Outage Probability and Fronthaul Usage Tradeoff Caching Strategy in Cloud-RAN

Zhun Ye*, Cunhua Pan†, Huiling Zhu† and Jiangzhou Wang†
*School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, 264209, P. R. China
Email: zhunye@sdu.edu.cn
†School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NT, United Kingdom
Email: {c.pan, h.zhu, j.z.wang@kent.ac.uk}

***Abstract*—In this paper, optimal content caching strategy is proposed to jointly minimize the cell average outage probability and fronthaul usage in cloud radio access network (Cloud-RAN). An accurate closed form expression of the outage probability conditioned on the user's location is presented, and the cell average outage probability is obtained through the composite Simpson's integration. The caching strategy for jointly optimizing the cell average outage probability and fronthaul usage is formulated as a weighted sum minimization problem, which is a nonlinear 0-1 integer NP-hard problem. In order to deal with the NP-hard problem, at first, two particular caching placement schemes are investigated: the most popular content (MPC) caching scheme and the proposed location-based largest content diversity (LB-LCD) caching scheme. Then a genetic algorithm (GA) based approach is proposed. Numerical results show that the performance of the proposed GA-based approach with significantly reduced computational complexity is close to the optimal performance achieved by exhaustive search based caching strategy.**

I. Introduction

Consisting of centralized base band processing resources, known as base band unit (BBU) pool, and distributed remote radio heads (RRHs), cloud radio access network (Cloud-RAN) becomes a new type of radio access network (RAN) architecture to support multipoint transmission and access point densification required by the fifth generation (5G) wireless mobile systems [1], [2]. However, existing fronthaul/backhaul of Cloud-RAN cannot meet the requirements of the emerging huge data and signaling traffic in terms of transmission bandwidth requirements, stringent latency constraints and energy consumption etc. [3], [4], which has become the bottleneck of the evolution towards 5G.

Content caching in RAN can be a promising solution to significantly reduce the fronthaul/backhaul traffic [5], [6]. During off-peak times, popular content files can be transferred to the cache-enabled access points (macro base stations, small cells etc.). If the files requested by mobile users are cached in the access points of the RAN, the files will be transmitted directly from the RAN's cache without being fetched from the core network, which can significantly reduce the fronthaul/backhaul traffic and meanwhile shorten the access latency of the files, thus improve users' quality of experience (QoE). In Cloud-RAN, thanks to the ongoing evolution of fronthaul technology and function splitting between the BBU and RRHs [4], there comes possibility to realize content caching in RRHs, which allows users fetching required content files directly from RRHs and thus can further reduce fronthaul traffic.

There are two stages related with content caching: caching placement stage and delivery stage [6]. Caching placement, or known as caching strategy, is the stage to determine which files should be stored in which cache-enabled access points, and delivery stage refers to transmitting the requested files from access points to mobile users. Caching strategy is of importance because it is the initial step to perform caching and obviously it will have an impact on the performance of the delivery stage. Hence, caching strategy should be optimized by taking into consideration the wireless transmission performance. However, the wireless transmission characteristics such as fading were not considered in most of the researches when designing caching strategies [7]–[9], i.e., it was assumed that the wireless transmission is error-free.

There are some papers considering wireless fading characteristics when designing caching strategy. In [10], optimal caching placement was obtained through a greedy algorithm to minimize the average bit error rate in a macro cell with many cache-enabled helpers. In [11], cache-enabled base stations are connected to a central controller via backhaul links. Caching strategy was proposed to minimize the average download delay. In [10] and [11], the authors only considered small scale Rayleigh fading by assuming that the user has the same large scale fading at any location. However, in reality, several RRHs will jointly serve the user in Cloud-RAN, and the distance between each RRH and the user will not be the same, so it is important to consider large scale fading in wireless transmission. The works in [10] and [11] cannot be extended to the case with the consideration of large scale fading. In addition, they focused on single-objective optimization without considering the fronthaul/backhaul usage.

The aim of caching in RRHs of Cloud-RAN is to significantly reduce the fronthaul traffic. Fronthaul usage, i.e., whether the fronthaul is used, is a metric which can reflect not only the fronthaul traffic and file delivery latency but also the energy consumption of the fronthaul. For example,

This work was supported by the China Scholarship Council (CSC) and the Fundamental Research Funds of Shandong University (No. 2015ZQXM008).

lower fronthaul usage implies there are more possibilities that mobile user can access the content files in near RRHs, which will shorten the file access latency, meanwhile the fronthaul cost (i.e, the energy consumption) will be lower. On the other hand, outage probability is an important performance metric of the system, which reflects the reliability of the wireless transmission, i.e., whether the requested content files can be successfully transferred to the user. If replicas of certain content files are cached in several RRHs, the outage probability will be reduced due to the transmit diversity in wireless transmissions, while the fronthaul usage will become higher because the total number of different files cached in the RRHs are reduced. On the other hand, caching different files in the RRHs will reduce the fronthaul usage, while the outage probability will become relatively higher due to the decrease of wireless diversity. Therefore, there exists tradeoff between fronthaul usage and outage probability.

In this paper, we investigate downlink transmission in a virtual cell in Cloud-RAN. The optimal caching strategy is proposed to jointly minimize the cell average outage probability and the fronthaul usage. A realistic fading channel is adopted, which includes path loss and small scale Rayleigh fading. The major contributions of this paper are:

1) Closed form expression of outage probability conditioned on the user's location is derived, and the cell average outage probability is obtained through the composite Simpson's integration. Simulation results show that the analysis is highly accurate.
2) The joint optimization problem is formulated as a weighted sum minimization of cell average outage probability and fronthaul usage, which is NP-hard. An effective genetic algorithm (GA) based approach is proposed to solve the problem, which can achieve almost the same performance as the optimal exhaustive search, while the computational complexity is significantly reduced.

II. System Model

It is assumed that there are N cache-enabled RRHs in a circular cell with radius R, and the set of RRH cluster is denoted as $\mathcal{N} = \{1, 2, \cdots, N\}$. The file library with a total of L content files is denoted as $\mathcal{F} = \{F_1, F_2, \cdots, F_L\}$, where F_l is the l-th ranked file in terms of popularity. The popularity distribution of the files follows the Zipf's law [12], i.e., the request probability of the l-th ranked content file is

$$P_l = \frac{l^{-\beta}}{\sum_{n=1}^{L} n^{-\beta}}, \tag{1}$$

where $\beta \in [0, +\infty)$ is the skewness factor.

Considering the BBU pool can be equipped with sufficient storage space, it is assumed that all the L content files are cached in the BBU pool, and all of them have the same size. Some of the content files can be further cached in the RRHs in order to improve the system's performance. The n-th RRH can cache M_n files, and generally $\sum_{n=1}^{N} M_n < L$. The caching placement of the content files in the RRHs can be denoted by a binary placement matrix $\mathbf{A}^{L\times N}$, with the (l, n)-th entry

$$a_{l,n} = \begin{cases} 1, & \text{the } n\text{-th RRH caches the } l\text{-th file} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

indicating whether the l-th content file is cached in the n-th RRH, and $\sum_{l=1}^{L} a_{l,n} = M_n, \ \forall n$.

Single user scenario is considered in this paper. However, the proposed algorithm can be applied in practical multiuser orthogonal frequency division multiple access (OFDMA) system, in which each user is allocated with different subcarriers [13]. It is assumed that the user can only requests for one file at one time, and all the RRHs caching the requested file will serve the user. If none of the RRHs caches the requested file, the file will be transferred to all the RRHs from the BBU pool through fronthauls, and then all the RRHs transmit the file to the user. The service RRH set for the user with respect to (w.r.t.) the l-th file is denoted as $\Phi_l = \{n | a_{l,n} = 1, n \in \mathcal{N}\}, (l \in \{1, 2, \cdots, L\})$, with cardinality $|\Phi_l| \in \{1, 2, \cdots, N\}$. The system model and file delivery scheme are illustrated in Fig.1. For example, when the user requests for the l_1-th file which is not cached in any of the RRHs, the user's service RRH set is $\Phi_{l_1} = \{1, 2, 3, 4\}$. When the user requests for the l_2-th file which is already cached in RRH 2 and RRH 3 via caching placement, the service RRH set is $\Phi_{l_2} = \{2, 3\}$.

Assuming that both the RRH and the user's device are equipped with single antenna, the user's received signal from the service RRH set when requesting for the l-th file can be expressed as

$$y = \sum_{n\in\Phi_l} \sqrt{p_T d_n^{-\alpha}} h_n + noise, \tag{3}$$

where p_T is the transmit power of each RRH, d_n is the distance between the n-th RRH and the user, α is the pass loss exponent, $h_n \sim \mathcal{CN}(0, 1)$ represents complex Gaussian small scale fading, and $noise$ denotes complex additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

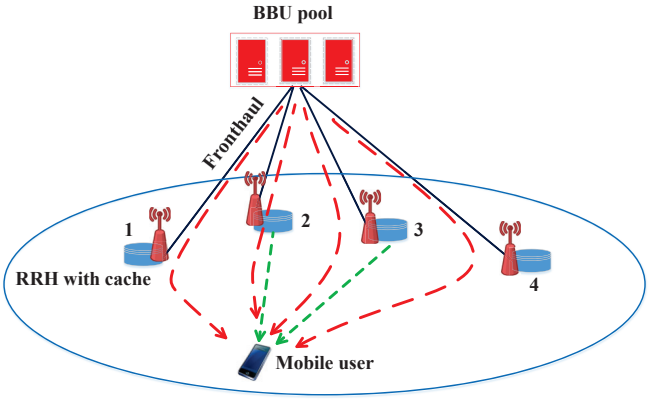

Fig. 1. System model and file delivery scheme. Red and green dashed lines represent the file fetching routes when user requests for the l_1-th and l_2-th content file, respectively.

III. Problem Formulation and Analysis

A. Problem Formulation

Define the normalized fronthaul usage w.r.t. the l-th file as

$$T_l(\mathbf{A}) = \prod_{n=1}^{N}(1-a_{l,n}) = \begin{cases} 1, & a_{l,n}=0 \text{ for } \forall n \\ 0, & \exists n \text{ such that } a_{l,n}=1 \end{cases}, \tag{4}$$

which indicates that if there is at least one copy of the requested file cached in the RRHs, there will be no fronthaul usage, i.e., $T_l = 0$, while if the requested file is not cached in any of the RRHs, there will be fronthaul usage, i.e., $T_l = 1$. Note that T_l does not depend on the user's location.

The caching strategy should be designed according to the long-term statistics over the user's locations and content file requests. The joint optimization problem can be formulated through a weighted sum of the objectives [14],

$$\min f_{obj}(\mathbf{A}) = \eta \underbrace{\sum_{l=1}^{L} P_l \mathbb{E}_{x_0}\left[P_{out}^{(l)}(x_0)\right]}_{\text{cell average outage probability}} + (1-\eta) \underbrace{\sum_{l=1}^{L} P_l T_l}_{\text{fronthaul usage}} \tag{5a}$$

$$s.t. \sum_{l=1}^{L} a_{l,n} = M_n, \tag{5b}$$

$$a_{l,n} \in \{0,1\}. \tag{5c}$$

where $\eta \in [0,1]$ is a weighting factor to balance the tradeoff between outage probability and fronthaul usage, $\mathbb{E}_{x_0}$ denotes expectation in terms of the user's location x_0, $P_{out}^{(l)}(x_0)$ is the outage probability when the user requests for the l-th file at location x_0. Constraint (5b) describes the caching limit of each RRH, and constraint (5c) indicates the joint optimization as a 0-1 integer problem.

Different values of η will lead to different balances between outage probability and fronthaul usage. Given η, the caching strategy can be determined through solving the optimization problem in (5). In practice, η is chosen by the decision maker (e.g., RAN's operator) according to the system's long-term statistics of outage probability and fronthaul usage.

B. Outage Probability Analysis

When the user requests for the l-th file at location x_0, the signal to noise ratio (SNR) of the received signal is given by

$$\gamma_l(x_0) = \sum_{n\in\Phi_l} \frac{p_T}{\sigma^2} d_n^{-\alpha} |h_n|^2 = \sum_{n\in\Phi_l} \gamma_0 S_n |h_n|^2 = \sum_{n\in\Phi_l} \gamma_n, \tag{6}$$

where $\gamma_0 = \frac{p_T}{\sigma^2}$ is SNR at the transmitter of each RRH, $S_n = d_n^{-\alpha}$ is the large scale fading, and $\gamma_n = \gamma_0 S_n |h_n|^2$ represents the received SNR from the n-th RRH. For a specific file, without ambiguity, we omit the subscript of file index l and the user's location x_0 in the following analysis.

In the service RRH set Φ with cardinality $|\Phi|$, the RRHs with the same distance to the user are grouped together. Assuming there are I ($I \le |\Phi|$) groups, the number of RRHs in the i-th group is denoted by J_i, and $\sum_{i=1}^{I} J_i = |\Phi|$. The distance between RRH and the user in the i-th group is denoted by d_i ($i \in \{1,2,3,\cdots,I\}$). Letting $\lambda_i = \frac{1}{\gamma_0 d_i^{-\alpha}}$, the probability density function (PDF) of the received SNR can be obtained as

$$f_\gamma(\gamma) = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{\lambda_i^j A_{ij}}{(j-1)!} \gamma^{j-1} e^{-\lambda_i \gamma}, \tag{7}$$

and the cumulative distribution function (CDF) is given by

$$F_\gamma(\gamma) = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{\lambda_i^{j-1} A_{ij}}{(j-1)!} \cdot \left[\frac{(j-1)!}{\lambda_i^{j-1}} - \left(e^{-\lambda_i\gamma} \sum_{k=0}^{j-1} \frac{(j-1)!}{(j-1-k)!\,\lambda_i^k} \gamma^{j-1-k}\right)\right], \tag{8}$$

where

$$A_{ij} = \frac{(-\lambda_i)^{J_i-j}}{(J_i-j)!} \frac{d^{J_i-j}}{ds^{J_i-j}} \left[M_\gamma(s)\left(1-\frac{1}{\lambda_i}\cdot s\right)^{J_i}\right]\Bigg|_{s=\lambda_i}. \tag{9}$$

The derivations of (7) and (8) are given in Appendix A.

The accuracy of the derived CDF of (8) is illustrated in Fig. 2 through three scenarios. Assuming there are 6 service RRHs for the user, and the distances between the service RRHs and the user are denoted by a vector $\mathbf{D}$. The three different scenarios are (1) scenario 1: $\mathbf{D_1} = [0.8R, 0.8R, 0.8R, 0.8R, 0.8R, 0.8R]$, ($R$ is the cell radius), i.e., all the RRHs are with the same distance to the user; (2) scenario 2: $\mathbf{D_2} = [0.6R, 0.7R, 0.7R, 0.8R, 0.8R, 0.8R]$, i.e., some of the RRHs have same distance with the user; (3) scenario 3: $\mathbf{D_3} = [0.5R, 0.6R, 0.7R, 0.8R, 0.9R, 1.0R]$, i.e., all the RRHs are with different distances to the user. It can be seen from Fig. 2 that the analytical results match the simulation results, which demonstrates the accuracy of the derived expression of (8).

The outage probability according to a certain SNR threshold γ_{th} is

$$P_{out}(\gamma_{th}) = F_\gamma(\gamma_{th}). \tag{10}$$

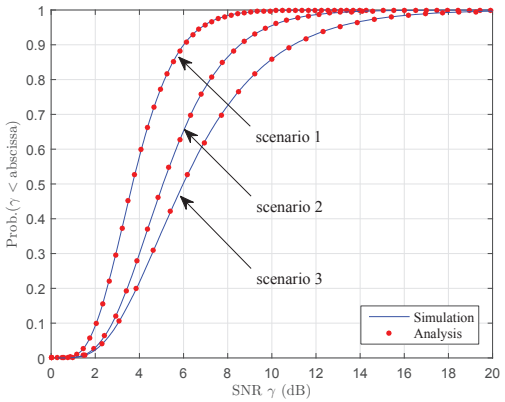

Fig. 2. CDF of the user's received SNR at a fixed location.

It is difficult to find a closed form solution of the cell average outage probability w.r.t. the l-th file, i.e., $\mathbb{E}_{x_0}[P_{out}^{(l)}(x_0)]$. However, we can use the composite Simpson's integration in forms of polar coordinates, where the user's location is denoted by (ρ, θ) and $x_0 = \rho e^{j\theta}$.

$$\begin{aligned}&\mathbb{E}_{x_0}\left[P_{out}^{(l)}(x_0)\right]\\&=\int_0^{2\pi}\int_0^{R} P_{out}^{(l)}(\rho,\theta) f_{x_0}(\rho,\theta)\rho d\rho d\theta\\&\approx\frac{\Delta h\Delta k}{9}\sum_{u=0}^{U}\sum_{v=0}^{V} w_{u,v}\rho_u P_{out}^{(l)}(\rho_u,\theta_v) f_{x_0}(\rho_u,\theta_v),\end{aligned} \tag{11}$$

where R is the cell radius, even integers U and V are chosen such that $\Delta h = R/U$ and $\Delta k = 2\pi/V$ meeting the requirement of calculation accuracy, $\rho_u = u\Delta h$, $\theta_v = v\Delta k$, $f_{x_0}(\rho, \theta)$ is the probability density function of the user's location, which is $1/\pi R^2$ when the user's location is uniformly distributed in the cell, and $\{w_{u,v}\}$ are constant coefficients given in [15].

Substituting (4), (8), (10) and (11) into (5a), the optimization problem is formulated as a function of the caching placement matrix $\mathbf{A}^{L\times N} = \{a_{l,n}\}$. However, the problem is a 0-1 integer nonlinear problem which is NP-hard, and it is difficult to obtain a closed form solution. The following section will focus on how to solve this problem.

IV. Caching Placement Schemes

In this section, firstly, the most popular content (MPC) caching placement and the largest content diversity (LCD) caching placement [11] are introduced, and we propose a location-based LCD (LB-LCD) caching placement for the tradeoff caching of outage probability and fronthaul usage in Cloud-RAN. Secondly, a genetic algorithm based approach is proposed to solve the joint optimization problem.

A. The MPC and LB-LCD Caching Placements

There are two particular caching placement schemes: one is the MPC caching, and the other one is the LCD caching. In MPC, each RRH caches the most popular files, i.e., the n-th RRH caches $\{F_l | l = 1, 2, \cdots, M_n\}$, which will have low outage probability while high fronthaul usage. In the LCD scheme, a total of $L' = \sum_{n=1}^{N} M_n$ $(< L)$ different most popular content files are cached in the RRHs, which can have lowest fronthaul usage while relatively high outage probability. If the LCD scheme is adopted in Cloud-RAN, the impact of locations of caching content files on the cell average outage probability needs to be considered. Assuming the locations of the user are uniformly distributed in the cell, caching the most popular files in the RRH nearest to the cell center will achieve better outage probability performance. Therefore, for Cloud-RAN, we improve the LCD scheme and propose a location-based LCD (LB-LCD) scheme which is described in Algorithm 1.

Algorithm 1: Proposed LB-LCD caching strategy

1 Sort the RRH set as $\mathcal{N}_s = \{n_i | i = 1, 2, \cdots, N, D_{n_1} \leq D_{n_2}, \cdots, \leq D_{n_N}\}$, where D_{n_i} denotes the distance between the n_i-th RRH and the cell center.

2 Fill the cache of the RRH set $\mathcal{N}_s$ in sequence from n_1 to n_N with content files $\{F_l | l = 1, 2, \cdots, \sum_{n=1}^{N} M_n\}$ in ascending order of l.

Since the MPC and LB-LCD caching schemes mainly focus on minimization of outage probability and fronthaul usage, respectively. In the following subsection, we propose a genetic algorithm based approach to jointly minimize the cell average outage probability and fronthaul usage.

B. Genetic Algorithm Based Approach

The genetic algorithm structure is shown in Fig. 3. Firstly, N_p candidate caching placement matrices are generated, known as the initial population (with population size N_p), and each matrix is called an individual. Then the objective value of each individual is evaluated through (5a). N_e individuals with best objective values are chosen as elites and passed into next generation (children of current generation population) directly. The rest of the next generation population are generated through crossover and mutation operations. The crossover function operates on two individuals (known as parents) and generates a crossover child, and the mutation function operates on a single individual and generates a mutation child. The number of individuals generated through crossover and mutation operations are denoted as N_c and N_m, respectively, where $N_e + N_c + N_m = N_p$, and the crossover fraction is defined as $f_c = \frac{N_c}{N_c+N_m}$. The selection function selects $2N_c$ and

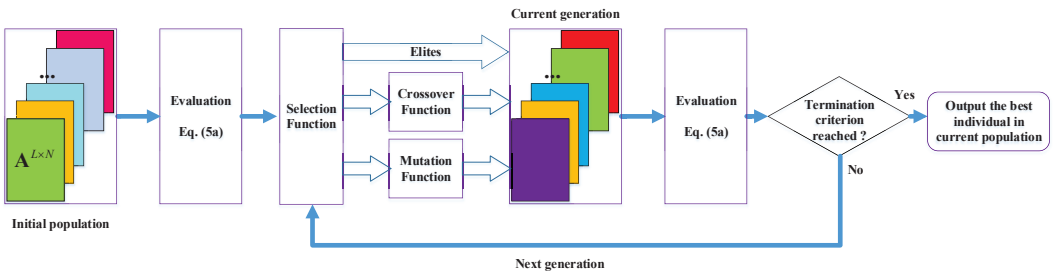

Fig. 3. Genetic algorithm structure.

N_m individuals from the current generation for the crossover and mutation function, respectively, where some individuals will be selected more than once. Stochastic uniform sampling selection [16] is adopted. Repeat the evaluation-selection-generation procedures until termination criterion is reached. Finally, the best individual in the current population is chosen as the output of the algorithm. The initial population, crossover function and mutation function of the proposed GA approach are described as follows.

1) Initial Population: The initial population is created as a set of $\{\mathbf{A}^{L\times N}\}$. For each column in each individual, M_n out of the first L' entries (i.e., $\{a_{1,n}, a_{2,n}, \cdots, a_{L',n}\}$) are set to be one randomly, and all the remaining entries are set to be zero, where

$$L' = \sum_{n=1}^{N} M_n < L \tag{12}$$

is based on the fact that the total different files with higher popularity can be cached in the RRHs are $\{F_l | l = 1, 2, \cdots, L'\}$. There is no benefit to cache files $\{F_l | l > L'\}$ with lower popularity. In addition, placement matrices of the MPC and the LB-LCD schemes are added into the initial population to further improve the performance.

2) Crossover Function: The crossover function generates a child $\mathbf{A}_c$ from parents $\mathbf{A}_1$ and $\mathbf{A}_2$. A two-point crossover function is used, which is described in Algorithm 2, in which steps 9 to 14 are heuristic operations to meet constraint (5b).

Algorithm 2: Crossover function

1 Get parent $\mathbf{A}_1 = \{a_{l,n}^{(1)}\}$ and $\mathbf{A}_2 = \{a_{l,n}^{(2)}\}$ from selection function, initialize their child $\mathbf{A}_c$ with entries $a_{l,n}^{(c)} = 0, \forall\{l, n\}$.
2 **for** $n = 1, 2, \cdots, N$ **do**
3 Generate random integers $l_1, l_2 \in [1, L')$, $l_1 \neq l_2$ according to uniform distribution.
4 **if** $l_1 < l_2$ **then**
5 Replace $a_{l,n}^{(1)},\ l = \{l_1+1, \cdots, l_2+1\}$ of $\mathbf{A}_1$ with $a_{l,n}^{(2)},\ l = \{l_1+1, \cdots, l_2+1\}$ of $\mathbf{A}_2$, and then set $a_{l,n}^{(c)} = a_{l,n}^{(1)},\ \forall l \in \{1, 2, \cdots, L\}$.
6 **else**
7 Replace $a_{l,n}^{(2)},\ l = \{l_2+1, \cdots, l_1+1\}$ of $\mathbf{A}_2$ with $a_{l,n}^{(1)},\ l = \{l_2+1, \cdots, l_1+1\}$ of $\mathbf{A}_1$, and then set $a_{l,n}^{(c)} = a_{l,n}^{(2)},\ \forall l \in \{1, 2, \cdots, L\}$.
8 **end**
9 **while** $\sum_{l=1}^{L} a_{l,n}^{(c)} > M_n$ **do**
10 Set nonzero $a_{l,n}^{(c)}$ to 0 in descending order of l.
11 **end**
12 **while** $\sum_{l=1}^{L} a_{l,n}^{(c)} < M_n$ **do**
13 Set zero $a_{l,n}^{(c)}$ to 1 in ascending order of l.
14 **end**
15 **end**

3) Mutation Function: The mutation function operates on a single individual and generates its mutation child. For each column of the individual, one of the first L' entries is randomly selected and the value is set to be the opposite (0 to 1 and vice versa), then steps 9 to 14 described in Algorithm 2 are executed to meet constraint (5b). The mutation operation reduces the probability that the algorithm converges to local minimums.

The number of objective function calculations w.r.t. a certain value of η is evaluated to measure the complexities of the proposed GA approach and the optimal exhaustive search method. The complexity of the proposed GA is $N_p N_g$, where N_p and N_g are the population size and the number of generations evaluated, respectively. The complexity of exhaustive search is $\prod_{n=1}^{N} \binom{L}{M_n}$. When $M_n = M,\ \forall n$, it is clear that the complexity of exhaustive search is exponential w.r.t. the number of RRHs, i.e., $\binom{L}{M}^N$.

V. Numerical Results

Some representative numerical results are given in this section. At first, the effectiveness of the proposed GA approach is verified by comparing its performance with exhaustive search. Then performances of different caching strategies are compared and the convergence behavior of the proposed GA is presented. Throughout the simulation, it is assumed that each RRH has the same cache size $M_n = M$. The transmit power of each RRH is $p_T = \frac{P}{N}$, where P is the total transmit power in the cell and $\frac{P}{\sigma^2} = 23$ dB. The received power attenuates 20 dB when the distance between the RRH and the user is R. In such setting, the outage probability does not depend on the absolute value of R, that is, R can be regarded as the normalized radius. The main simulation parameters are summarized in Table I.

TABLE I
Simulation parameters

Parameter	Value
Path loss exponent α	3
P/σ^2	23 dB
SNR threshold γ_{th}	3 dB
User location distribution	uniform
U and V in Simpson's integration	$6, 6$
Population size N_p in GA	50
Selection function	stochastic universal sampling
Number of elites N_e	10
Crossover fraction f_c	0.85

Fig. 4 shows the optimal tradeoffs between the cell average outage probability and the fronthaul usage with different cache size M. There are three RRHs, and the polar coordinates of which are $\left(\frac{R}{4}, 0\right)$, $\left(\frac{R}{3}, \frac{2\pi}{3}\right)$, and $\left(\frac{R}{2}, \frac{4\pi}{3}\right)$, respectively. There are $L = 9$ content files, and the popularity skewness factor $\beta = 1.5$. It can be seen from the figure that the minimum cell average outage probability is achieved at point A_1 when $M = 1$, A_2 when $M = 2$, and A_3 when $M = 3$, respectively. The corresponding caching placements of the three points are the MPC scheme. On the other hand, the minimum fronthaul usage is achieved at point B_1 when $M = 1$, B_2 when $M = 2$, and B_3 when $M = 3$, respectively. The corresponding caching placements of the three points are the LB-LCD scheme. The

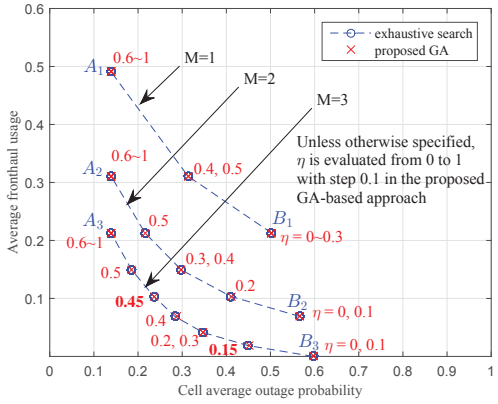

Fig. 4. Cell average outage probability and fronthaul usage tradeoff. $L = 9, M = \{1, 2, 3\}, N = 3, \beta = 1.5$.

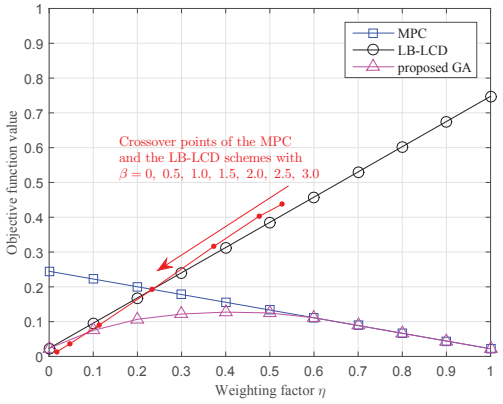

Fig. 5. Objective function value and weighting factor η. $L = 50, M = 5, N = 7$.

results obtained through the proposed GA approach is almost the same as exhaustive search, which means that the proposed GA approach can achieve near-optimal performance.

Fig. 5 shows the objective value of different caching strategies with $M = 5$. There are $L = 50$ files, $N = 7$ RRHs with one RRH located at the cell center and the other 6 RRHs evenly distributed on the circle with radius $2R/3$, and $\beta = 1.5$. It can be seen from the figure that as the weighting factor η increases, i.e., more focus on minimization of outage probability, the objective value of MPC decreases linearly, while the objective value of the LB-LCD scheme increases linearly. The horizontal coordinate of the crossover point of the MPC and LB-LCD scheme approaches zero as the popularity skewness factor β increases. That is, as β increases, the MPC scheme will dominate with most values of η. This can be explained as follows. When β increases, the average fronthaul usage will depend more and more on the few files with higher ranks. These files can be cached in the RRHs under both of the MPC and the LB-LCD schemes, thus the MPC and the LB-LCD schemes are equivalent in terms of fronthaul usage, while the MPC can achieve lower outage probability. Therefore the MPC scheme is superior to the LB-LCD scheme.

According to the above evaluations in Fig. 4 and Fig. 5, the MPC and LB-LCD caching schemes are two special solutions of the joint optimization problem when $\eta = 1$ and $\eta = 0$, respectively. The former can achieve the lowest cell average outage probability while the latter can achieve the minimum fronthaul usage. The proposed GA-based approach can achieve different tradeoffs between the cell average outage probability and fronthaul usage according to different weighting factors, which can achieve better performance than the MPC and LB-LCD schemes.

Fig. 6 shows the convergence behavior of the proposed GA approach. It can be seen from the figure that the mean objective value of the population converges within average 8 generations. The computational complexity is $N_g N_p = 8 \times 50 = 400$. While the computational complexity of the exhaustive search is $\binom{50}{5}^7 = 1.92 \times 10^{44}$, which is not feasible in practice.

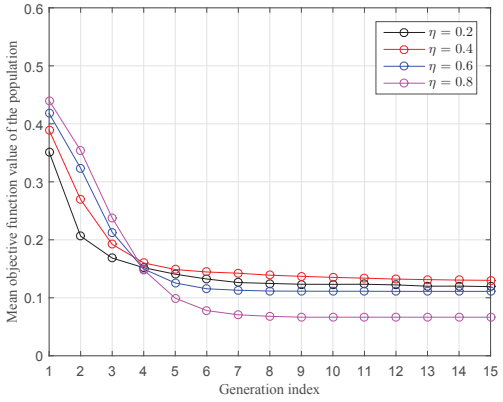

Fig. 6. Convergence behavior of the proposed GA approach. $L = 50, M = 5, N = 7, \beta = 1.5$.

VI. Conclusion

In this paper, we have investigated optimal caching strategy in Cloud-RAN for future mobile communications. In order to jointly minimize the cell average outage probability and fronthaul usage, the optimization problem is formulated as a weighted sum of the two objectives, with weighting factor η (and $1 - \eta$). Analytical expressions of outage probability has been presented and verified through simulations. Performances of two particular caching strategies have been analyzed, namely the MPC and the LB-LCD schemes. When the minimization of the cell average outage probability is more focused on, the MPC scheme is superior to the LB-LCD scheme, while the latter is superior to the former in the opposite situation, i.e., where the reduction of average fronthaul usage is more focused on. When the content files' popularity skewness factor β is larger, the MPC scheme will dominate in a wide range of η. A genetic algorithm based approach is proposed to solve the joint optimization problem, which can achieve nearly the same optimal performance of exhaustive search, while the computational complexity is significantly reduced.

APPENDIX A
DERIVATIONS OF (7) AND (8)

In (6), $|h_n|^2 \sim \chi^2(2)$, where $\chi^2(2)$ is the central Chi-squared distribution with 2 degrees of freedom, and the PDF is given by

$$f_{|h_n|^2}(x) = \exp(-x),\ x > 0. \tag{A.1}$$

Then the PDF of $\gamma_n = \gamma_0 S_n |h_n|^2$ is

$$f_{\gamma_n}(\gamma) = \frac{1}{\gamma_0 S_n} \exp\left(-\frac{\gamma}{\gamma_0 S_n}\right),\ \gamma > 0,\ n \in \Phi. \tag{A.2}$$

The moment generation function (MGF) [17] of the random variable γ_n is

$$M_{\gamma_n}(s) = \int_0^\infty f_{\gamma_n}(\gamma) e^{s\gamma} d\gamma = \frac{1}{1 - \gamma_0 S_n \cdot s}\ . \tag{A.3}$$

Since the RRHs are distributed at different locations, $\{\gamma_n, n \in \Phi\}$ is independent of each other, the MGF of received SNR $\gamma = \sum_{n\in\Phi} \gamma_n$ is given by

$$M_\gamma(s) = \prod_{n\in\Phi} M_{\gamma_n}(s) = \prod_{n\in\Phi} \frac{1}{1 - \gamma_0 S_n \cdot s}\ . \tag{A.4}$$

Since there are I distinct distances $d_1 \neq d_2 \neq \cdots \neq d_i \neq \cdots \neq d_I$ between the service RRHs and the user, and the i-th distance has multiplicity of J_i, (A.4) can be rewritten as

$$M_\gamma(s) = \frac{1}{\left(1 - \frac{1}{\lambda_1}s\right)^{J_1} \left(1 - \frac{1}{\lambda_2}s\right)^{J_2} \cdots \left(1 - \frac{1}{\lambda_I}s\right)^{J_I}}\ , \tag{A.5}$$

where $\lambda_i = \frac{1}{\gamma_0 d_i^{-\alpha}},\ i \in \{1, 2, \cdots, I\}$ is the i-th pole of multiplicity J_i of $M_\gamma(s)$, using partial fraction expansion, $M_\gamma(s)$ can be expressed as

$$M_\gamma(s) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{A_{ij}}{\left(1 - \frac{1}{\lambda_i}s\right)^j}\ , \tag{A.6}$$

where $\{A_{ij}\}$ are the undetermined coefficients. Multiplying $(1 - \frac{1}{\lambda_i}s)^{J_i}$ to both sides of (A.6), then calculating the $(J_i - j)$-th order derivate for both sides and let $s = \lambda_i$, we have

$$\begin{aligned}
&\frac{d^{J_i - j}}{ds^{J_i - j}} \left[M_\gamma(s) \left(1 - \frac{1}{\lambda_i}s\right)^{J_i} \right] \Bigg|_{s=\lambda_i} \\
=&\frac{d^{J_i - j}}{ds^{J_i - j}} \left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{A_{ij}}{\left(1 - \frac{1}{\lambda_i} \cdot s\right)^j} \left(1 - \frac{1}{\lambda_i}s\right)^{J_i} \right] \Bigg|_{s=\lambda_i} \\
=&(J_i - j)! \left(-\frac{1}{\lambda_i}\right)^{J_i - j} A_{ij}\ .
\end{aligned} \tag{A.7}$$

Thus A_{ij} is obtained as (9).

The PDF of γ can be obtained by inversely transforming the MGF in (A.6). Considering a general form of the PDF,

$$f(\gamma) = \gamma^n e^{-a\gamma},\quad \gamma \geq 0, \tag{A.8}$$

where integer $n \geqslant 0$ and a is a positive real number. The MGF of $f(\gamma)$ can be obtained by continuously using the method of integration by parts.

$$M(s) = \int_0^\infty \gamma^n e^{-a\gamma} e^{s\gamma} d\gamma = \frac{n!}{(a - s)^{n+1}}\ . \tag{A.9}$$

The CDF can be calculated in the same manner,

$$\begin{aligned}
F(\gamma) &= \int_0^\gamma \gamma^n e^{-a\gamma} d\gamma \\
&= \frac{1}{a} \left[\frac{n!}{a^n} - \left(e^{-a\gamma} \sum_{k=0}^{n} \frac{n!}{(n-k)!\, a^k} \gamma^{n-k} \right) \right]\ .
\end{aligned} \tag{A.10}$$

According to (A.6), (A.8) and (A.9), the PDF of the received SNR is obtained, as shown in (7). According to (7), (A.8) and (A.10), the CDF of the received SNR is obtained as shown in (8).

REFERENCES

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks —a technology overview," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[2] C. L. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, 2014.

[3] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: a survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[4] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: beyond baseband samples and point-to-point links," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 90–97, October 2015.

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

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

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

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

[9] S. Wang, X. Zhang, K. Yang, L. Wang, and W. Wang, "Distributed edge caching scheme considering the tradeoff between the diversity and redundancy of cached content," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, November 2015, pp. 1–5.

[10] J. Song, H. Song, and W. Choi, "Optimal caching placement of caching system with helpers," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 1825–1830.

[11] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, December 2015, pp. 1–6.

[12] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, vol. 1, March 1999, pp. 126–134.

[13] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems —part II: joint chunk, power and bit allocation," *IEEE Transactions on Communications*, vol. 60, no. 2, pp. 499–509, February 2012.

[14] M. Ehrgott, *Multicriteria Optimazation*, 2nd ed. Springer, 2005.

[15] R. L. Burden and J. D. Faires, *Numerical Analysis*, 9th ed. Brooks/Cole, Cengage Learning, 2011.

[16] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1998.

[17] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*. Hoboken: John Wiley & Sons, 2005.