j\neq m}^{M} \text{Tr}\left(\mathbf{W}_j \mathbf{H}_k\right) + \sigma_k^2} \geq \gamma_m, \quad \forall k \in \mathcal{G}_m, \forall m \tag{19b}$$

$$\mathbf{W}_m \succeq \mathbf{0}, \quad \forall m \tag{19c}$$

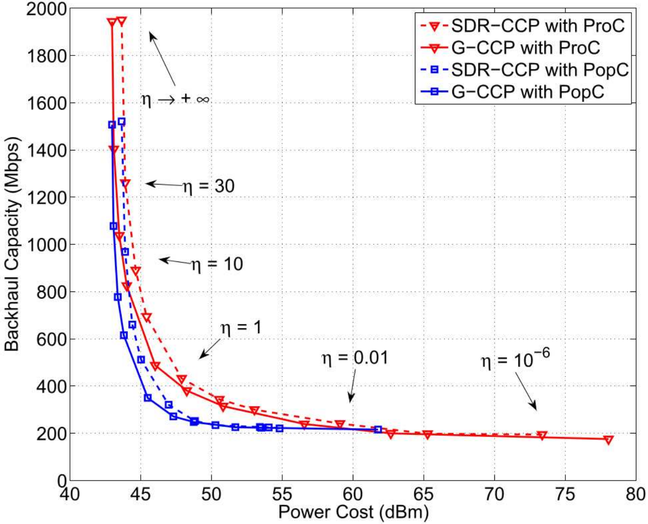
- The optimal solution $\{\mathbf{W}_m\}$ of $\mathcal{P}_{\text{INI}}$ can be used directly as a feasible starting point for the SDR-based CCP algorithm.
- For the generalized CCP algorithm, randomization and scaling need to be used if the optimal solution to $\mathcal{P}_{\text{INI}}$ is not of rank 1.
- If the problem $\mathcal{P}_{\text{INI}}$ turns out to be infeasible, then the original problem $\mathcal{P}_0$ is infeasible and both algorithms will terminate.

Simulation Results

- Backhaul-power tradeoff comparison between SDR-CCP and G-CCP.
- Performance comparison of three heuristic caching strategies for unequal content popularity (Zipf distribution is used for content popularity).
  - Popularity-aware caching.
  - Random caching.
  - Probabilistic caching.
- Performance comparison between unicast and multicast transmission.
- Effect of smooting approximation of ℓ_0-norm in the objective function.

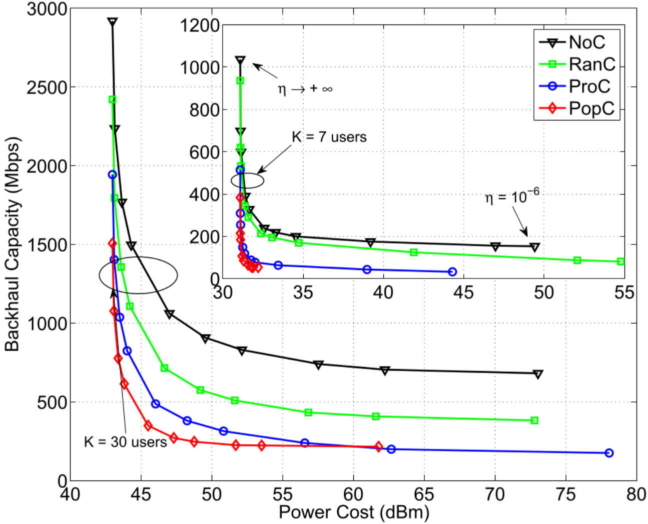
Simulation Results contd.

- Backhaul-power tradeoff comparison between SDR-CCP and G-CCP.



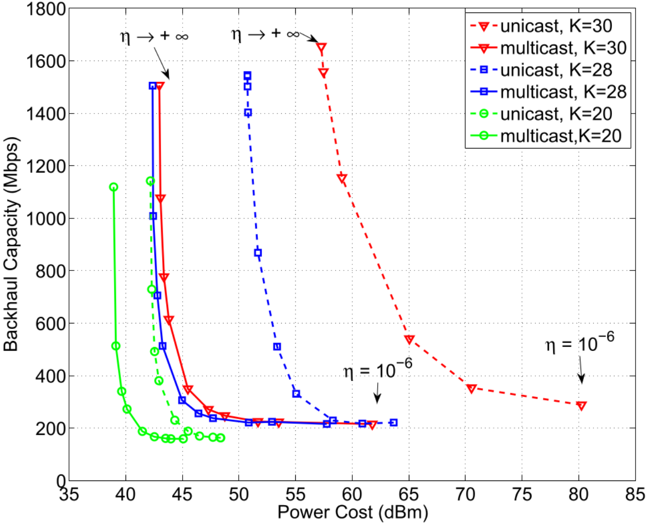
Simulation Results contd.

- Performance comparison of three heuristic caching strategies for unequal content popularity (Zipf distribution is used for content popularity).



Simulation Results contd.

- Performance comparison between unicast and multicast transmission.



Simulation Results contd.

- Effect of smooting approximation.