Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks

Zhongyuan Zhao, Member, IEEE, Mugen Peng, Senior Member, IEEE, Zhiguo Ding, Member, IEEE, Wenbo Wang, Senior Member, IEEE, and H. Vincent Poor, Fellow, IEEE

Abstract

In cloud radio access networks (C-RANs), a huge amount of data exchanging on both backhaul and fronthaul cause high power consumption and unideal quality of service (QoS) guarantees of real-time services. To solve this problem, a cluster content caching scheme is proposed in this paper, which can take full advantages of distributed caching and centralized processing. In particular, redundant traffic on backhaul can be reduced since cluster content caches provide a part of required content objects for the remote radio heads (RRHs) connected to a common edge cloud. Then the explicate expressions of both effective capacity and energy efficiency performance are derived, which show that our proposed scheme can improve the QoS guarantees with a lower power cost of local storage. Moreover, to fully explore the potential of our studied cluster content caching schemes, the joint design of resource unit and RRH allocations is optimized, and two distributed algorithms are proposed. Finally, simulation results are provided to verify the accuracy of the analytical results and to show the performance gains resulting from cluster content caching in C-RANs.

Index Terms

Content caching, energy efficiency, effective capacity, cloud-radio access networks, resource allocation

Zhongyuan Zhao (e-mail: zyzhao@bupt.edu.cn), Mugen Peng (e-mail: pmg@bupt.edu.cn), and Wenbo Wang (e-mail: wbwang@bupt.edu.cn) are with the Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing, China. Zhiguo Ding (z.ding@lancaster.ac.uk) is with the School of Computing and Communications, Lancaster University, Lancaster, LA1 4YW, UK. H. Vincent Poor (poor@princeton.edu) is with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA.
I. INTRODUCTION

The cloud radio access network (C-RAN) is an energy-efficient network architecture for mobile operators to provide high data rate service, in which remote radio heads (RRHs) connect with a cloud-based baseband unit (BBU) via fiber fronthaul links [1]. In addition to savings on capital and operational expenditures and reducing power consumption, the centralization of BBUs enables large-scale signal processing and resource management, and the theoretical performance limits have been studied in [2]. In practical transmission situations, the capacity of both the fronthaul and the backhaul may not be able to support the significant number of low-latency data exchanges between RRHs and the BBU pool. Although some efficient signal compression methods are proposed in C-RANs, such as [3] and [4], it is insufficient to satisfy the dramatically increasing requirements from mobile users for real-time services with high quality of service (QoS) guarantees. To solve this problem, some evolved network architectures of C-RANs have been proposed, such as heterogeneous cloud radio access networks (H-CRANs) in [5] and edge cloud radio access networks (EC-RANs) in [6]. In particular, H-CRANs and EC-RANs can balance the loadings of the fronthaul and the backhaul by separating the controller from the cloud center and decentralizing the BBU pools, respectively. The performance gains of H-CRANs and EC-RANs have been evaluated in [7] and [8], which demonstrated that they have great potentials to improve both spectral and energy efficiencies.

A key network architecture feature of C-RANs and their related extensions is that the distance between a user and its access node is shortened significantly, which shows that wireless networks are evolving from a base station-centric architecture to a user-centric architecture, as well as from a connection-centric purpose to a content-centric purpose. Compared with the conventional network architectures, content-centric networks pay more attention to QoS guarantees, and some dynamic storage units, named as content caches are employed. In [9] and [10], proactive caching schemes are proposed in heterogeneous networks (HetNets), where the content caches are deployed at the edge of networks, such as the small cells and the users. The performance improvement achieved by proactive caching has been verified in [11]. However, these edge caching strategies might be not applicable in C-RANs due to the absence of BBU at the RRHs. In particular, the contents in edge caches should be transmitted to the cloud BBU pools, and then the radio frequency version of information will be sent back to the RRHs. Therefore, the content delivery
with edge caches cannot be accomplished between the RRHs and the users only, which will cause bad delay experience. In the paradigm of C-RANs, content caches are equipped at the cloud center to fully exploit the computational capability. In [14], a centralized content caching scheme was proposed, and a test content-centric network based on the architecture of C-RANs was introduced accordingly. A hierarchical content caching scheme has been proposed in [12], and the coordination between the cloud cache and the edge caches at the base stations was optimized in [13].

Although content caching schemes seem to be valid approaches to improve the QoS and the energy efficiency of wireless networks, it is not easy to find a suitable metric to evaluate the performance gains precisely, which is an interesting issue. The existing works, such as in [2], cannot characterize the impacts of content caching on the QoS in C-RANs. Based on the constant source arrival rate assumption, effective capacity has been defined as a link-level QoS metric for a wireless channel in [15]. The concept of effective capacity has been shown to be an efficient criterion to capture the delay experience of data services, which has been studied in the basic scenarios of wireless communications [16]–[18].

Content-centric C-RANs have substantial potential for future wireless communication systems, but there are some challenges: Firstly, although the existing content-centric C-RAN concept can reduce the service delay, it is not easy to decide where to deploy the content caches since it is a dilemma to keep a balance centralized processing and distributed caching. Secondly, the existing studies of QoS evaluation in wireless systems are in the interference-free scenarios, and are not suitable for analyzing the performance of content caching in C-RANs, which are typically interference limited networks. Thirdly, due to the limited storage volume, it is an open issue that how to fully explore the potential of content caches in C-RANs.

Therefore, motivated by the necessity of network architecture enhancement and QoS metric establishment, a cluster content caching scheme in C-RANs is studied in this paper, and the performance analysis and optimization algorithm design are studied as well. Our main contributions can be summarized as follows.

Firstly, a cluster content caching scheme is proposed in edge cloud-radio access networks (EC-RANs), in which RRHs in the same cluster share a common local cluster content cache. Our proposed scheme can take full advantages of centralized processing and distributed caching in C-RANs, and the capacity constraint of backhaul links is analyzed to show the improvement
of delay experience achieved by our proposed scheme.

Secondly, by formulating a stochastic geometry-based network model, both the effective capacity and the energy efficiency of the proposed cluster content caching scheme are analyzed, and tractable expressions are derived. The analytical results show that our proposed scheme can improve the QoS guarantees in an energy-efficient way, since a part of requests can be responded locally with a low cost of storage in each cluster.

Thirdly, to further improve the performance gains of cluster content caching in C-RANs, the joint design of radio resource unit (RRU) and RRH allocations is optimized in this paper, which can be modeled as a nested coalition formation game. Moreover, to reduce the computational complexity, another utility function for the RRU allocation problem is given based on the Shapley value.

The rest of this paper is organized as follows. Section II describes the system model, and the cluster content caching scheme in C-RANs is also proposed, which shows that the proposed scheme can reduce the loadings on backhaul links. In Section III, both the effective capacity and the energy efficiency are studied, and the tractable expressions will be provided. In Section IV, the joint design of RRU and RRH allocations is studied, and a nested coalition formation game-based algorithm is provided. To reduce the computational complexity, a sub-optimal algorithm is given in Section V. The simulation results are shown in Section VI, followed by the conclusion in Section VII.

II. SYSTEM MODEL AND PROTOCOL DESCRIPTION

To improve QoS guarantees of real time services, content caches are employ in C-RANs. In the conventional C-RANs, a cloud based content cache is equipped at the cloud center, where both a cloud BBU pool and a centralized controller are located as illustrated in Fig. 1(a). RRHs connect to the BBU pool via fronthaul, while the BBU connects to the content cloud through backhaul. In this fully centralized network architecture, heavy burdens are put on both backhaul and fronthaul due to a huge amount of data exchanged between cloud center and RRHs, which increases power consumption and lengthen transmission latency. To solve this problem, a paradigm of EC-RANs with cluster content caching are studied in this paper. As shown in Fig. 1(b), the loadings on both fronthaul and backhaul links can be balanced by partially decentralizing the baseband processing and the controlling functions into the edge clouds. Moreover, by employing a cluster content
cache in each cluster, the QoS guarantees and energy efficiency can be further improved, since a part of requests can be responded in the cluster immediately with a low cost on local storage.

Without loss of generality, we focus on a typical cluster $C_T$ in Fig. 1(b), where RRHs are connected to a common edge cloud, and the cluster-scale joint management, such as scheduling and resource allocations, can be implemented. Both a cluster content cache $U_T$ and a cloud content cache $U_C$ are deployed to fully exploit the potential of content caching in C-RANs. The content requests from served users are aggregated at the edge cloud in $C_T$, and can be treated by using the following strategy: First, $C_T$ checks its cluster content cache $U_T$, and the requests can be served immediately if the desired content is available at $U_T$. Otherwise, the requests will be forward to the cloud content cache, and then the corresponding content can be provided through the backhaul link from the content cloud. Then the requests can be handled similarly to the case in which the content is with the cluster content cache $U_T$.

For convenience, we assume that the cloud content cache $U_C$ stores the set of all the content objects required by the users in $C_T$, which is denoted as a limited set $\Omega_C = \{S_1, \ldots, S_L\}$, with $S_1, \ldots, S_L$ all of the same size $B_S$. Moreover, the content objects kept in the local cluster cache $U_T$, which are selected randomly from $\Omega_C$, can be treated as the members belonging to a subset of $\Omega_C$, i.e., $\Omega_T = \{S_{T_1}, \ldots, S_{T_K}\} \subseteq \Omega_C$. A radio resource unit (RRU) is defined as a set of resources in time and frequency domains that are allocated to accomplish the transmission of a specific content object through radio access channels. Each RRH can serve only one content
object in each RRU due to the single-antenna deployment. To provide a tractable expression of channel capacity for a link from RRH to user in C-RANs, the observation at a typical user \( u_T \) in \( C_T \) can be expressed as

\[
y_T = \sqrt{\rho} h_m d_m^{-\beta/2} s_m + \sum_{R_j \in I_R} \sqrt{\rho} h_j d_j^{-\beta/2} s_j + n_T,
\]

where \( s_m \) is the desired message for \( u_T \) from its serving RRH \( R_m \), the channels are modeled by including Rayleigh fading and path loss in this paper, i.e., \( h_m \) denotes the flat Rayleigh fading for the link from \( R_m \) to \( u_T \), \( d_m \) is the distance between \( R_m \) and \( u_T \), \( \beta \) is the path loss exponent, \( I_R \) denotes a set of all the interfere RRHs, \( s_j, h_j \) and \( d_j \) are defined similarly for a specific interfere RRH \( R_j \), \( j \neq m \), \( h_m, h_j \sim \mathcal{CN}(0,1) \), \( n_T \) is the additive Gaussian noise with unit covariance, and \( \rho \) is the average signal-to-noise ratio (SNR). Due to (1), the channel capacity can be expressed as follows:

\[
C = \mu W \log(1 + \gamma), \quad \text{where } \gamma = \frac{\rho d_m^{-\beta} |h_m|^2}{\sum_{R_j \in I_R} \rho d_j^{-\beta} |h_j|^2 + \sigma^2},
\]

where \( \mu \) denotes the spectral efficiency. In particular, \( \mu \) evaluates that how many content objects can be served in each RRU, i.e., \( \mu = \frac{L B_S}{(N W T)} \) bit/s/Hz when \( L \) content objects can be served in \( N \) RRUs, where the time length and the spectrum bandwidth of an RRU are denoted as \( T \) and \( W \), respectively.

As shown in (2), channel capacity is based on the ideal delay assumption, and thus cannot characterize the QoS of requested content. To provide a tractable QoS metric of a wireless channel, the concept of effective capacity is introduced as follows.

### A. The Theory of Effective Capacity in Wireless Channels

It is challenging to evaluate QoS that can be supported by a wireless link due to its inconstant transmission condition. In [15], the concept of effective capacity is provided as a feasible solution of this problem. In particular, assuming there exists a queue of infinite size with a constant arrival rate at the source, effective capacity can characterize the maximum arrival rate that can be supported by a wireless channel with a specific QoS guarantee, which is defined as follows [15]:

\[
E(\theta) = -\lim_{t \to \infty} \frac{1}{\theta t} \log \mathbb{E}\{e^{-\theta S(t)}\},
\]
Fig. 2. Queueing model of content transmissions in C-RANs.

where $S(t) = \sum_{0=t_0 < t_1 < \cdots < t_n = t} \int_{t_{i-1}}^{t_i} r(\tau) d\tau$ is the delivered service through a wireless channel in bits over the time interval $[0, t)$, $r(\tau)$ denotes the channel capacity at time $\tau$, and $\theta$ is the QoS exponent. To specify the delay constraint, $\theta$ is defined as the decay rate of the tail distribution with a stochastic queue length $Q$,

$$\theta = \lim_{q \to \infty} \frac{\log \Pr\{Q > q\}}{q}.$$  

(4)

For a large value of threshold $q_{\text{max}}$, the buffer violation probability can be approximated as $\Pr\{Q > q_{\text{max}}\} \approx e^{-\theta q_{\text{max}}}$, and the delay violation probability can be bound as $\Pr\{D > d_{\text{max}}\} \leq c \sqrt{\Pr\{Q > q_{\text{max}}\}}$, where $c$ is a positive constant related to arrival rate [15]. Therefore, a smaller $\theta$ implies a looser QoS requirement, while larger $\theta$ denotes a more strict QoS requirement. Due to (3) and (4), effective capacity can show the relationship between delay experience and wireless channel capacity.

Under the block fading channel assumption, each channel coefficient is a constant in each RRU, and the effective capacity can be further derived as [17]

$$E(\theta) = -\frac{1}{\theta W} \ln \mathbb{E}\{(1 + \gamma)^{-\mu \theta W T}\},$$  

(5)

where $\tilde{T} = T / \ln 2$.

**B. Capacity Constraint of Backhaul in Cluster Content Caching Scheme**

Since some objects can be obtained locally, cluster content caching improves the QoS guarantees by reducing delay, and migrates the loading on backhaul links. As shown in Fig. 2, our studied content objects are sent to users via multi-hop tandem links. In particular, a content object can be transmitted to the edge cloud in each C-RAN cluster through a wired backhaul link or from a local cluster content cache, whose arrival data rates are fixed and can be denoted as $r_{\text{BH}}$ and $r_{\text{CC}}$, respectively. Then it is delivered to RRHs through fiber fronthaul links, and
finally can be sent to the users by using wireless channels. The corresponding experienced delay can be given as the following proposition.

**Proposition 1:** (Proposition 5, [18]) Assume that a network carries packetized traffic, and consists of $N_h$ hops. Given an external arrival process with constant data rate $r$ and constant packet size $B$, the end-to-end delay $D$ experienced by the traffic can be expressed as

$$
\lim_{D_{\max} \to \infty} \frac{\log \Pr\{D > D_{\max}\}}{D_{\max} - (N_h B)/r} = -\theta,
$$

(6)

where $\theta$ denotes the QoS exponent defined by (4), and the effective capacity satisfies $E(\theta/r) = r$. Based on (6), the following approximation can be obtained for a large delay threshold $D_{\max}$:

$$
\Pr\{D > D_{\max}\} \approx e^{-\theta(D_{\max} - \frac{N_h B}{r})},
$$

(7)

Proposition 1 implies that a higher arrival data rate $r$ leads to a smaller QoS exponent $\theta$, and improves delay experience. The QoS exponents of a content object $S_l$ that is responded by the cluster content cache $U_T$ and the cloud cache $U_C$ are denoted as $\theta^T_l$ and $\theta^C_l$, respectively. Based on Proposition 1, the following theorem of the relationship between $\theta^T_l$ and $\theta^C_l$ can be provided.

**Theorem 2:** To achieve the same delay experience, the QoS exponents of content cloud and cluster caching to deliver a specific content object $S_l$ should satisfy the following condition:

$$
\theta^C_l = \frac{1}{1 - \frac{2B}{rBH} \theta^T_l}.
$$

(8)

**Proof:** Based on (6), the QoS exponent of cloud content caching can be expressed as

$$
\theta^C_l = -\lim_{D_{\max} \to \infty} \frac{\log \Pr\{D > D_{\max}\}}{D_{\max} - (2B)/r_{BH}},
$$

(9)

Similarly, the QoS exponent of cluster content caching can be given as

$$
\theta^T_l = -\lim_{D_{\max} \to \infty} \frac{\log \Pr\{D > D_{\max}\}}{D_{\max} - (2B)/r_{CC}} \approx -\lim_{D_{\max} \to \infty} \frac{\log \Pr\{D > D_{\max}\}}{D_{\max}},
$$

(a)

(10)

where the approximation (a) in (10) follows the fact that the arrival rate $r_{CC}$ goes infinity since the content can be obtained locally in the C-RAN cluster, and the delay caused by arrival process can be ignored. When both the cloud and the cluster caching schemes share a common delay experience, $\theta^C_l$ in (9) can be further derived as

$$
\frac{1}{\theta^C_l} = -\lim_{D_{\max} \to \infty} \frac{D_{\max} - (2B)/r_{BH}}{\log \Pr\{D > D_{\max}\}} = \frac{1}{\theta^T_l} + \lim_{D_{\max} \to \infty} \frac{(2B)/r_{BH}}{\log \Pr\{D > D_{\max}\}}.
$$

(11)
Recalling (7) in Proposition 1, the probability of delay $D$ exceeding a threshold $D_{\text{max}}$ can be given as $\lim_{D_{\text{max}} \to \infty} \Pr\{D > D_{\text{max}}\} \approx e^{-\theta_{l}^{T} D_{\text{max}}}$. Then (11) can be given as

$$\frac{1}{\theta_{l}^{T}} = \frac{1}{\theta_{l}^{T}} + \frac{(2B_{S})/r_{\text{BH}}}{\log e^{-\theta_{l}^{T} D_{\text{max}}}} = \left(1 - \frac{2B_{S}}{r_{\text{BH}} D_{\text{max}}}\right) \frac{1}{\theta_{l}^{T}}. \quad (12)$$

And the proof has been finished.

Theorem 2 shows that the cluster content caching can reduce the loading on the backhaul links. In particular, due to (8), the capacity of backhaul links must satisfy the following constraint:

$$c_{\text{BH}} \geq r_{\text{BH}} = \frac{2B_{S}}{D_{\text{max}}(1 - \frac{\theta_{l}^{T}}{\theta_{l}^{T}})}. \quad (13)$$

This implies that the only way to reduce the performance gap of QoS guarantee with and without cluster content caching is to enlarge the capacity of backhaul link. In particular, as the QoS guarantee of cloud content caching approaches that of cluster content caching, i.e., $\frac{\theta_{l}^{T}}{\theta_{l}^{T}} \to 1$, it requires that the backhaul link capacity goes infinity. Therefore, the deployment of cluster content caches can migrate the loadings of backhaul links, and improve the QoS guarantees.

Although some non-cacheable contents cannot be stored locally, our studied cluster caching scheme still can improve QoS guarantees with low power consumption in practical wireless networks, where both cacheable and non-cacheable contents are requested. In particular, by storing the some popular cacheable contents in local cluster caches, the delay caused by obtaining non-cacheable contents from cloud content cache can be reduced due to the loading mitigation of the links between local cloud centers and the data data centers, and thus better QoS guarantees of non-cacheable contents can be supported.

III. PERFORMANCE ANALYSIS BASED ON STOCHASTIC GEOMETRY

To evaluate the performance of cluster content caching in C-RANs, the effective capacity is studied in this section. To provide an accurate analytical model, the locations of RRHs are modeled as a homogenous Poisson point process (PPP) $\Psi_{\text{R}}$ with a given density $\lambda_{\text{R}}$. The locations of users are modeled as a homogenous marked PPP $\Phi_{\text{u}}(M_{n})$ with a given density $\lambda_{\text{u}}$, where the mark $M_{n}$ is the type of content that the $n$-th user $U_{n}$ in $\Phi_{\text{u}}(M_{n})$ requests, and required content objects of users are independent with each other. Compared with other complicated point processes, homogeneous PPP is more suitable for characterizing the seamless coverage of C-RANs, while the edge effect exists in the clustering point process. Moreover, the difference
between homogeneous PPP and clustering point process is their densities of points only. Expect
the derivation of density functions, the analysis of all these point processes will follow a similar
mathematical paradigm, and our derived analytical results can be extended to other PPP model
by replacing the density functions. Note that the caching coordinations are implemented by
sharing a common cache in each cluster, while the coordinations of signal processing, such as
coordinated multiple point (CoMP) transmissions and network beamforming, are not considered.
The reason that we consider a simple case without joint signal processing is to provide some
meaningful insights of content caching in C-RANs.

A. Effective Capacity of A Typical User Associated with A Specific RRH

Each user in $C_T$ can be treated as a typical user since its co-channel interference follows an
identical distribution based on the introduced PPP model. Without loss of generality, we consider
$U_i$ in $C_T$ as a typical user, which accesses a given RRH that serves its requested content.
Recalling (5), the key step of analyzing the effective capacity is to study the expectation of
$Z = (1 + \gamma)^{-\mu\theta_j T}$, which can be derived by using the following approximation. In particular,
the range of receive signal-to-interference-plus-noise ratio (SINR) $\gamma$ can be divided into $N$
disjoint intervals, i.e., the $n$-th intervals can be expressed as $I_n = [\gamma_n, \gamma_{n+1}]$, $n = 1, \ldots, N$,
$0 = \gamma_1 < \cdots < \gamma_{N+1} = \gamma_{\max} < \infty$, and $\gamma$ can be represented by the $n$-th typical value $\bar{\gamma}_n$
when $\gamma \in I_n$. Then $\mathbb{E}\{Z\}$ can be approximately expressed as

$$\mathbb{E}\{Z\} \approx \sum_{n=1}^{N} \left( \Pr\{\gamma < \gamma_{n+1}\} - \Pr\{\gamma < \gamma_n\}\right) (1 + \bar{\gamma}_n)^{-\mu\theta_j W T}. \tag{14}$$

This approximation is similar to the scalar quantization of analog signals, and the optimum
setting of typical value is $\bar{\gamma}_n = (\gamma_n + \gamma_{n+1})/2$ based on the quantization theory [30]. Based on
(14), an explicit expression of effective capacity can be given as the following theorem.

**Theorem 3:** Considering a typical user $U_i$, which accesses a specific RRH $R_m$ that provides
its required content $S_j$, its effective capacity can be expressed as

$$E_{i,m}(\theta_j, d_m) = -\frac{1}{\theta_j W T} \ln(G(\theta_j, d_m)), \tag{15}$$

where $\theta_j$ is the QoS exponent of $S_j$, $d_m$ is the distance between $U_i$ and $R_m$.
$G(\theta_j, d_m)$ is given as

$$G(\theta_j, d_m) = \sum_{n=1}^{N} \left( e^{-2\pi A(\beta)\gamma_n^2 \lambda_R d_m^2} - \frac{\gamma_n d_m^2 \sigma^2}{\rho} - e^{-2\pi A(\beta)\gamma_{n+1}^2 \lambda_R d_m^2} - \frac{\gamma_{n+1} d_m^2 \sigma^2}{\rho} \right) (1 + \bar{\gamma}_n)^{-\mu\theta_j T}, \tag{16}$$
\( A(\beta) = \frac{1}{\beta} \Gamma\left(\frac{2}{\beta}\right) \Gamma\left(1 - \frac{2}{\beta}\right) \), and \( \Gamma(x) \) denotes the gamma function.

**Proof:** To obtain a tractable expression of typical user effective capacity, the outage probability \( \Pr\{\gamma < \gamma_n\} \) is required. By following a similar mathematical paradigm in [19], it can be derived as

\[
\Pr\{\gamma < \gamma_n\} = \mathbb{E}_{\Psi_R, h_m, R_j \in \Psi_R/\{R_m\}} \left\{ \Pr \left\{ \frac{|h_m|^2}{\rho} < \frac{\gamma_n d_m^\beta}{\sigma^2} (I + \sigma^2) \right\} \right\} 
= 1 - \mathbb{E}_{\Psi_R, R_j \in \Psi_R/\{R_m\}} \left\{ e^{-\frac{\gamma_n d_m^\beta}{\rho} (I + \sigma^2)} \right\} 
= 1 - e^{-\frac{\gamma_n d_m^\beta \sigma^2}{\rho}} \mathbb{E}_{\Psi_R} \left\{ \prod_{R_j \in \Psi_R/\{R_m\}} \frac{1}{1 + \gamma_n d_m^\beta \gamma_j^{-d_j^{-\beta}}} \right\}, \quad (17)
\]

where \( I = \sum_{R_j \in \Psi_R/\{R_m\}} \rho d_j^{-\beta} |h_j|^2 \) denotes the co-channel interference, and \( |h_i|^2 \) of each channel follows independent and identically exponential distribution. Based on the probability generating functional (PGFL) of PPP, \( K_1 \) can be expressed as

\[
K_1 = \prod_{l=1}^L e^{-\frac{\gamma_n d_m^\beta \sigma^2}{\rho} J_1}, \quad (18)
\]

where

\[
J_1 = \exp \left[ -2\pi \lambda_R \int_0^\infty \left( 1 - \frac{1}{1 + \gamma_n d_m^\beta \gamma_j^{-d_j^{-\beta}}} \right) d_j dd_j \right] 
= \exp \left( -2\pi \lambda_R \gamma_n^\frac{\beta}{2} d_m^2 \int_0^\infty \frac{y}{y^{\beta} + 1} dy \right) = e^{-2\pi \lambda_R \frac{A(\beta) \gamma_n^2 d_m^2}{\rho}}. \quad (19)
\]

Based on (17), (18), and (19), \( \Pr\{\gamma < \gamma_n\} \) can be given as

\[
\Pr\{\gamma < \gamma_n\} = 1 - e^{-2\pi A(\beta) \frac{2}{\gamma_n^2 \lambda_R d_m^\beta \sigma^2}} \gamma_n d_m^\beta \gamma_j^{-d_j^{-\beta}}, \quad (20)
\]

Substituting (20) into (14), Theorem 1 is proved.

The approximation of (14) is a solid approach based on quantization theory, which has been widely applied in signal compression. The accuracy of quantization is mainly decided by the squared-error distortion, which can be improved by dividing smaller intervals. In this paper, the quantization intervals are divided equally, and the simulation results in Section VI show that the analytical results can match with the Monte Carlo results perfectly when the maximum value of \( \gamma \) is set as \( \gamma_{max} = 5 \times 10^4 \) and the number of intervals is \( N = 10^6 \), respectively.
B. Average Effective Capacity of A Typical Cluster $C_T$

The average effective capacity is proposed as a metric to evaluate the performance of cluster content caching in C-RANs. In this part, each user is required to access its nearest RRH that provides its desired content object so that the received signal power can be maximized. First, the hit ratio of the cluster content cache in $C_T$ is denoted as $P_{hit}$, which characterizes the probability that the requests can be satisfied at $U_T$ [28]

$$P_{hit} = \frac{\text{Number of requests served by } U_T}{\text{Total number of user requests}} = \sum_{l \in U_T} P_l,$$

(21)

where $P_l$ is the probability that the content object $S_l$ is requested by the users, which can capture the popularity of $S_l$. Note that the summation of popular ratios is 1, and thus the hit ratio follows the constraint $P_{hit} = \sum_{l \in U_T} P_l \leq \sum_{l \in \Omega_T} P_l = 1$, which is a probability measure. And the average effective capacity of content object $S_l$ can be expressed as

$$\bar{E}_l = P_{hit} \bar{E}(\theta^T_l) + (1 - P_{hit}) \bar{E}(\theta^C_l),$$

(22)

where $\bar{E}(\theta^T_l)$ and $\bar{E}(\theta^C_l)$ are the average effective capacity of content object $S_l$ that is served by $U_T$ and $U_C$, respectively. Due to the differences in the popularity of content objects, the average effective capacity of $C_T$ can be given as $\bar{E}_T = \sum_{l=1}^{L} P_l \bar{E}_l$. Moreover, RRHs belonging in $\Psi_R$ can be divided into $L$ disjoint partitions to serve $L$ different content objects. The formulation of an RRH set $\Psi_l$ that serves $S_l$ can be treated as an independent thinning process of a homogenous PPP $\Psi_R$, which is also a homogenous PPP with a fixed density $\lambda_l$, and $\sum_{l=1}^{L} \lambda_l = \lambda_R$.

Based on the aforementioned assumptions and the results given in Theorem 1, tractable expressions of $\bar{E}(\theta^T_l)$ and $\bar{E}(\theta^C_l)$ in (22) can be obtained, and the following corollary of average cluster effective capacity is provided.

Corollary 4: When each user accesses to the nearest RRH that serves its desired content object, the average effective capacity of a typical cluster $C_T$ can be written as

$$\bar{E}_T = P_{hit} \sum_{l=1}^{L} \bar{E}(\theta^T_l) + (1 - P_{hit}) \sum_{l=1}^{L} \bar{E}(\theta^C_l),$$

(23)

where $\bar{E}(\theta^T_l)$ and $\bar{E}(\theta^C_l)$ can be expressed as

$$\bar{E}(\theta_l) = P_l \sum_{n=1}^{N} \left[ \mathcal{L}_i(\gamma_n) - \mathcal{L}_i(\gamma_{n+1}) \right] (1 + \gamma_n)^{-\mu W_T}, \quad \theta_l = \theta^T_l, \theta^C_l,$$

(24)
\( P_l \) denotes the popularity of \( S_l \), and \( L_l(\gamma_n) \) can be given as

\[
L_l(\gamma_n) = 1 - 2\pi \lambda_l \int_0^\infty d_m e^{-(2\pi A(\beta)\gamma_n^2 (\lambda_R - \lambda_l) + \pi \lambda_l u(\gamma_n, \beta) + \pi \lambda_l)} d_m e^{\frac{-\gamma_m d_m^2 \sigma^2}{\rho}} d\gamma,
\]

where \( u(\gamma_n, \beta) = \gamma_n^{2/\beta} \int_{\gamma_n^{-2/\beta}}^{\infty} (1 + x^{2/\beta})^{-1} dx \). In an interference limited C-RAN scenario, an explicit expression for \( L_l(\gamma_n) \) is given by

\[
L_l(\gamma_n) = 1 - \frac{1}{2A(\beta)\gamma_n^{2/\beta}(q_l - 1) + u(\gamma_n, \beta) + 1},
\]

where \( q_l = \lambda_R / \lambda_l \).

**Proof:** Due to the definition of \( \bar{E}(\theta^T_l) \) and \( \bar{E}(\theta^C_l) \) in (22), they can be expressed as

\[
\bar{E}(\theta^T_l) = \sum_{i,m} \left\{ E_{l,m}(\theta^T_l, d_m) \right\}, \quad \theta^T_l = \theta_l^T, \theta_l^C.
\]

Then the key step is to derive a tractable expression for \( L_l(\gamma_n) \) in (24). Unlike Theorem 3, each user is required to access its nearest serving RRH, which establishes a constraint of the locations of RRHs serving \( S_l \). Then \( \mathcal{K}_1 \) in (18) should be expressed as

\[
\mathcal{K}_1 = e^{-\frac{\gamma_m^d \sigma^2}{\rho}} \mathcal{J}_2 \prod_{k \neq l} \mathcal{J}_1,
\]

where \( \mathcal{J}_1 \) is given in (19), and \( \mathcal{J}_2 \) can be expressed as

\[
\mathcal{J}_2 = \exp \left[ -2\pi \lambda R \int d_j \left( 1 - \frac{1}{1 + \gamma^m d_m^2 d_j^2} \right) d_j d_j \right] = e^{-\pi \lambda_l u(\gamma_n, \beta) d_m^2}.
\]

The probability density function (PDF) of \( d_m \) can be given as \( f(d_m) = 2\pi \lambda_l d_m e^{-\pi \lambda_l d_m^2} \), and \( L_l(\gamma_n) \) can be further derived as follows,

\[
L_l(\gamma_n) = \Pr \{ \gamma < \gamma_n \} = 1 - 2\pi \lambda_l \int_0^\infty d_m e^{-(2\pi A(\beta)\gamma_n^2 (\lambda_R - \lambda_l) + \pi \lambda_l u(\gamma_n, \beta) + \pi \lambda_l)} d_m e^{\frac{-\gamma_m d_m^2 \sigma^2}{\rho}} d\gamma.
\]

In an interference limited scenario, the impacts of noise can be ignored, and thus \( \sigma^2 \) in (29) can be set as zero. Then \( L_l(\gamma_n) \) can be written as

\[
L_l(\gamma_n) = 1 - 2\pi \lambda_l \int_0^\infty d_m e^{-(2\pi A(\beta)\gamma_n^2 (\lambda_R - \lambda_l) + \pi \lambda_l u(\gamma_n, \beta) + \pi \lambda_l)} d_m d_m = 1 - \frac{1}{2A(\beta)\gamma_n^{2/\beta}(q_l - 1) + u(\gamma_n, \beta) + 1}.
\]

And Corollary 4 has been proved.

When \( P_{hit} = 0 \), \( \bar{E}_T = \sum_{l=1}^L \bar{E}(\theta^C_l) \) is the effective capacity of no-cluster content caching strategy in C-RANs. As introduced previously, the QoS exponents follow a constraint \( \theta^L_l \leq \theta^C_l \).
Then the performance improvement of effective capacity achieved by cluster content caching scheme in C-RANs can be written as follows based on Corollary 4:

\[
\Delta \bar{E}_T = \bar{E}_T - \bar{E}_T|_{P_{hit}=0} = P_{hit} \left[ \sum_{l=1}^{L} (\bar{E}(\theta^T_l) - \bar{E}(\theta^C_l)) \right]
\]

\[
= P_{hit} \left\{ \sum_{l=1}^{L} P_l \left[ \sum_{n=1}^{N} (\mathcal{L}(\gamma_n) - \mathcal{L}(\gamma_{n+1})) \left( (1 + \bar{\gamma}_n)^{-\mu \theta^T_l} - (1 + \bar{\gamma}_n)^{-\mu \theta^C_l} \right) \right] \right\}.
\]

(31)

As shown in (31), the performance gains \(\Delta \bar{E}_T\) increases as the hit ratio of local cluster caching \(P_{hit}\) increases. In particular, when all content can be served by the cluster content cache \(\mathcal{U}_T\), i.e., \(M = L\) in (21), the local content caching can approach its performance limits. Assuming that all the required content objects have the same QoS exponents of cloud and cluster content caching schemes in \(C_T\), \(\mathcal{U}_T\) should choose to keep the most popular content objects in its coverage area, which contribute more to the improvement of effective capacity.

**C. Energy Efficiency Performance of A Typical Cluster \(C_T\)**

The average effective capacity to the average power consumption ratio of a typical cluster \(C_T\) with a cluster content cache is defined as follows, which can characterize the energy efficiency with a specific QoS requirement,

\[
\eta_T = \frac{\bar{E}_T}{\bar{P}_T} = \frac{P_{hit} \sum_{l=1}^{L} \bar{E}(\theta^C_l) + (1 - P_{hit}) \sum_{l=1}^{L} \bar{E}(\theta^T_l) }{\lambda R \pi r_T^2 P_R + KP_{CC} + (1 - P_{hit})P_{BH}},
\]

(32)

where \(\bar{E}_T\) denotes the average effective capacity of \(C_T\) given in Corollary 4, \(\bar{P}_T\) is the average total power consumption with cluster content caching, \(P_R\) denotes the power consumption of radio transmission and baseband processing for each RRH, \(r_T\) is the radius of considered cluster \(C_T\), \(P_{CC}\) is the power consumption of keeping a content object in \(\mathcal{U}_T\), \(K\) is the size of \(\mathcal{U}_T\), and \(P_{BH}\) is the power consumption of obtaining desired content objects through backhaul. Similarly, the average effective capacity to the average power consumption ratio of without cluster content caching can be expressed as follows when \(P_{hit} = 0\)

\[
\eta_T|_{P_{hit}=0} = \frac{\bar{E}_T|_{P_{hit}=0}}{\bar{P}_T|_{P_{hit}=0}} = \frac{\sum_{l=1}^{L} \bar{E}(\theta^C_l) }{\lambda R \pi r_T^2 P_R + P_{BH}}.
\]

(33)

And the comparison result of power consumption between two different schemes is

\[
\bar{P}_T - \bar{P}_T|_{P_{hit}=0} = KP_{CC} - P_{hit}P_{BH} \leq 0.
\]

(34)
The last inequity in (34) follows the fact $P_{CC} \ll P_{BH}$, i.e., $P_{BH}$ is larger than 10 W [20], while $P_{CC}$ is smaller than 1 W [13]. Moreover, recalling (31), it shows that our studied cluster content caching scheme can achieve higher effective capacity. Therefore, employing cluster caches is an energy efficient approach to improve QoS guarantees in C-RANs.

IV. JOINT DESIGN OF RRU AND RRH ALLOCATIONS BASED ON A NESTED COALITION FORMATION GAME

Based on the study in Section III, the QoS guarantees of cluster content caching are impacted by the radio resource allocation strategy, such as the serving RRH allocation and the RRU allocation. The management of the two categories of radio resource can be modeled as a coupled integer programming problem, which is an NP-hard problem. To provide an efficient solution, both serving RRH allocation and RRU allocation can be formulated as coalition formation games, and the joint design of them can be modeled as a nested coalition formation game.

A. Serving RRH Allocation Strategy

Assuming that there exists a subset of content set $\Omega_C$, i.e., $\Omega_j = \{S_{j_1}, \ldots, S_{j_M}\}$, whose members share the same RRU $RU_j$, then RRHs in $C_T$ will form $M$ disjoint sets, which are denoted as $R_{j_1}, \ldots, R_{j_M}$ to transmit the corresponding content objects in $\Omega_j$. Denoting $\mathcal{W}_{jm}$ as the set of users that require $S_{jm}$ in $C_T$, then the expected effective capacity of content object $S_{jm}$, which is served by RRHs in $R_{jm}$, can be expressed as follows when the location information of the members in $R_{jm}$ and $\mathcal{W}_{jm}$ are available:

$$
\bar{E}(R_{jm}) = \begin{cases} 
\sum_{U_i \in \mathcal{W}_{jm}} E(\theta_{T_{jm}}, d_i), & S_{jm} \text{ is in cluster content cache } \mathcal{U}_T, \\
\sum_{U_i \in \mathcal{W}_{jm}} E(\theta_{C_{jm}}, d_i), & S_{jm} \text{ is in cloud content cache } \mathcal{U}_C 
\end{cases}
$$

(35)

where $E(\theta_{T_{jm}}, d_i)$ and $E(\theta_{C_{jm}}, d_i)$ follow the effective capacity of a specific user given in Theorem 3, $d_i$ denotes the distance between $U_i$ and its serving RRH in $R_{jm}$, and $\theta_{T_{jm}}$ and $\theta_{C_{jm}}$ are the QoS exponents of content object $S_{jm}$ served by local cluster cache and cloud cache, respectively. As shown in (35), the QoS guarantees of users in $\mathcal{W}_{jm}$ are mainly decided by RRHs in $R_{jm}$, which have great impacts on the path loss that characterized by $d_i$. 

1) Utility Function Formulation: The RRHs can be treated as a scarce resource in the studied C-RAN scenario, which are competed for by the users to improve their QoS guarantees. Therefore, the serving RRH allocation can be modeled as a coalition formation game, in which RRHs in $C_T$ can be treated as players, and each RRH allocation result can be denoted as a partition $\Pi = \{R_{j_1}, \ldots, R_{j_M}\}$. The payoff of the $k$-th RRH $R_k$ to join $R_{jm}$ can be defined as the effective capacity improvement achieved by its participation, and the corresponding utility function is given as follows:

$$
\phi_k(R_{jm}) = \begin{cases} 
\bar{E}(R_{jm} \cup R_k) - \bar{E}(R_{jm}) - c_{RH} \left( P_R + \frac{1}{O(R_{jm})} P_{CC} \right), & S_{jm} \text{ is in } U_T \\
\bar{E}(R_{jm} \cup R_k) - \bar{E}(R_{jm}) - c_{RH} \left( P_R + \frac{1}{O(R_{jm})} P_{BH} \right), & S_{jm} \text{ is in } U_C 
\end{cases},
$$

(36)

where $O(A)$ denotes the number of elements in set $A$, $\tau_k$ in (36) is the cost of RRH $k$ to join $R_{jm}$, which is determined by the total power consumption. In particular, $P_R$ denotes the power consumption of an RRH spending on radio frequency and baseband processing, $P_{CC}$ and $P_{BH}$ follow the denotation of (32), and $c_{RH}$ is the energy efficiency coefficient to control the impact of the cost part $\tau_k$. Note that each content object is required by multiple RRHs in $C_T$, and thus the cost of acquiring it should be split among the corresponding RRHs. Then the total utility of each coalition can be written as

$$
v(R_{jm}) = \sum_{R_k \in R_{jm}} \phi_k(R_{jm}).
$$

(37)

2) A Distributed Coalition Formation Algorithm: Based on the utility function given in (36), the payoff of $R_k$ to join $R_{jm}$ is decided by only the members in $R_{jm}$. Thus the proposed RRH allocation problem can be modeled as a hedonic coalitional formation game, and the coalition formation process is accomplished by using preference relations. Based on the definitions in [21], the preference relation indicates that an RRH strictly prefers to join a specific coalition rather than the other one. For example, consider two coalitions $R_{jm}$ and $R_{jn}$, $R_k$ will choose to join $R_{jm}$ if the two given coalitions satisfy the preference relation $R_{jm} \succ_k R_{jn}$. Due to [21], the criteria of preference relation determination can be established as follows:

$$
R_{jm} \succ_k R_{jn} \iff \begin{cases} 
C1: \phi_k(R_{jm}) > \phi_k(R_{jn}) \\
C2: v(R_{jm}) + v(R_{jn} \setminus \{R_k\}) > v(R_{jn}) + v(R_{jm} \setminus \{R_k\})
\end{cases}.
$$

(38)
Algorithm 1 (A hedonic coalition formation-based serving RRH allocation algorithm)

**Initialization:** For a given subset of content objects $\Omega_j^{\text{sub}} = \{S_{j1}, \ldots, S_{jM}\}$, all RRHs in $C_T$ can be divided into $M$ disjoint coalitions $\mathcal{R}_j, \ldots, \mathcal{R}_{jM}$.

**Repeat:** For each RRH coalition $\mathcal{R}_{jm}$

1. Each RRH in $\mathcal{R}_{jm}$, i.e., $R_k$, negotiates with other coalitions $\mathcal{R}_{jn}$, $j_m \neq j_n$, and obtains the individual and the total coalition utility values based on (36) and (37), respectively;
2. If the obtained utility values satisfy the preference relation criteria given in (38), then $R_k$ leaves $\mathcal{R}_{jm}$ and joins $\mathcal{R}_{jn}$;

**Termination:** When the members of $\mathcal{R}_{j1}, \ldots, \mathcal{R}_{jM}$ do not change.

where C1 implies that $R_k$ pursues higher individual payoff, while C2 can avoid the loss of total utility caused by the move of $R_k$.

The conditions in (38) ensure that each C-RAN cluster can achieve considerable QoS performance gains via a coalition formation game-based RRH allocation strategy, and the corresponding distributed RRH allocation algorithm can be described in Algorithm 1. In particular, RRHs are assigned to serve the content objects $S_{j1}, \ldots, S_{jM}$ in the initialization phase, and RRHs take turns to negotiate with other potential coalitions, and determine whether to join a new one based on the criteria given in (38) until the obtained partition results converge. The following proposition asserts that the proposed algorithm will converge to a Nash stable partition.

**Proposition 5:** (Theorem 1, [21]) Starting from any initial partition, Algorithm 1, which is based on a hedonic solution, will always converge to a Nash stable partition.

**B. RRU Allocation Strategy**

Besides the serving RRHs allocation, the performance of radio-access channels is also impacted by the RRU allocation strategy. Similar to RRH allocation, RRU allocation can also be formulated as a coalition formation game. In particular, the content objects $S_1, \ldots, S_L$ can be treated as players, which negotiate with each other to form disjoint coalitions, and the content objects in the same coalition share a common RRU.

1) **Utility Function Formulation:** Without loss of generality, we focus on a coalition $\mathcal{T}_i = \{S_{i1}, \ldots, S_{iF}\}$, whose members are the content objects sharing the $i$-th RRU. The payoff of
coalition $\mathcal{T}_i$ can be defined as the sum effective capacity of the considered cluster $C_T$, Then the corresponding utility function can be expressed as

$$\psi(T_i) = \left[ \sum_{S_{in} \in T_i} \bar{E}(R_{in}) - c_{RU} \left( \sum_{S_{in} \in T_i} O(R_{in}) P_R + O(U_T) P_{CC} + O(U_C) P_{BH} \right) \right]^+, \quad (39)$$

where $\bar{E}(R_{in})$ is the expected effective capacity of content object $S_{in}$ given in (35), $\varrho_i$ is the cost part of forming coalition $\psi(T_i)$, $c_{RU}$ is an energy efficiency coefficient to control the impact of cost in (39), and $(a)^+ = \max(a, 0)$.

2) **A Distributed Merge and Split Algorithm**: Unlike the serving RRH allocation problem, the payoff of each content in RRU allocation is not related only to the members in its own coalition, since the spectral efficiency is determined by the partition result via the coalition formation process. Therefore, the hedonic coalition formation algorithm in the previous subsection is not applicable. Moreover, due to the existence of costs, our studied game is non-superadditive with an empty core, and thus the grand coalition cannot be formed [22]. To solve the RRU allocation problem efficiently, a distributed merge and split method is studied in this subsection.

The key idea of the proposed algorithm is to form the coalitions of content objects only through merge and split operations. Without loss of generality, we take the first $l$ coalitions $\mathcal{T}_1, \ldots, \mathcal{T}_l$ as an example, and the rules for merge and split operations are shown as follows:

- **Merge rule**: If the utility functions of coalitions $\mathcal{T}_1, \ldots, \mathcal{T}_l$ satisfy $\sum_{j=1}^l \psi(T_j) < \psi(\bigcup_{j=1}^l T_j)$, then $\mathcal{T}_1, \ldots, \mathcal{T}_l$ merge as one cluster $\bigcup_{j=1}^l T_j$.

- **Split rule**: If there exists a partition of a coalition $\mathcal{T}_j = \{ P_1, \ldots, P_m \}$, whose utility functions satisfy $\psi(T_j) < \sum_{i=1}^m \psi(P_i)$, then $\mathcal{T}_j$ splits into $m$ disjoint coalitions $P_1, \ldots, P_m$.

C. **A Nested Coalition Formation Game-Based Algorithm**

Recalling (39), the average sum effective capacity of content transmission in $C_T$ is jointly decided by the RRH and the RRU allocations, and thus the two formulated coalitional games are coupled. To obtain a tractable method solving this problem, a nested coalition formation game-based algorithm is studied. In particular, the effective capacity of each content object $\bar{E}(R_{jm})$ given in (35), which is the payoff part in (39), can be interpreted as the accumulation
of effective capacity improvement achieve by RRHs joining the coalition $\overline{E}(R_{jm})$ gradually, i.e.,

$$\overline{E}(R_{jm}) = \overline{E}(R_{jm}) - \overline{E}(\emptyset) = \sum_{k \in R_{jm}} [\overline{E}(R_{jm} \cup R_k) - \overline{E}(R_{jm})].$$  \hspace{1cm} (40)

Such an interpretation coincides with the process of RRH allocation. Therefore, when the coefficients of cost control in (36) and (39) satisfy $c_{RH} = c_{RU} = c_0$, the following relationship between the utility functions of two studied coalitional games can be established:

$$\psi(T_i) = \left[ \sum_{S_{in} \in T_i} v(R_{in}) \right]^+ = \left[ \sum_{S_{in} \in T_i} \sum_{R_k \in R_{in}} \phi_k(R_{in}) \right]^+. \hspace{1cm} (41)$$

Equation (41) shows that the utility of RRU allocation can be obtained by using the utility of RRH allocation. Moreover, due to the preference relation criteria given in (38), it ensures that the partition results of serving RRH allocation provide reliable utility for the coalitional game of RRU allocation. Then a nested coalitional formation can be proposed for the joint design of RRU and RRH allocations. As shown in Algorithm 2, all the required content objects are randomly assigned to orthogonal RRUUs, and form $N$ disjoint coalitions in the initialization phase. During the iteration phase, the content coalitions take arbitrary merge and split operations. The proposed hedonic coalitional game-based RRH allocation algorithm given in the previous part is nested to provide the total utility values of the studied C-RAN cluster $C_T$ in each iteration. After obtaining a final partition result, each RRH in $C_T$ should be check if it is the nearest serving RRH for any users, and those unnecessary ones are turned into sleep mode to improve energy efficiency.

Unlike the conventional coalition formation games, the utility value of the studied nested coalition formation game is jointly determined by the partition results of RRU and RRH allocations, which may impact the convergent performance. For example, $T_i$ and $T_j$ merge as one coalition $T_{i\cup j}$, since their utility functions follow the merge rule during a specific iteration. However, the Nash stable partition resulting from coalitional game-based RRH allocation algorithm may not be unique, and different initializations of RRH partitions or negotiation orders will lead to different utility values of content coalitions, which may change the relationship between $\psi(T_{i\cup j})$ and $\psi(T_i) + \psi(T_j)$. Therefore, there exists the possibility that $T_{i\cup j}$ may split into $T_i$ and $T_j$ once more, and the proposed algorithm may not converge. To avoid the instability of utility values, the initialization and the negotiation order should be given in the RRH allocation algorithm. In particular, during both the initialization and the iteration phases, the negotiation order of the
considered set $\mathcal{B} = \{b_{k_1}, \ldots, b_{k_L}\}$ is given by following the increasing order of the superscripts of its members, i.e.,

$$\Lambda : b_{k_1} \to \cdots \to b_{k_i} \to \cdots \to b_{k_L}, \quad b_{k_1}, \ldots, b_{k_L} \in \mathcal{B}, \text{ and } k_1 < \cdots < k_i < \cdots < k_L.$$  \hspace{1cm} (42)

Then the initializations of RRH partitions in Algorithm 2 can be written as follows, in which RRHs are encouraged to join a coalition that can maximize its utility during the initialization phase of RRH allocation:

$$R_k \in \mathcal{R}_{jn}, \text{ s.t. } \phi_k(\mathcal{R}_{jn}) = \max_{S_{jn} \in \Omega_j} C(\phi_k(\mathcal{R}_{jn})), \quad R_k \in \mathcal{R} \times \Lambda,$$

where $\mathcal{R} \times \Lambda$ denotes an ordered set that generate based on $\mathcal{R}$ with a given order $\Lambda$ defined in (42). To verify the convergence performance of the proposed joint allocation algorithm, which is based on a nested coalition formation game, the following theorem is provided.

**Theorem 6:** The proposed joint allocation algorithm, which is obtained by solving a nested coalition formation game, converges to a $\mathcal{D}_{hp}$-stable partition, which means that all the possible partitions cannot recur with additional merge or split operations.

**Proof:** Assuming that the final partition result is not $\mathcal{D}_{hp}$-stable, there exists one partition $\Pi^* = \{T^*_1, \ldots, T^*_N\}$ that can recur during the iteration process, and the corresponding transformation can be expressed as

$$\Pi^* \rightarrow \Pi_1 \rightarrow \cdots \rightarrow \Pi_i \cdots \rightarrow \Pi^*.$$  \hspace{1cm} (44)

The total utility value of content transmissions keeps increasing via the merge and split operations due to the rules introduced previously. Therefore, based on the transitivity of coalition utility caused by the partition transform given in (44), there exists a subset $\{T^*_j, \ldots, T^*_k\}$ of $\Pi^*$, and the utility values of its members satisfy the following relationship:

$$\sum_{p=1}^{l} \psi_1(T^*_{jp}) < \sum_{p=1}^{l} \psi_2(T^*_{jp}),$$

where $\psi_1(T^*_{jp})$ denotes the utility of $T^*_{jp}$ when $\Pi^*$ first shows up in (44), and $\psi_2(T^*_{jp})$ is defined similarly for the second appearance of $\Pi^*$. Due to (45), there exists at least one content coalition $T^*_j$, whose utility follows the relationship $\psi_1(T^*_j) < \psi_2(T^*_j)$, $T^*_j \in \{T^*_1, \ldots, T^*_N\}$. The spectral efficiency stays the same during the two appearances of $\Pi^*$, since the RRU allocation results are identical. Then the difference between $\psi_1(T^*_j)$ and $\psi_2(T^*_j)$ must be caused by different serving
Algorithm 2 (A nested coalition formation game-based algorithm)

Step 1. Joint allocation of RRU and RRH

Initialization: Formulate $N$ disjoint coalitions of content objects $\mathcal{T}_1, \ldots, \mathcal{T}_N$, $1 \leq N \leq L$;

Repeat: For each content coalition $\mathcal{T}_i$ ($\mathcal{T}_i \neq \emptyset$)

1) Merge operation: Negotiate with other coalitions, i.e., $\mathcal{T}_{j_1}, \ldots, \mathcal{T}_{j_l}$, $j_1, \ldots, j_l \neq i$;
   - Obtain the utility values of $\psi(\mathcal{T}_i)$, $\psi(\mathcal{T}_{j_1}), \ldots, \psi(\mathcal{T}_{j_l})$ and $\psi(\mathcal{T}_i \cup \mathcal{T}_{j_1} \cup \cdots \cup \mathcal{T}_{j_l})$ based on (41) and Algorithm 1, in which the initial RRH partition is given in (43), and both the RRH coalitions and their members follow the negotiation order defined in (42);
   - If $\psi(\mathcal{T}_i) + \sum_{p=1}^{l} \psi(\mathcal{T}_{j_p}) < \psi(\mathcal{T}_i \cup \mathcal{T}_{j_1} \cup \cdots \cup \mathcal{T}_{j_l})$, $\mathcal{T}_i = \{\mathcal{T}_i \cup \mathcal{T}_{j_1} \cup \cdots \cup \mathcal{T}_{j_l}\}$, $\mathcal{T}_{j_1} = \cdots = \mathcal{T}_{j_l} = \emptyset$;

2) Split operation: For each subset $\mathcal{T}_{i_{\text{sub}}}$ in $\mathcal{T}_i$
   - Obtain the utility values of $\psi(\mathcal{T}_i)$, $\psi(\mathcal{T}_{i_{\text{sub}}})$ and $\psi(\mathcal{T}_{i}/\mathcal{T}_{i_{\text{sub}}})$ based on (41) and Algorithm 1, in which the initial RRH partition is given in (43), and both the RRH coalitions and their members follow the negotiation order defined in (42);
   - If $\psi(\mathcal{T}_{i_{\text{sub}}}) + \psi(\mathcal{T}_{i}/\mathcal{T}_{i_{\text{sub}}}) > \psi(\mathcal{T}_i)$, $\mathcal{T}_i = \mathcal{T}_{i}/\mathcal{T}_{i_{\text{sub}}}$, and formulate a new coalition $\mathcal{T}_k = \mathcal{T}_{i_{\text{sub}}}$;

Termination: When the members of each coalition do not change.

Step 2. Identify RRHs that are not required by any user, and put them into sleep mode.

RRH allocations, which can be denoted as $\Xi_1 = \{\mathcal{R}_1^{1}, \ldots, \mathcal{R}_k^{1}\}$ and $\Xi_2 = \{\mathcal{R}_1^{2}, \ldots, \mathcal{R}_k^{2}\}$, respectively. As described in Algorithm 2, both the initialization and the negotiation order have been given, and thus $\Xi_2$ can be obtained via limited steps of partition transforms from $\Xi_1$, i.e.,

$$\Xi_1 \rightarrow \Xi_{s_1} \rightarrow \cdots \rightarrow \Xi_2.$$  \hspace{1cm} (46)

As shown in (46), there must exist an RRH that wants to leave its current coalition and join a new one due to the preference relation defined in (38), and thus $\Xi_1$ is not Nash-stable, which is the final RRH partition result of the first appearance of $\Pi^*$. Such a result contradicts the conclusion given in Proposition 4. Therefore, Algorithm 2 converges to a $\mathbb{D}_{hp}$-stable partition result, and the proof is complete.

Compared with centralized schemes, in which a heavy burden is imposed on the fronthaul
links to obtain global information, our proposed algorithm can allocate the serving RRHs and the RRUs distributively, and only local information is required. Therefore, our studied algorithm can achieve considerable performance gains with low computational complexity. Moreover, to further improve the performance of cluster content caching, the global joint processing, such as dynamic cluster formation and inter-cluster interference coordination scheduling, should be considered, as well as the content correlation for the placement and the management of both the cluster and the cloud caches. These global joint processing can be supported by our studied system model due to the existence of cluster controller and the cloud cache at the cloud center.

V. A SUBOPTIMAL ALGORITHM WITH LOWER COMPUTATIONAL COMPLEXITY

Although Algorithm 2 can converge to a stable solution with a limited number of iterations, the nested structure implies that the total utility of content transmissions is obtained by solving a coalitional game-based RRH allocation problem in each iteration. Therefore, to reduce the process complexity, an efficient method is to decouple the problem as two independent coalition formation games. In this section, a suboptimal RRU allocation algorithm is studied, whose utility does not depend on the partition results of the serving RRH allocation.

A. Utility Function Formulation Based on Shapley Value

To reduce the computational complexity, an alternative utility function for RRU allocation should be established. The interests of RRHs might be conflicting among the content objects, which should be considered for the RRU allocation. In particular, the content objects with less interest conflict on RRH allocation tend to share a common RRU. Therefore, to characterize the interest of a specific content object \( S_i \) on each RRH, we need to evaluate the importance of each RRH on \( S_i \), and the Shapley value is applied for the utility function formulation.

In our studied RRU allocation problem, the Shapley value can be interpreted as the expected marginal contribution of each RRH for the transmission of \( S_i \). In particular, a grand coalition \( \mathcal{N} \) formed by all RRHs in the typical C-RAN cluster \( C_T \) to transmit \( S_i \), and the Shapley value of \( R_j \) is defined as its expected marginal contribution when it joins the grand coalition in a random order, which can be expressed as follows:

\[
v_{i,j} = \sum_{\mathcal{N}_s \subseteq \mathcal{N} \setminus \{R_j\}} \frac{O(\mathcal{N}_s)! (O(\mathcal{N}) - O(\mathcal{N}_s) - 1)!}{O(\mathcal{N})} [\bar{E}(\mathcal{N}_s \cup R_j) - \bar{E}(\mathcal{N}_s)],
\]

(47)
where $\bar{E}(N_s \cup \text{RRH}_j) - \bar{E}(N_s)$ is the marginal contribution of R$_j$ in coalition $N_s$ based on (35), and $\frac{O(N_s)!O(N_s)-O(N_s)-1)!}{O(N)!}$ is the probability that such a subset $N_s$ occurs when RRHs join the grand coalition under a random order.

The importance of each RRH for a specific content object can be evaluated by its Shapley value, i.e., R$_j$ with a higher Shapley value $v_{i,j}$ is more important to $S_i$. Then the utility function of RRU allocation can be formulated based on the interest conflicts among the content objects that share the same RRU, which can be expressed as follows for a given content object $S_j$ to join a coalition $T_i = \{S_{i_1}, \ldots, S_{i_F}\}$, $S_j \notin T_i$:

$$\varphi_j(T_i) = \sum_{S_{i_n} \in T_i} \left( \sum_{k=1}^{K} |v_{j,k} - v_{i_n,k}| \right) - \varrho_i,$$

(48)

where the cost part $\varrho_i$ follows the definition given in (39). In particular, when $v_{j,k} = v_{i_n,k}$ for each RRH, it implies that $S_j$ and $S_{i_n}$ have the same interests on RRU allocation, and the payoff is set as zero since it is the most competitive case of RRU allocation. As the differences of the Shapley values increase, the interest conflicts of RRHs between $S_j$ and $S_{i_n}$ can be mitigated, and thus the payoff of $S_j$ increases. Note that cost of RRU sharing is the performance loss caused by co-channel interference, which becomes severer as the cardinality of the consider coalition $T_i$ increases. Therefore, the grand coalition formation, which implies that all the content objects share a common RRU, is not the best choice in most instances, and the cost part in (39) can control the scale of each coalition efficiently.

B. A Suboptimal RRU Allocation Algorithm

Based on (48), the payoff function of each content object is related to the members in its own coalition only, and thus can be solved by a hedonic method as well. To establish the preference relation criteria given in (38), the total utility of $T_i$ can be written as

$$u(T_i) = \sum_{S_j \in T_i} \varphi_j(T_i).$$

(49)

Then a hedonic RRU allocation algorithm can be given, which is similar to Algorithm 1, and the Nash stability of its partition results can be guaranteed by Proposition 5. Algorithm 3 provides a sub-optimal solution with lower computational complexity, in which the joint design of RRH and RRU allocations are decoupled as two independent coalition formation games.
**Algorithm 3 (A suboptimal coalitional game-based algorithm)**

**Step 1. RRU allocation**

- **Initialization**: All content objects in $C_T$ are divided into $N$ disjoint coalitions $T_1, \ldots, T_N$;
- **Repeat**: For each coalition $T_k$
  - $S_j$ in $T_k$ negotiates with other coalitions $T_m, m \neq k$, and $\varphi_j(T_k)$ and $\varphi_j(T_m)$ can be obtained due to (48), and the total utility values of each coalitions is based on (49);
  - If the obtained utility values satisfy the preference relation criteria given in (38), then $S_j$ transfers into a new coalition $T_m$;
- **Termination**: When the members of $T_1, \ldots, T_N$ do not change.

**Step 2. RRH allocation by following Algorithm 1.**

**Step 3. Identify RRHs that are not required by any user, and put them into sleep mode.**

In Step 1 of Algorithm 2 and Step 1-2 of Algorithm 3, where utility values of coalition formation games are based on the analytical results, we assumed that all the RRHs are active, and thus all of them should transmit at the same time. Moreover, each user accesses its nearest RRH that serves its desired content. Therefore, these assumptions are consistent to those in Section II, which do not change the distribution of interference. The turning-off process is just an additional step to avoiding power consumption on keeping unnecessary RRHs active, which is decided by the partition result only.

**C. Comparison of Computational Complexity of Algorithm 2 and Algorithm 3**

1) **Computational Complexity Analysis of Algorithm 2**: In Algorithm 2, both RRH and RRU are allocated by solving coalition formation games. The computational complexity of proposed algorithm is mainly decided by the communication between different coalitions and the computation of utility function. Denoting the number of RRHs in $C_T$ as $D$, the obtained partition results of the RRH allocation in Algorithm 2 are maximum coalitions in [23], and its computational complexity can be estimated as $O((D+M)^{2D+M})$. Then we analyze the computation complexity of RRU allocation, which is based on the pairwise negotiation process. As introduced in [24], the RRU allocation can be accomplished in $O(L^2)$ steps of negotiation. In each negotiation operation, a coalition formation game-based RRH allocation problem should be solved due to the nested
structure. Note that \( M \) increases linearly as \( L \) increases, i.e., \( M = aL, \) \( 0 < a < 1. \) Therefore, the computational complexity of Algorithm 2 can be estimated as \( O(L^2(D + aL)2^{2D+aL}). \)

2) **Computational Complexity Analysis of Algorithm 3:** Although the computation of Shapley value is NP-Hard, it can be approximated efficiently and accurately, such as the linear approximation method in [25], where the computational complexity can be estimated as \( O(D^3) \) in our studied problem. Similar to RRH allocation in Algorithm 2, \( N \) RRU's can be allocated among \( L \) content objects by solving a hedonic coalition formation game, and its computational complexity can be estimated as \( O((1 + b)L2^{(2+b)L}) \), where \( N = bL, \) \( 0 < b < 1. \) Since the computation of Shapley value and the allocations of RRH and RRU can be executed independently, the computational complexity of Algorithm 3 can be estimated as \( O(D^3+(1+b)L2^{(2+b)L}+(D+aL)2^{2D+aL}), \) which is lower than that of Algorithm 2 due to the leading term analysis.

The joint design of RRU and RRH in Section IV and V are related to the previous section, since the utility functions are formulated based on the derived effective capacity provided by Theorem 3. Moreover, by using effective capacity as a metric, the resource allocation results are jointly decided by the channel capacity and the quality of service (QoS). In particular, RRH allocation strategy is taken as an example. In the conventional allocation schemes, each RRH should serve the content requested by its nearest user on the criterion of throughput. However, it is different in our studied caching-based model. When its nearest user requests a content that are not stored locally, a RRH may choose to serve other contents in its local cluster cache instead, which can achieve higher effective capacity due to the improvement of QoS. There exist similar cases in RRB allocation. Therefore, the employment of caches has great impacts on the resource allocation in our studied scenario. Moreover, our proposed resource allocation algorithms are also applicable for the non-cacheable contents, which can improve the delay experience.

**VI. Simulation Results**

In this section, simulation results are provided to verify the accuracy of the performance analysis and to show the performance gains of the proposed scheme with optimized resource allocation. In particular, the channel are assumed to be block fading in each RRU, where \( T = 1 \) ms and \( W = 1 \) kHz. The locations of RRHs and users are generated by applying homogenous PPPs in a disc region with a given radius 1000 m, and the densities of RRHs and users are given as \( \lambda_R = \lambda_U = 5 \times 10^{-6}. \) The cardinality of the required content set is \( L = 5, \) the size of each
content object is set as $B_S = 1$ Mbits, and the popularity of content objects follows Zipf’s law [29]. The power consumptions of RRHs in the active and the sleep modes are set as $P_{R}^{\text{act}} = 104$ W and $P_{R}^{\text{sle}} = 56$ W, respectively [27], and the power consumption of cluster content caching and backhaul transmissions defined in are set as $P_{CC} = 0.15$ W and $P_{BH} = 10$ W.

To verify the accuracy of our theoretical analysis, the effective capacity of a typical user is plotted in Fig. 3, where the path loss exponent is set as $\beta = 4, 6, 8$. As shown in the figure, the numerical results based on the analytical results given in Theorem 3 match the Monte Carlo results perfectly, which shows the validity of our theoretical derivations. Moreover, as the value of the path loss exponent $\beta$ increases, the effective capacity of a typical user increases. The reason
is that the receive SINR can be improved as the signal to interference path loss ratio increases, i.e., $\gamma \sim \frac{d_m^{-\beta}}{(\sum_{j \neq m} d_j^{-\beta})} = 1/(\sum_{j \neq m} (d_m/d_j)^\beta)$. Note that $d_m \leq d_j$ the serving RRH is usually nearer to the user than the interfere RRHs, and thus $\gamma$ is an increasing function with respect to the path loss exponent $\beta$, and so is the effective capacity.

The performance of a typical cluster is evaluated in Fig. 4, where the Zipf exponent is set as $s = 0, 0.5, 1, 2$, respectively, and the QoS exponents are set as $\theta^T = 0.1$ and $\theta^C = 0.6$, respectively. As shown in the figure, both the effective capacity and the energy efficiency increase as the size of cluster cache is enlarged, since more requests can be responded locally with short delay. In particular, compared with no-caching scheme, i.e., $K = 0$, the average effective capacity and the average effective capacity to average power consumption ratio can be improved by 0.57 Mbit/s/Hz and 0.004 Mbit/Joule when $K = 5$. Moreover, the performance of cluster content caching improves faster as $s$ increases, since the user interests converge to fewer popular contents stored in cluster cache.

To further evaluate the performance improvement of cluster content caching, the performance of the proposed RRU and RRH allocation algorithms is provided in Fig. 5, where the effective capacity and the effective capacity to power consumption ratio are plotted, respectively. In particular, the average effective capacity to power consumption ratio is provided to show the energy efficiency of cluster content caching, which can be defined as follows for the $i$-th RRU:

$$\eta_i = \frac{\sum_{S_l} \bar{E}(R_{S_l})}{\sum_{S_l} \mathcal{O}(R_{S_l}) P_{R}^{\text{act}} + (N_T - \sum_{S_l} \mathcal{O}(R_{S_l})) P_{R}^{\text{le}} + \mathcal{O}(U_T) P_{CC} + \mathcal{O}(U_C) P_{BH}}, \quad S_l \in \mathcal{T}_i. \quad (50)$$

As shown in Fig. 5, both the effective capacity and the energy efficiency increase as $M$ increases and $\theta^C$ decreases. Compared with the simulation results in Fig. 4, the performance gains of cluster content caching can be enlarged to 0.95 Mbit/s/Hz and 0.0055 Mbit/Joule when $K = 5$.

In Fig. 6, the performance comparison of different RRU and RRH allocation schemes is provided. Two previous allocation schemes are selected as two comparable schemes, which are the orthogonal RRU allocation scheme and the RRU full-reusing scheme. As shown in Fig. 5(a), Algorithm 2 can always achieve the best average effective capacity performance. There exists a performance gap between Algorithm 2 and Algorithm 3, since the utility function of RRU allocation in Algorithm 3 cannot ensure each content object makes the best choice. As $\lambda_R$ increases, the performance of RRU full-reusing scheme approaches that of Algorithm 2, which implies that the denser RRH coverage can alleviate the interest conflict among content objects,
and thus all of them tend to transmit in a common RRU. Although increasing the density of RRHs can always improve the effective capacity performance, it is not the best choice from an energy efficiency perspective. As shown in Fig. 5(b), the effective capacity to power consumption ratio increases in the low \( \lambda_R \) region, while it keeps decreasing in the high \( \lambda_R \) region. Moreover, the case of multi-casting can be considered as a special case of our studied scheme, i.e., when each content object is served by using orthogonal RRUs, and the performance of the orthogonal RRU allocation scheme in Fig.6 can be treated as a performance lower bound of multi-casting.

VII. Conclusion
In this paper, a cluster content caching scheme has been proposed in C-RANs, in which some requested content could be stored in local cluster content caches. By using a stochastic geometry-based network model, the effective capacity, which is defined as a link-level QoS metric, has been extended to our studied C-RAN scenario with content caching. Tractable expressions for the effective capacity and the energy efficiency have been derived to verify the performance gains of our proposed scheme. Then the joint design of RRU and RRH allocations has been studied to further improve the performance of cluster content caching, and two coalition formation game-based algorithms have been designed. The simulation results show that the effective capacity and the energy efficiency can be improved by 0.57 Mbit/s/Hz and 0.004 Mbit/Joule at most when the number of required content objects is 5, due primely to the relieving loading on backhaul of the proposed scheme. By employing our proposed optimization algorithms, the performance gains can be enlarged to 0.95 Mbit/s/Hz and 0.0055 Mbit/Joule, respectively.

REFERENCES

[1] “C-RAN: The road towards green RAN,” White Paper 2.5 ed. Beijing, China, China Mobile Res. Inst., 2011.
[2] A. Sanderovich, S. Shamai, and Y. Steinberg, “Distributed MIMO receiver-Achievable rates and upper bounds,” IEEE Trans. Inform. Theory, vol. 55, no. 10, pp. 4419-4438, Oct. 2009.
[3] S. Park, O. Simeone, O. Sahin, and S. Shitz, “Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory,” IEEE Signal Process. Mag., vol. 31, no. 6, pp. 69-79, Nov. 2014.
[4] Y. Zhou and W. Yu, “Optimized backhaul compression for uplink cloud radio access network,” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1295-1307, June 2014.
[5] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, “Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies,” IEEE Wireless Commun., vol. 21, no. 6, pp. 126-135, Dec. 2014.
[6] M. Peng, S. Yan, K. Zhang, and C. Wang, “Fog computing based radio access networks: Issues and challenges,” accepted by IEEE Networks, [Online]. Available: http://arxiv.org/pdf/1506.04233.pdf.
[7] V. Suryaprakash, P. Rost, and G. Fettweis, “Are heterogeneous cloud-based radio access networks cost effective?” to appear in IEEE J. Sel. Areas Commun., [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7110548.
[8] Z. Zhao, M. Peng, Z. Ding, C. Wang, and H. V. Poor, “Cluster formation in cloud-radio access networks: Performance analysis and algorithms design,” in Proc. IEEE ICC 2015, London, U.K., pp. 3903-3908, June 2015.
[9] E. Bastug, M. Bennis and M. Debbah, “Living on the edge: The role of proactive caching in 5G wireless networks,” IEEE Commun. Mag., vol. 52, no. 8, pp. 82-89, Aug. 2014.
[10] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, “Green Delivery: Proactive content caching and push with energy harvesting based small cells,” IEEE Commun. Mag., vol. 53, no. 4, pp. 142-149, Apr. 2015.
[11] E. Bastug, M. Bennis, M. Kountouris, and M. Debbah, “Cache-enabled small cell networks: Modeling and tradeoffs,” EURASIP Journal on Wireless Commun. and Network., vol. 2015, no. 1, pp. 1-11, Feb. 2015.
[12] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, “Cache in the air: Exploiting content caching and delivery techniques for 5G systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131-139, Feb. 2014.

[13] Y. Xu, Y. Li, Z. Wang, T. Lin, G. Zhang, and S. Ci, “Coordinated caching model for minimizing energy consumption in radio access network,” in *Proc. IEEE ICC 2014*, Sydney, Australia, pp. 2406-2411, June 2014.

[14] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, “On content-centric wireless delivery networks,” *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 118-125, Dec. 2014.

[15] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630-643, July 2003.

[16] D. Qiao, M. C. Gursoy, and S. Velipasalar, “Effective capacity of two-hop wireless communication systems,” *IEEE Trans. Inform. Theory*, vol. 59, no. 2, pp. 873-885, Feb. 2013.

[17] C. Zhong, T. Ratnarajah, K.-K. Wong, and M.-S. Alouini, “Effective capacity of correlated MISO channels,” in *Proc. IEEE ICC 2011*, Kyoto, Japan, pp. 1-5, June 2011.

[18] D. Wu and R. Negi, “Effective capacity-based quality of service measures for wireless networks,” *J. ACM Mobile Networks and Applications*, vol. 11, no. 1, pp. 91-99, Feb. 2006.

[19] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, “Modeling and analysis of $K$-tier downlink heterogeneous cellular networks,” *IEEE J. Selected Areas in Commun.*, vol. 30, no. 3, pp. 550-560, Apr. 2012.

[20] Y. A. Sambo, M. Z. Shakir, K. A. Qaraqe, E. Serpedin, and M. A. Imran, “Expanding cellular coverage via cell-edge deployment in heterogeneous networks: Spectral efficiency and backhaul power consumption perspectives,” *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 140-149, June 2014.

[21] W. Saad, Z. Han, T. Basar, A. Hjorungnes, and J. B. Song, “Hedonic coalition formation games for secondary base station cooperation in cognitive radio networks,” in *Proc. IEEE WCNC 2010*, Sydney, Australia, pp. 1-6, Apr. 2010.

[22] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, “Coalitional game theory for communication networks,” *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 77-97, Sep. 2009.

[23] T. Shrot, Y. Aumann, and S. Kraus, “Easy and hard coalition resource game formation problems-A parameterized complexity analysis,” in *Proc. AAMAS 2009*, Budapest, Hungary, pp. 433-440, May 2009.

[24] O. Shehory and S. Kraus, “Coalition formation among autonomous agents: Strategies and complexity (preliminary report),” *Lecture Notes in Computer Science*, vol. 957, pp. 55-72, June 2005.

[25] S. S. Fatima, M. Wooldridge, and N. R. Jennings, “A linear approximation method for the Shapley value,” *Artificial Intelligence*, vol. 172, no. 14, pp. 1673-1699, Sep. 2008.

[26] G. Auer, V. Giannini, C. Desset, I Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, “How much energy is needed to run a wireless network?” *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40-49, Oct. 2011.

[27] U. Lee, I. Rimac, and V. Hilt, “Greening the internet with content-centric networking,” in *Proc. ACM the 1st International Conf. on Energy-Efficient Comput. and Network.*, Passau, Germany, pp. 179-182, Apr. 2010.

[28] A. Hospodor, “Hit ratio of caching disk buffers,” in *Proc. IEEE CompCon Spring ’92*, San Francisco, USA, pp. 427-432, Feb. 1992.

[29] L. Breslau, “Web caching and zipf-like distributions: Evidence and implications,” in *Proc. IEEE INFOCOM 1999*, New York, USA, pp. 126-134, Mar. 1999.

[30] R. G. Gallager, *Principles of Digital Communication*, Cambridge University Press, Cambridge, UK, 2008.