IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 25, 2021, accepted December 10, 2021, date of publication December 21, 2021,
date of current version January 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3137148

Caching and Multicasting for Fog Radio Access Networks

ALAA BANI-BAKR[1], MHD NOUR HINDIA[1], (Member, IEEE), KAHARUDIN DIMYATI[1], (Member, IEEE), ZATI BAYANI ZAWAWI[2], AND TENGKU FAIZ TENGKU MOHMED NOOR IZAM[1]

[1]Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Independent Researcher, Kuala Lumpur 40470, Malaysia

Corresponding author: Tengku Faiz Tengku Mohmed Noor Izam (tengkufaiz@um.edu.my)

This work was supported by the Fundamental Research Grant Scheme (FRGS) under Grant FRGS/1/2020/TK0/UM/02/39 (FP19-2020).

ABSTRACT Cache-enabled Fog radio access network (F-RAN) is a promising technology to alleviate the traffic congestion and boost the contents delivery success rate. The efficiency of disseminating the cached contents in F-RAN can be boosted by enabling multicast service at the fog access points (F-APs). This paper proposes a joint random caching and multicasting optimization scheme for wireless backhauled F-RAN. Using tools from stochastic geometry, the expression of the successful transmission probability (STP) is derived by carefully analyzing the different types of serving F-APs and interferers. Then, a closed-form expression of the asymptotic STP in the high SNR region is derived to reduce the complexity. The joint caching and multicasting optimization problem is formulated to maximize the STP. The optimization problem is complex and non-convex in general. A novel projected cuckoo search algorithm (PCSA) is proposed to obtain the optimal content placement that maximizes the STP. The numerical simulation results show that PCSA outperforms the original cuckoo search algorithm (CSA) and the proposed asymptotically joint caching and multicasting scheme outperforms the benchmark caching schemes by up to 15% higher STPs.

INDEX TERMS Backhaul, caching, cuckoo search algorithm, F-RAN, optimization, multicast, stochastic geometry.

I. INTRODUCTION

Cloud radio access network (C-RAN) has been regarded as a promising network architecture to support caching technology in the fifth generation communication systems (5G). However, the rapid growth in the number of smart mobile devices and services along with the centralized architecture of C-RAN have led to many limitations, including traffic congestion, high end-to-end latency, and processing of massive amount of data [1]. Evolving from C-RAN, fog radio access network (F-RAN) architecture has been recently proposed by Cisco to alleviate the aforementioned limitations [1]–[3]. F-RAN is a distributed computing network architecture in which the fog access points (F-APs) are provided with limited computing and cache capabilities [4]. In F-RAN, bringing the cache resources and thus the most popular contents closer to the end-users at the edge of the network can effectively reduce traffic congestion and communication latency. Moreover, enabling multicast transmission at the F-APs is considered an efficient technique to disseminate contents to multiple end-users that are simultaneously requesting the same content [5]. Due to the cache resources constraint, the joint optimization of the content placement and multicasting service at F-APs is of great importance to boost the performance of F-RAN.

The caching problem in the cloud has been addressed by [6], wherein the authors proposed optimal algorithms to minimize the total transfer and caching costs in the online and off-line cases. In [7], a federated learning framework was proposed to predict the user demands of a specific edge content in order to provide the exact location for its placement. The caching optimization in F-RAN has been addressed by several papers. In [8], transmission-aware caching strategies were proposed for the distributed and centralized modes of F-RAN aiming at minimizing the mean download delay of the end-users. In [9], the authors proposed a probabilistic caching

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaofan He.

strategy that maximizes the hit probability in self-backhauled millimeter-wave F-RAN. The optimal cache design presented in [10] maximises the fractional offloaded traffic (FOT) and successful transmission probability (STP) for joint transmission and parallel transmission strategies. In [11], an edge caching scheme was proposed to improve the cache hit rate by predicting contents popularity and learning the end-users preferences. In [12], the authors proposed a proactive cache placement strategy to optimize the STP in wireless backhauled F-RAN, where the optimal caching design was obtained using projection gradient method. In [13], a dynamic distributed edge caching strategy was proposed to reduce the fronthaul traffic load and request service delay in ultra-dense F-RAN. In [14], the authors used actor–critic reinforcement learning to optimize the service delay in a fog-enabled IoT network by jointly considering the computing, caching, and radio resources. In [15], the association of end-user, prediction of content popularity, and placement of content were optimized using machine learning based algorithms. In [16], in order to mitigate the inter-F-AP interference, the F-APs were self-organized into multiple clusters and the joint radio and cache resource management optimization problem was formulated as Stackelberg game. A graph based cooperative caching scheme to maximize the offload traffic was presented in [17]. In [18], a hybrid caching scheme was proposed to balance the delay and energy efficiency. In [19], the authors proposed a delay-aware cache update scheme that takes into account the mobility of the end-users in F-RAN, wherein dueling deep-Q-network framework was utilized to minimize the average transmission delay. The alternating direction method of multipliers algorithm was used to optimize the cache placement of the capacity-aware edge caching strategy proposed in [20]. In [21], the authors formulated the edge caching problem to optimize the FOT, STP, and delay, where an improved fruit fly optimization algorithm was used to find the optimal content cache placement. In [22], the authors proposed a joint power allocation and proactive cooperative caching scheme that minimizes the latency in F-RAN. In [23], [24], the authors proposed multi-objective caching schemes for wireless backhauled F-RAN. Nevertheless, none of the aforementioned contributions considers the multicast transmission scheme.

The joint caching and multicasting in F-RAN was investigated in [25]–[27]. In [25], the authors tackled the optimization of non-orthogonal multiple access (NOMA) multicast transmission in F-RAN, wherein the problem was formulated to reduce the delivery latency of the downlinks from the content provider to end-users via the fog nodes. However, the problem of optimizing the cache placement was not addressed. In [26], the authors proposed a joint caching and node association strategy to maximize the energy efficiency in a multicast-enabled heterogeneous F-RAN. However, the wireless backhauling of the fog nodes was not taken into consideration. In the joint caching and multicasting scheme in [27], the caching optimization problem was formulated as a mixed-integer non-linear programming (MINLP) problem aiming at minimizing the transmission time of cloud processor. Then, the problem was relaxed and transformed into a tractable one. However, the fronthaul links from the F-APs to the end-users and end-users preferences were not taken into account. The optimization of joint caching and multicasting in traditional cellular network was addressed in [28]–[31], wherein the problem of caching and multicasting design was formulated to improve the STP. It is noted that [28]–[31] did not take into account the wireless backhauling of the base stations.

Motivated by the previous discussions, this paper proposes a joint random caching and multicasting scheme to maximize the STP in F-RAN. The proposed scheme is based on the content combinations and takes into consideration the stochastic nature of the channel, multicast transmission, and the wireless backhauling of the F-APs. The main contributions of this paper are summarized as follows:

- First, expressions of the association probabilities with the direct and transit F-APs with respect to the requested content are derived. Then, an expression of the STP in the general SNR region is derived using stochastic geometry tool and by carefully analyzing the different types of serving F-APs and interferers.
- A closed-form expression of the asymptotically STP in the high SNR is derived to reduce the complexity.
- The optimization problem is formulated to obtain the optimal cache placement that maximizes the STP.
- A novel projected cuckoo search algorithm (PCSA) is developed to obtain the optimal caching placement.
- Finally, the numerical simulations show that the proposed PCSA outperforms the original cuckoo search algorithm (CSA) and converges faster than it. The results also show that the proposed asymptotically caching scheme outperforms the traditional caching schemes.

The rest of this paper is organized as follows. In Section II, we present the system model, including the network, caching, multicasting, and association models. The performance analysis and the problem optimization are presented in Section III. The numerical results and discussions are provided in Section IV. The concluding remarks are delivered in Section V. A list of the key notations used throughout the paper is provided in Table 1.

II. SYSTEM MODEL

A. NETWORK MODEL

Consider a downlink cache-enabled F-RAN consisting of a tier of F-APs connected through wireless backhaul links with a tier of cloud access points (C-APs). The locations of the limited cache capacity F-APs in $\mathbb{R}^2$ are modeled as an independent homogeneous Poisson point process (PPP) Φ_F with density λ_F. Whereas, the C-APs are modeled as another independent homogeneous PPP Φ_C with density λ_C, such that $\lambda_F > \lambda_C$. The locations of the end-users are modeled as an independent homogeneous PPP Φ_U with density λ_U. With no loss of generality, we focus on the typical end-user

TABLE 1. Key notations.

Notation	Description
Φ_F	point process of F-APs
Φ_C	point process of C-APs
Φ_U	point process of end-users
Φ_m	point process of F-APs cache content m
Φ_{-m}	point process of F-APs that do not cache content m
$\Phi_{a,m}$	point process of available F-APs when m is requested
$\Phi_{-a,m}$	point process of unavailable F-APs when m is requested
λ_F	density of Φ_F
λ_C	density of Φ_C
λ_U	density of Φ_U
a_m	probability of randomly requesting content m by an end-user
$\boldsymbol{p}$	caching distribution of combination
p_j	probability of caching combination j at a F-AP
$\boldsymbol{T}$	caching distribution of contents
T_m	probability of caching content m at a F-AP
b_μ	inactive probability of the content μ
Λ_m	probability of available F-APs when m is requested
$F_{m,0}$	direct F-AP with respect to content m
$F_{a,0}$	transit F-AP with respect to content m
C_0	nearest C-AP to $F_{a,0}$
K	cache size of a F-AP
$\mathcal{K}_0$	content load of the serving F-AP
M	total number of contents
$SINR_{m,0}$	SINR at u_0 when it is served by $F_{m,0}$
$SINR_{a,0}$	SINR at u_0 when it is served by $F_{a,0}$
$SINR_{C,a}$	SINR at $F_{a,0}$ when u_0 is served by $F_{a,0}$
$D_{0,0}$	distance between u_0 and $F_{m,0}$
$D_{\ell,0}$	distance between u_0 and access point ℓ
$D_{a,0}$	distance between u_0 and $F_{a,0}$
$D_{C,a}$	distance between $F_{a,0}$ ad C_0
$D_{\ell,a}$	distance between $F_{a,0}$ and access point ℓ
$h_{0,0}$	channel coefficient between u_0 and $F_{m,0}$
$h_{\ell,0}$	channel coefficient between u_0 and access point ℓ
$h_{a,0}$	channel coefficient between u_0 and $F_{a,0}$
$h_{C,a}$	channel coefficient between $F_{a,0}$ and C_0
$h_{\ell,a}$	channel coefficient between $F_{a,0}$ and access point ℓ
$q_{m,0}(\boldsymbol{p})$	STP of content m when u_0 is served by $F_{m,0}$
$q_{C,0}(\boldsymbol{p})$	STP of content m when u_0 is served by $F_{a,0}$
$q_{a,0}(\boldsymbol{p})$	STP of content m over the link from $F_{a,0}$ to u_0
$q_{C,a}(\boldsymbol{p})$	STP of content m over the link from C_0 to $F_{a,0}$
$q(\boldsymbol{p})$	STP of u_0
$q_\infty(\boldsymbol{p})$	asymptotic STP of u_0 as $P/N_0 \to \infty$

u_0 located on origin [32]. It is assumed that all the F-APs and C-APs are equipped with a single antenna with transmission power P. The total transmission bandwidth of each F-AP and C-AP is W_F and W_C, respectively. Each end-user has one receive antenna. For the propagation model, both large-scale and small-scale fading are considered. For large-scale fading, a signal propagates distance D is attenuated by $D^{-\alpha}$, where α is the path loss exponent. Rayleigh fading is assumed to model the small-scale fading, wherein each small-scale fading coefficient h follows $|h|^2 \overset{d}{\sim} \exp(1)$.

B. CACHING AND MULTICASTING

Let $\mathcal{M} = \{1, 2, ..M\}$ stand for the set of M contents in the network. All the contents are assumed to be of the same size and have an identical apriori known popularity distribution among all the end-users for sake of tractability. Let $\boldsymbol{a} = (a_m)_{m\in\mathcal{M}}$ represent the content popularity distribution, where $a_m \in (0, 1)$, such that $\sum_{m=1}^{M} a_m = 1$, denotes the probability of randomly requesting content m by an end-user. It is assumed that a_m is characterized by Zipf distribution, i.e., $a_m = m^{-\gamma} / \sum_{m\in\mathcal{M}} m^{-\gamma}$, where γ is the skew parameter. With no loss of generality, the contents are assumed to be indexed according to $\boldsymbol{a}$ ranked from the most popular content to the least one (i.e., $a_1 \geq a_2 \geq \cdots \geq a_M$).

It is assumed that all the contents are cached at each C-AP. Whereas, in the tier of limited cache capacity F-APs, each F-AP is provided with a cache of size $K \leq M$. Accordingly, each F-AP caches a combination of K different contents out of M contents. Therefore, the total number of combinations in the F-AP tier is $J = \binom{M}{K}$. Let $\mathcal{J} = \{1, 2, \cdots, J\}$ stand for the set of J combinations. Combination $j \in \mathcal{J}$ can be represented as a vector $\boldsymbol{\chi}_j = (\chi_{j,m})_{m\in\mathcal{M}}$ with dimension M, where $\chi_{j,m} = 1$ if combination j contains content m, otherwise $\chi_{j,m} = 0$. Hence, the set of K contents in combination j can be represented as $\mathcal{M}_j = \{m;\ \chi_{j,m} = 1\}$.

The considered random content placement strategy is based on the combination of contents as in [28]–[31]. Let $\boldsymbol{p} = (p_j)_{j\in\mathcal{J}}$ denote the caching distribution of combinations, where p_j is the probability of caching combination j at each F-AP, such that p_j satisfies

$$0 \leq p_j \leq 1, \quad j \in \mathcal{J}, \tag{1}$$

$$\sum_{j\in\mathcal{J}} p_j = 1 \tag{2}$$

Based on the caching distribution of combinations $\boldsymbol{p}$, the caching distribution of contents can be defined as $\boldsymbol{T} = (T_m)_{m\in\mathcal{M}}$, where T_m denotes the probability of caching content m at each F-AP and given as follows

$$T_m = \sum_{j\in\mathcal{J}_m} p_j, \quad m \in \mathcal{M} \tag{3}$$

satisfying

$$0 \leq T_m \leq 1, \quad m \in \mathcal{M}, \tag{4}$$

$$\sum_{m\in\mathcal{M}} T_m = K \tag{5}$$

where $\mathcal{J}_m = \{j \in \mathcal{J} : \chi_{j,m} = 1\}$ is the set of $J_m = \binom{M-1}{K-1}$ combinations containing content m.

For an efficient content dissemination in the fronthaul link from the F-APs to the end-users, this paper adopts multicast transmission service at the F-APs. Multicast is a content-centric transmission scheme, in which a F-AP with L_0 associated end-users requesting K_0 (i.e., $L_0 \geq K_0$) different contents cached by it, serves all the users requesting the same content using a single transmission, in which each requested content is transmitted once using frequency division multiple access (FDMA) over $1/K_0$ of the total bandwidth [28]–[31], which results in a more efficient utilization of the total bandwidth compared with the connection-based transmission scheme known as unicast, in which the F-AP transmits a single content to each one of the L_0 requesters over $1/L_0$ of the total bandwidth. Note that, by adopting multicast transmission service the transmission rate of the F-AP is improved by L_0/K_0 compared with unicast transmission service. Whereas, for sake of tractability, broadcast

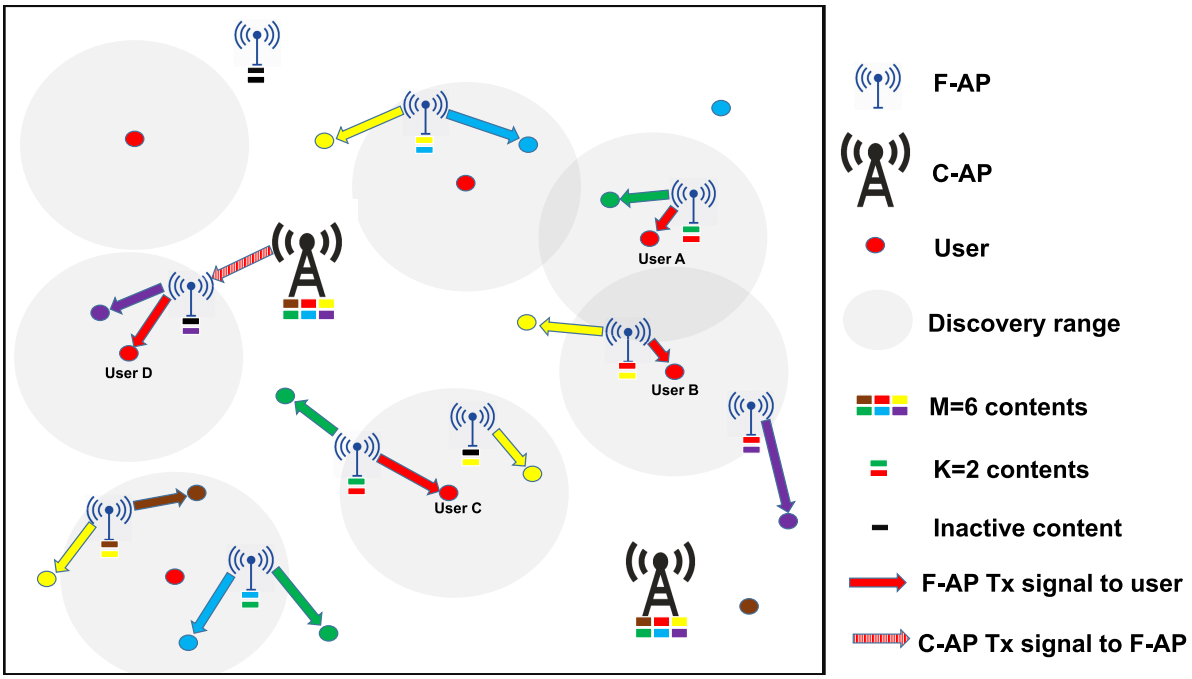

FIGURE 1. Illustration of the association mechanism.

transmission service is adopted by the C-APs, such that the transmission bandwidth of a C-AP is evenly shared by the content library, i.e., each content is transmitted over $1/M$ of the total bandwidth.

C. ASSOCIATION MODEL

This paper considers the content-centric association mechanism illustrated in Fig. 1, under which the serving F-AP is statistically determined by the caching distribution $\boldsymbol{p}$ and may not be the geographically nearest F-AP to the requester. By taking into account both the content-centric and physical layer proprieties, the considered association mechanism is more flexible and has a higher probability of association than the traditional connection-based association mechanism that only considers the physical layer proprieties and under which the requester is always associated with the geographically nearest F-AP to its location. It is assumed that the C-APs can only be accessed by the F-APs, i.e., there is no direct communication between the C-APs and end-users. Let R denote the discovery range (i.e., maximum communication range) of the typical end-user u_0. When u_0 randomly requests content m, the considered association mechanism can be demonstrated as

1) If content m is cached by a F-AP within R, u_0 will be associated with nearest F-AP that caches content m to serve it directly, e.g., end-users A, B, and C in Fig. 1. Here, the serving F-AP is denoted as $F_{m,0}$ and called 'direct F-AP'. Note that all the F-APs cache content m can be direct F-APs. Thus, the point process of the direct F-APs is $\Phi_m \subseteq \Phi_F$, which is a thinned PPP of density $T_m\lambda_F$. Then, the probability of serving u_0 by a direct F-AP within R when content m is requested can be obtained as follows

$$\mathcal{A}_{m,0} = 1 - \exp\left(-\pi T_m \lambda_F R^2\right) \tag{6}$$

Please refer to Appendix A for the proof of (6).

2) If there is no F-AP caches content m within R, the typical end-user u_0 will search within R for the nearest available F-AP $F_{a,0}$ to fetch content m from the nearest C-AP C_0, e.g., end-user D in Fig. 1. Due the 2-hop transmission $F_{a,0}$ is called 'transit F-AP'.
Note that the available F-AP is a F-AP caches at least one inactive content (i.e., a content not requested by any end-user associated with it). Let the random variable $Y_\mu \in \{0, 1\}$ denote whether content μ cached at a F-AP is being requested by end-users associated with it, where $Y_\mu = 1$ represents the event of requesting content μ by any associated end-user, and $Y_\mu = 0$ otherwise. The probability mass function (PMF) of Y_μ is influenced by the probability density function (PDF) of the Voronoi cell size of the F-AP caching content μ. Then, using proposition 1 of [33], the probability of content μ being inactive b_μ can be obtained as follows

$$b_\mu = \mathbb{P}\left[Y_\mu = 0\right] = \left(1 + \frac{a_\mu \lambda_U}{3.5\, T_\mu \lambda_F}\right)^{-3.5} \tag{7}$$

where notation $\mathbb{P}[.]$ stands for the probability.

Accordingly, the probability of the transit F-APs with respect to content m can be obtained as follows

$$\Lambda_m = \frac{1 - T_m}{K - T_m} \sum_{\mu \in \mathcal{M} \setminus m} T_\mu b_\mu \tag{8}$$

Please refer to Appendix B for the proof of (8).

Let $\Phi_{a,m} \subseteq \Phi_F$ denote the point process of the available F-APs with respect to content m. Then, by noting that $\Phi_{a,m}$ is a thinned PPP with density $\Lambda_m \lambda_F$, the probability of serving of u_0 when it requests content m by a transit F-AP within R can be obtained as follows

$$\mathcal{A}_{a,0} = \exp\left(-\pi T_m \lambda_F R^2\right)\left(1 - \exp\left(-\pi \Lambda_m \lambda_F R^2\right)\right) \tag{9}$$

Please refer to Appendix C for the proof of (9).

III. PERFORMANCE ANALYSIS AND PROBLEM OPTIMIZATION

A. PERFORMANCE ANALYSIS

The STP, defined by the probability that the end-users successfully receive their desired contents, is the performance metric considered by this paper. Therefore, when the typical user u_0 randomly requests content m, it can be successfully received and decoded if the channel capacity of u_0, which depends on the signal-to-interference-plus-noise ratio (SINR) and transmission bandwidth, exceeds the threshold transmission rate τ. However, under the adopted caching and multicasting strategy both the SINR of u_0 and the dedicated bandwidth to transmit content m are influenced by the combinations caching distribution $\boldsymbol{p}$. Thus, it is necessary to carefully analyze the impact of the caching distribution on both when u_0 is associated with the different types of serving F-APs to derive the expression of the STP.

When a direct F-AP $F_{m,0}$ within R serves u_0 when it requests content m, the SINR of u_0 can be expressed as in (10), as shown at the bottom of the next page, where $D_{0,0}$ denotes the distance between $F_{m,0}$ and u_0, $h_{0,0}$ denotes the small-scale channel coefficient between $F_{m,0}$ and u_0, N_0 is the noise power, $\ell \in \Phi_m \setminus F_{m,0}$ denote all the F-APs caching content m except the serving F-AP $F_{m,0}$, $\ell \in \Phi_{-m}$ denote all the F-APs that do not cache content m, $\ell \in \Phi_C$ denote all the C-APs, $D_{\ell,0}$ stands for the distance between access point ℓ and u_0, and $h_{\ell,0}$ denotes the channel coefficient between access point ℓ and u_0.

Let the discrete random variable $\mathcal{K}_0 \in \{1, \cdots, K\}$ represent the content load of the serving F-AP (i.e., the set of contents downloaded by the end-users associated with $F_{m,0}$). Bearing in mind that $F_{m,0}$ disseminates each one of the $\mathcal{K}_0$ contents over a bandwidth of $\frac{W_F}{\mathcal{K}_0}$ due to the adopted multicasting scheme. Thus, the STP of content m when u_0 is being served by $F_{m,0}$ is given as follows

$$q_{m,0}(\boldsymbol{p}) = \mathbb{P}\Bigg[\underbrace{\frac{W_F}{\mathcal{K}_0} \log_2\left(1 + SINR_{m,0}\right)}_{\triangleq \mathbb{C}_{m,0}} \geq \tau\Bigg] \tag{11}$$

where $\mathbb{C}_{m,0}$ represents the channel capacity of the link between $F_{m,0}$ and u_0. Generally, the content load $\mathcal{K}_0$ is correlated in a complex manner with the SINR since the larger association region of the F-APs results in higher content load due to the higher number of associated end-user and lower SINR owing to the larger F-AP to end-user distance [34]. As in [28]–[31], the correlation and dependence are ignored for analytical tractability. Hence, (11) can be rewritten as in (12), as shown at the bottom of the next page, where $\mathbb{P}[\mathcal{K}_0 = \kappa]$ represents the PMF of $\mathcal{K}_0$, which is given in (13), as shown at the bottom of the next page, and proved in Appendix D, and $q_{\kappa,m,0}(\boldsymbol{p})$ is the STP of content m when it is served by $F_{m,0}$ with a content load of κ.

Next, stochastic geometry tools are used to obtain $q_{\kappa,m,0}(\boldsymbol{p})$ as in the following theorem.

Theorem 1: The STP of content m when u_0 is served by the direct F-AP $F_{m,0}$ with a content load of κ can be expressed as in (14), as shown at the bottom of the next page, where

$$\begin{aligned}\mathcal{U}_\kappa(\boldsymbol{p}) = 1 &+ \frac{2}{\alpha}\left(2^{\frac{\kappa\tau}{W_F}} - 1\right)^{\frac{2}{\alpha}} \beta'\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}, 2^{\frac{-\kappa\tau}{W_F}}\right) \\ &+ \frac{2}{\alpha}\left(\frac{1 - T_m}{T_m}\right)\left(2^{\frac{\kappa\tau}{W_F}} - 1\right)^{\frac{2}{\alpha}} \beta\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right) \\ &+ \frac{2}{\alpha}\left(\frac{\lambda_C}{T_m \lambda_F}\right)\left(2^{\frac{\kappa\tau}{W_F}} - 1\right)^{\frac{2}{\alpha}} \beta\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)\end{aligned} \tag{15}$$

where $\beta'(x, y, z) \triangleq \int_z^1 u^{x-1}(1-u)^{y-1} du$ denotes the complementary incomplete Beta function, and $\beta(x, y) \triangleq \int_0^1 u^{x-1}(1-u)^{y-1} du$ denotes the Beta function.

Proof: See Appendix E

Noting that a 2-hop transmission is required for successfully delivering content m to u_0 when it is served by a transit F-AP $F_{a,0}$ within R. Thus, the STP of content m can be expressed as follows

$$\begin{aligned}q_{C,0}(\boldsymbol{p}) = &\mathbb{P}\Bigg[\underbrace{\frac{W_F}{\mathcal{K}_0} \log_2\left(1 + SINR_{a,0}\right)}_{\triangleq \mathbb{C}_{a,0}} \geq \tau\Bigg] \\ &\times \mathbb{P}\Bigg[\underbrace{\frac{W_C}{M} \log_2\left(1 + SINR_{C,a}\right)}_{\triangleq \mathbb{C}_{C,a}} \geq \tau\Bigg]\end{aligned} \tag{16}$$

where $\mathbb{C}_{a,0}$ is the channel capacity of the links between $F_{a,0}$ and u_0, $\mathbb{C}_{C,a}$ is the channel capacity of the links between C_0 and $F_{a.0}$, $SINR_{a,0}$ represents the SINR of u_0, and $SINR_{C,a}$ is the SINR of $F_{a,0}$.

By carefully analyzing the different types of interfering access points, $SINR_{a,0}$ and $SINR_{C,a}$ can be obtained as in (17) and (18), as shown at the bottom of the next page, respectively, where $D_{a,0}$ and $h_{0,0}$ are the distance and the channel coefficient between $F_{a,0}$ and u_0, respectively. $\ell \in \Phi_{a,m} \setminus F_{a,0}$ represent all the available F-APs with respect to content m except $F_{a,0}$, $\ell \in \Phi_{-a,m}$ denote all the unavailable F-APs, i.e., $\Phi_{-a,m} \triangleq \Phi_F \setminus \Phi_{a,m}$, $D_{C_0,a}$ is the distance between C_0 and $F_{a,0}$, $h_{C_0,a}$ is the channel coefficient between C_0 and $F_{a,0}$, $\ell \in \Phi_C \setminus C_0$ represent all the C-APs except C_0, $\ell \in \Phi_F \setminus F_{a,0}$

denote all the F-APs except $F_{a,0}$, $D_{\ell,a}$ is the distance between access point ℓ and $F_{a,0}$, and $h_{\ell,a}$ is the channel coefficient between access point ℓ and $F_{a,0}$.

Taking into consideration the random nature of the content load of the serving transit F-AP $F_{a,0}$, equation (16) can be rewritten as shown in (19), as shown at the bottom of the page, where $q_{\kappa,a,0}(\boldsymbol{p})$ is the STP of content m over the link from $F_{a,0}$ to u_0 when the content load of $F_{a,0}$ is κ, $q_{C,a}(\boldsymbol{p})$ is the STP of content m over the link from C_0 to $F_{a,0}$, and $q_{\kappa,C,0}(\boldsymbol{p})$ represents the STP of content m conditioned on the Content load of $F_{a,0}$ over the link from C_0 to u_0 via $F_{a,0}$. Utilizing tools from stochastic geometry, $q_{\kappa,C,0}(\boldsymbol{p})$ can be obtained as in Theorem 2.

Theorem 2: The STP of content m when u_0 is served by the transit F-AP $F_{a,0}$ with a content load of κ can be obtained as in (20), as shown at the bottom of the next page, where

$$\begin{aligned}\mathcal{V}_\kappa(\boldsymbol{p}) = 1 &+ \frac{2}{\alpha}\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}\beta'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{\frac{-\kappa\tau}{W_F}}\right)\\ &+\frac{2}{\alpha}\left(\frac{1-\Lambda_m}{\Lambda_m}\right)\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\\ &+\frac{2}{\alpha}\left(\frac{\lambda_C}{\Lambda_m\lambda_F}\right)\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\end{aligned} \tag{21}$$

and

$$\begin{aligned}\mathcal{G} = 1 &+ \frac{2}{\alpha}\left(2^{\frac{M\tau}{W_C}}-1\right)^{\frac{2}{\alpha}}\beta'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{\frac{-M\tau}{W_C}}\right)\\ &+\frac{2}{\alpha}\left(\frac{\lambda_F}{\lambda_C}\right)\left(2^{\frac{M\tau}{W_C}}-1\right)^{\frac{2}{\alpha}}\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\end{aligned} \tag{22}$$

Proof: See Appendix F.

Bearing in mind that there are M distinct contents in the system. Then, the STP of u_0 denoted as $q(\boldsymbol{p})$ can be expressed as in the following theorem.

Theorem 3: The STP of the typical end-user u_0 can be obtained as in (23), as shown at the bottom of the next page.

Proof: The proof is straightforward using total probability theorem and noting that the system contains M different contents and each one of them can be delivered to u_0 by either a transit or direct F-AP.

To reduce the complexity of the STP, the following corollary considers the asymptotic scenario of $\frac{P}{N_0}\to\infty$, which represents the scenario of interference-limited network.

Corollary 1 (Asymptotic Performance): When $\frac{P}{N_0}\to\infty$, the asymptotic STP $q_\infty(\boldsymbol{p})$ can be obtained as in (24), as shown at the bottom of the next page, where $\mathcal{U}_\kappa(\boldsymbol{p})$, $\mathcal{V}_\kappa(\boldsymbol{p})$, and $\mathcal{G}$ are given by (15), (21), and (22), respectively.

$$SINR_{m,0} = \frac{D_{0,0}^{-\alpha}\,|h_{0,0}|^2}{\sum_{\ell\in\Phi_m\setminus F_{m,0}} D_{\ell,0}^{-\alpha}\,|h_{\ell,0}|^2 + \sum_{\ell\in\Phi_{-m}} D_{\ell,0}^{-\alpha}\,|h_{\ell,0}|^2 + \sum_{\ell\in\Phi_C} D_{\ell,0}^{-\alpha}\,|h_{\ell,0}|^2 + \frac{N_0}{P}} \tag{10}$$

$$\begin{aligned}q_{m,0}(\boldsymbol{p}) &= \sum_{\kappa=1}^{K}\mathbb{P}\left[\mathcal{K}_0=\kappa\right]\mathbb{P}\left[\frac{W_F}{\kappa}\log_2\left(1+SINR_{m,0}\right)\geq\tau\right]\\ &= \sum_{\kappa=1}^{K}\mathbb{P}\left[\mathcal{K}_0=\kappa\right]\,q_{\kappa,m,0}(\boldsymbol{p})\end{aligned} \tag{12}$$

$$\mathbb{P}\left[\mathcal{K}_0=\kappa\right] = \sum_{j\in\mathcal{J}_m}\frac{p_j}{T_m}\sum_{\chi\in\{\mathcal{S}\subseteq\mathcal{M}_{j,-m}:|\mathcal{S}|=\kappa-1\}}\prod_{\mu\in\chi}\left(1-b_\mu\right)\prod_{\mu\in\mathcal{M}_{j,-m}/\chi}b_\mu \tag{13}$$

$$q_{\kappa,m,0}(\boldsymbol{p}) = 2\pi T_m\lambda_F\int_0^R d\exp\left(-\pi T_m\lambda_F\,\mathcal{U}_\kappa\left(\boldsymbol{p}\right)\,d^2-\frac{N_0}{P}\left(2^{\frac{\kappa\tau}{W_F}}-1\right)d^\alpha\right)\mathrm{d}d \tag{14}$$

$$SINR_{a,0} = \frac{D_{a,0}^{-\alpha}\,|h_{a,0}|^2}{\sum_{\ell\in\Phi_{a,m}\setminus F_{a,0}} D_{\ell,0}^{-\alpha}\,|h_{\ell,0}|^2 + \sum_{\ell\in\Phi_{-a,m}} D_{\ell,0}^{-\alpha}\,|h_{\ell,0}|^2 + \sum_{\ell\in\Phi_C} D_{\ell,0}^{-\alpha}\,|h_{\ell,0}|^2 + \frac{N_0}{P}} \tag{17}$$

$$SINR_{C,a} = \frac{D_{C_0,a}^{-\alpha}\,|h_{C_0,a}|^2}{\sum_{\ell\in\Phi_C\setminus C_0} D_{\ell,a}^{-\alpha}\,|h_{\ell,a}|^2 + \sum_{\ell\in\Phi_F\setminus F_{a,0}} D_{\ell,a}^{-\alpha}\,|h_{\ell,a}|^2 + \frac{N_0}{P}} \tag{18}$$

$$\begin{aligned}q_{C,0}(\boldsymbol{p}) &= \sum_{\kappa=1}^{K}\mathbb{P}\left[\mathcal{K}_0=\kappa\right]\underbrace{\mathbb{P}\left[\frac{W_F}{\kappa}\log_2\left(1+SINR_{m,0}\right)\geq\tau\right]}_{q_{\kappa,a,0}(\boldsymbol{p})}\underbrace{\mathbb{P}\left[\frac{W_C}{M}\log_2\left(1+SINR_{C,a}\right)\geq\tau\right]}_{q_{C,a}(\boldsymbol{p}}\\ &= \sum_{\kappa=1}^{K}\mathbb{P}\left[\mathcal{K}_0=\kappa\right]\,q_{\kappa,C,0}(\boldsymbol{p})\end{aligned} \tag{19}$$

Proof: As $\frac{P}{N_0} \to \infty$, *the terms containing it in the exponential functions in* $q_{\kappa,m,0}$, $q_{\kappa,a,0}$, *and* $q_{C,a}$ *approach zero. Thus,* $q_{\kappa,m,0}$ $q_{\kappa,a,0}$, *and* $q_{C,a}$ *can be written in the form of* $\int_0^a bd \exp\left(-cd^2\right)dd = \frac{b}{2c}\left(1 - \exp\left(-ca^2\right)\right)$, *where a, b, and c are constants. Therefore, we complete the proof.*

B. PROBLEM OPTIMIZATION

Since the STP is fundamentally influenced by the caching and multicasting model via the combinations caching distribution $\boldsymbol{p}$, this paper aims at maximizing the STP by optimizing the combinations caching distribution $\boldsymbol{p}$. Accordingly, the caching optimization problem is formulated as follows

Problem 1 (Caching Optimization)

$$\max_{\boldsymbol{p}} \; q\left(\boldsymbol{p}\right) \quad s.t.\ (1), (2) \tag{25}$$

Solving Problem 1 involves very high computational complexity as a result of the very complex form of $q(\boldsymbol{p})$. However, the complexity can be reduced by considering $q_\infty\left(\boldsymbol{p}\right)$ as an objective function as in Problem 2.

Problem 2 (Asymptotic Optimization when $\frac{P}{N_0} \to \infty$*)*

$$\max_{\boldsymbol{p}} \; q_\infty\left(\boldsymbol{p}\right) \quad s.t.\ (1), (2) \tag{26}$$

The convexity of $q\left(\boldsymbol{p}\right)$ and $q_\infty\left(\boldsymbol{p}\right)$ cannot be ensured due to their complex form. Thus, Problems 1 and 2 are non-convex in general.

This paper proposes PCSA to obtain optimal solution of the caching optimization problem. PCSA is a novel modified version of the original CSA. CSA is a meta-heuristic algorithm that shows high efficiency in solving the non-convex optimization problems [35]. Inspired form the nature, the original CSA was proposed by Yang and Deb in 2009 to mimic the obligate brood parasitism of cuckoo birds by generating new nests by Lévy flights and random walks in the first and second generation of each iteration, where the nest represents a solution [36]. CSA has attracted high attention lately because it outperforms the other nature-inspired optimization methods in terms of simplicity, quality solution, success rate, number of evaluations of the objective function, and execution time [37], [38]. However, the fast convergence of CSA cannot be assured due to the random walk-based search for the best nest. Moreover, the original CSA utilizes penalty method to treat constrained problems. However, it is not easy to guarantee the feasibility of the optimal solution using penalty method when there are equality constraints. Also, the selection of the user-defined penalty factors significantly affects the quality of the optimization. Moreover, when the optimization problem has multiple constraints, it is difficult to determine the penalty factor for each constraint as each one might has a different scale. To overcome these limitations, we develop PCSA outlined in Algorithm 1, where t is the iteration index, N_C is the maximum number of iterations, N_P is the population size, and $P_a \in (0, 1)$ is the fraction of nest to be abandoned.

In PCSA, the feasibility of the randomly generated nests is assured by the projection optimization performed at steps 3 and 7. The projection optimization is a search for the nearest nest to the projected nest $\breve{\boldsymbol{p}}^i$ in the set of feasible nests satisfying the problem's constraints. Therefore, the projection optimization problem is formulated as

Problem 3 (Projection Optimization)

$$\min_{\boldsymbol{p}^i} \; ||\boldsymbol{p}^i - \breve{\boldsymbol{p}}^i||^2 \;\; s.t.\ (1), (2) \tag{27}$$

where $||.||$ stands for the Euclidean norm. The projection optimization problem is convex and the optimal solution of which can be obtained using Matlab Optimization Toolbox, wherein the optimal solution is computed by as $p_j^i = \min\{\left[\breve{p}_j^i - \xi\right]^+, 1\}$, $\forall j \in \mathcal{J}$, here ξ satisfies $\sum_{j\in\mathcal{J}} \min\{\left[\breve{p}_j^i - \xi\right]^+, 1\} = 1$, where $[x]^+ \triangleq \max\{x, 0\}$.

$$q_{\kappa,C,0}(\boldsymbol{p}) = \underbrace{2\pi\,\Lambda_m\lambda_F \int_0^R d \exp\left(-\pi\,\Lambda_m\lambda_F\,\mathcal{V}_\kappa(\boldsymbol{p})\,d^2 - \frac{N_0}{P}\left(2^{\frac{\kappa\tau}{W_F}} - 1\right)d^\alpha\right)\mathrm{d}d}_{=q_{\kappa,a,0}(\boldsymbol{p})}$$
$$\times \underbrace{2\pi\lambda_C \int_0^\infty d \exp\left(-\pi\,\lambda_C\mathcal{G}\,d^2 - \frac{N_0}{P}\left(2^{\frac{M\tau}{W_C}} - 1\right)d^\alpha\right)\mathrm{d}d}_{=q_{C,a}(\boldsymbol{p})} \tag{20}$$

$$q\left(\boldsymbol{p}\right) = \sum_{m\in\mathcal{M}} a_m \sum_{\kappa=1}^{K} \mathbb{P}\left[\mathcal{K}_0 = \kappa\right]\left(\mathcal{A}_{m,0}\, q_{\kappa,m,0}\left(\boldsymbol{p}\right) + \mathcal{A}_{a,0}\, q_{\kappa,C,0}\left(\boldsymbol{p}\right)\right) \tag{23}$$

$$q_\infty\left(\boldsymbol{p}\right) \triangleq \lim_{\frac{P}{N_0}\to\infty} q\left(\boldsymbol{p}\right)$$
$$= \sum_{m\in\mathcal{M}} a_m \sum_{\kappa=1}^{K} \mathbb{P}\left[\mathcal{K}_0 = \kappa\right]\left(\mathcal{A}_{m,0}\,\frac{1 - \exp\left(-\pi\,T_m\lambda_F\,\mathcal{U}_\kappa\left(\boldsymbol{p}\right)R^2\right)}{\mathcal{U}_\kappa\left(\boldsymbol{p}\right)} + \mathcal{A}_{a,0}\,\frac{1 - \exp\left(-\pi\,\Lambda_m\lambda_F\,\mathcal{V}_\kappa\left(\boldsymbol{p}\right)R^2\right)}{\mathcal{G}\,\mathcal{V}_\kappa\left(\boldsymbol{p}\right)}\right) \tag{24}$$

Algorithm 1 Projected Cuckoo Search Algorithm (PCSA)

1: **initialization** set t, N_C, N_P, and P_a
2: generate a random initial population of N_P nests $\{\breve{\boldsymbol{p}}^i : i \in \{1, 2, \cdots, N_P\}\}$
3: obtain a feasible population $\{\boldsymbol{p}^i : i \in \{1, 2, \cdots, N_P\}\}$ by projecting each $\breve{\boldsymbol{p}}^i$ onto the set of the variables satisfying (1) and (2)
4: calculate the fitness value $q_\infty\left(\boldsymbol{p}^i\right)$, $\forall i \in \{1, 2, \cdots, N_P\}$
5: **while** $t \leq N_C$ **do**
6: generate new nests $\tilde{\boldsymbol{p}}^i_{new}$, $\forall i \in \{1, 2, \cdots, N_P\}$, by Lévy flights using (27)
7: obtain a feasible set of the new nest $\boldsymbol{p}^i_{new}$ by projecting each $\tilde{p}^i_{new}$ onto the set of the variables satisfying (1) and (2)
8: choose a nest j randomly among $\{\boldsymbol{p}^j : j \in \{1, 2, \cdots, N_P\}\}$
9: **if** $q_\infty\left(\boldsymbol{p}^j\right) < q_\infty\left(\boldsymbol{p}^i_{new}\right)$ **then**
10: $\boldsymbol{p}^j \leftarrow \boldsymbol{p}^i_{new}$
11: **end if**
12: abandon randomly P_a fraction of worse nests and obtain new ones by sorting the components of each abandoned nest descendingly
13: find the best nest
14: **end while**

In step 6 of Algorithm 1, the new nests in the first generation of each iteration are generated via Lévy flights as follows

$$\tilde{\boldsymbol{p}}^i_{new} = \boldsymbol{p}^i + \delta \otimes \boldsymbol{L}(\vartheta) \tag{28}$$

where $\tilde{\boldsymbol{p}}^i_{new}$ and $\boldsymbol{p}^i$ are the locations of the i-th new and current nest, respectively, δ is the scaling factor of the step size, which depends on the scale of the problem and should be chosen wisely to avoid flying too far, $\otimes$ denotes the entry-wise multiplication of two vectors, $\vartheta \in [0.3, 1.99]$ represents the Lévy distribution index, and $\boldsymbol{L}(\vartheta) = (L_j(\vartheta))_{j\in\mathcal{J}}$ denotes the Lévy random vector. According to Mantegna's algorithm [39], the components of $\boldsymbol{L}(\vartheta)$ can be calculated by

$$L_j(\vartheta) = \frac{\omega}{|\nu|^{1/\vartheta}}, \qquad \forall j \in \mathcal{J} \tag{29}$$

where $\omega \overset{d}{\sim} \mathcal{N}(0, \sigma^2_\omega)$ and $\nu \overset{d}{\sim} \mathcal{N}(0, \sigma^2_\nu)$ are random numbers drawn from normal distribution with a zero mean and variance of

$$\sigma^2_\omega = \left[\frac{\sin(\pi\vartheta/2)\,\Gamma(1+\vartheta)}{\vartheta\, 2^{(\vartheta-1)/2}\,\Gamma((1+\vartheta)/2)}\right]^{1/\vartheta} \tag{30}$$

and

$$\sigma^2_\nu = 1 \tag{31}$$

respectively, where $\Gamma(.)$ denotes gamma function.

The scaling factor δ is assumed to be decreasing with iteration index to encourage the localization of the Lévy's search, i.e.,

$$\delta = 1 - \frac{1 - (t/20M)}{N_C} \tag{32}$$

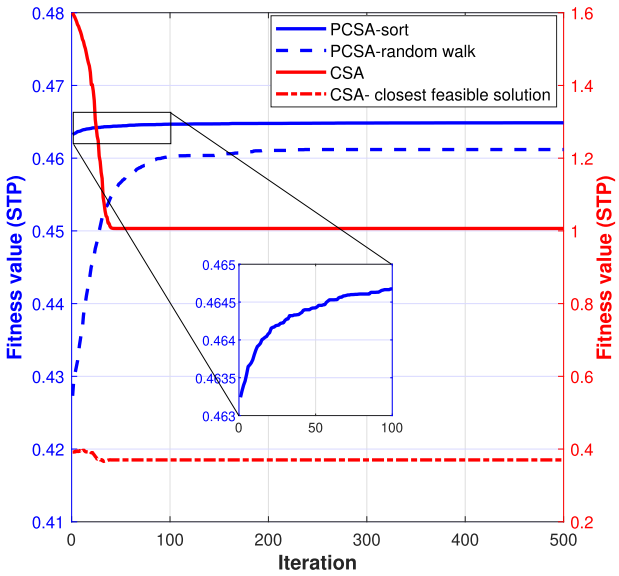

FIGURE 2. Fitness value (STP) versus iteration index at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.3$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 100$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = W_C = 10$ MHz, and $\tau = 0.5$ Mbps.

As the most popular content are more probable to be cached, in step 11, we encourage PCSA to converge faster by generating the second generation of the nests by sorting the components of the abandoned nest in a descending order instead of generating it via random walk as in the original CSA.

In each iteration of PCSA, the time complexity of the projection optimization in steps 3 and 7 is at most $\mathcal{O}(N_P \times I)$, where I is the maximum number of function evaluations in the projection optimization operation. The time complexity of the sort operation in step 12 is $\mathcal{O}(N_P \times J^2)$. whereas, the time complexity of the rest of the algorithm is $\mathcal{O}(N_P \times J)$. However, it is essential that $I > J^2$ to efficiently obtain a feasible nest by the projection optimization. Thus, the time complexity of PCSA in each iteration after neglecting the low order terms is $\mathcal{O}(N_P \times I)$. As there are N_C iterations, the overall time complexity of PCSA becomes $\mathcal{O}(N_C \times N_P \times I)$.

IV. NUMERICAL RESULTS

In this section, we first evaluate the performance of proposed PCSA. Then, we verify the asymptotic approximation of the STP presented in the previous section. Finally, the performance of the proposed joint random caching and multicasting scheme is evaluated.

The parameters of PCSA used in this work are $N_C = 500$, $N_P = 100$, $\vartheta = 1.5$, and $P_a = 0.25$. To evaluate the performance of PCSA, it is compared with CSA presented in [40] with penalty factor of 100 and the closest feasible solution to the solution obtained using CSA. The parameters of CSA were set to be the same as PCSA. Fig. 2 plots the average fitness value (i.e., STP) of the asymptotic caching and multicasting scheme obtained over 20 trails versus the

TABLE 2. Statistical results of PCSA and CSA, where * indicates an infeasible solution.

	PCSA		CSA	
	Sort	Random walk	Original	Closest feasible
Average	0.4647	0.4593	1.0337	0.3712
Median	0.4648	0.4614	1.0634	0.3712
St. Dev.	0.0032	0.0065	0.2159	0.0362
Worst	0.4645	0.4535	1.2968*	0.2841
Best	0.4651	0.4646	1.0423*	0.4276

iteration index. The figure shows that the average fitness values of CSA are infeasible (i.e., $STP > 1$). Whereas, the closest feasible nests to the nests obtained by CSA has a poor performance compared with the proposed PCSA. The figure also depicts that the PCSA with sorting the components of the abandoned nest in a descending order converges faster and outperforms the PCSA with random walk-based search. The statistical results of the performance of PCSA and CSA provided in Table 1 indicate that PCSA with sorting the components of the abandoned nest in a descending order outperforms the other algorithms.

Fig. 3 plots the STP versus the transmission SNR P/N_0. The simulation curves obtained by performing Monte Carlo simulations 50000 times show that the proposed caching strategy with multicast transmission service always achieves higher STPs than unicast transmission scheme. Moreover, the figure shows that the curves of the analytical and simulation curves of the proposed caching and multicasting scheme with general transmit SNR are close. Also, it shows that the analytical curve of the general transmit SNR caching and multicasting scheme approaches the analytical asymptotic curve, which verifies Theorem 3 and Corollary 1 and reveals that $q_\infty(\boldsymbol{p})$ provides a good approximation of $q(\boldsymbol{p})$ when the transmission SNR is high, i.e., $P/N_0 \geq 40$ dB. Accordingly, we focus on the asymptotic caching optimization as $P/N_0 \to \infty$ in the rest of this section.

In Fig. 4–11, the performance of the proposed asymptotic caching and multicasting strategy for F-RAN using PCSA is compared with two benchmark strategies. The first benchmark is 'Uniform', which refers to caching strategy proposed by [41], in which the F-APs randomly cache a combination according to the uniform distribution, i.e., the probability of caching each combination is $p_j = \frac{1}{J}, \forall j \in \mathcal{J}$. The second benchmark is 'Popular' caching strategy, in which only the most K popular contents are cached at each F-AP [42]. It is assumed that the benchmark schemes adopt the same association, multicasting, and wireless backhauling models of the proposed scheme. Note that the requested contents except the K top ranked contents are served by a transit F-AP in Popular scheme. Whereas, the contents are often served by the direct F-APs in Uniform scheme.

Fig. 4 illustrates the relationship between the STP and threshold transmission rate τ. The figure shows a decrease in the STP with the threshold transmission rate for all the considered schemes. It is observed that Uniform scheme achieves lower STPs than Popular scheme when $\tau \geq 0.2$ Mbps, which

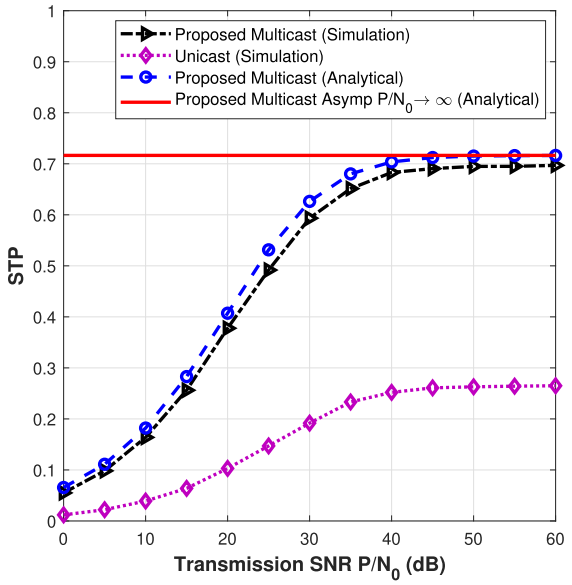

FIGURE 3. STP versus transmission SNR P/N_0 at $M = 5$ contents, $K = 3$ contents, $\lambda_U = 0.03$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 30$ m, $\gamma = 0.5$, $\alpha = 4$, $W_F = W_C = 10$ MHz, and $\tau = 0.5$ Mbps.

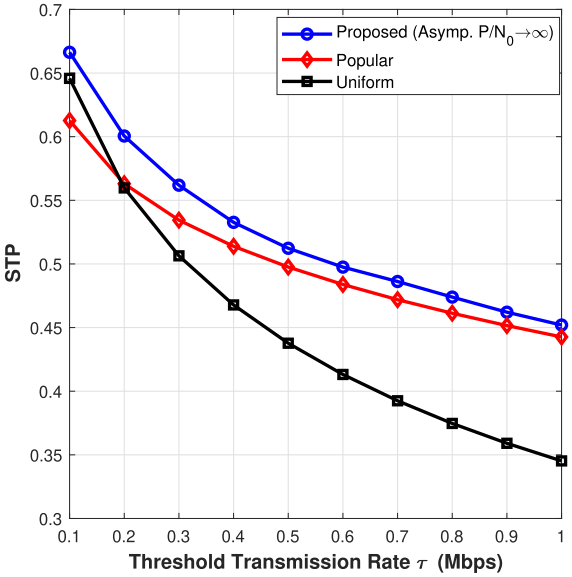

FIGURE 4. STP versus threshold transmission rate τ at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, and $W_C = 50$ MHz.

is due to the content that is served by direct F-APs with high separation distance from the requester in Uniform scheme has a lower STP than severing it by a 2-hop transmission using a closer transit F-AP as in Popular scheme. It is also observed that the proposed asymptotic caching schemes outperforms the benchmark schemes as a result of optimizing the cache placement.

Fig. 5 illustrates the relationship between the STP and discovery range R. We can observe a logistic growth in the

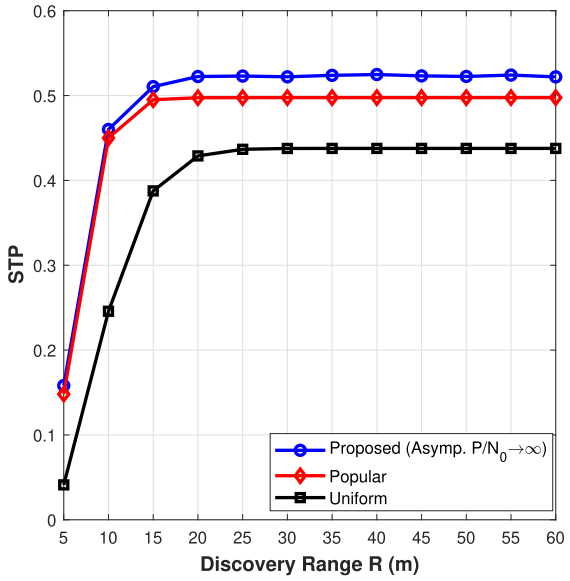

FIGURE 5. STP versus discovery range R at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

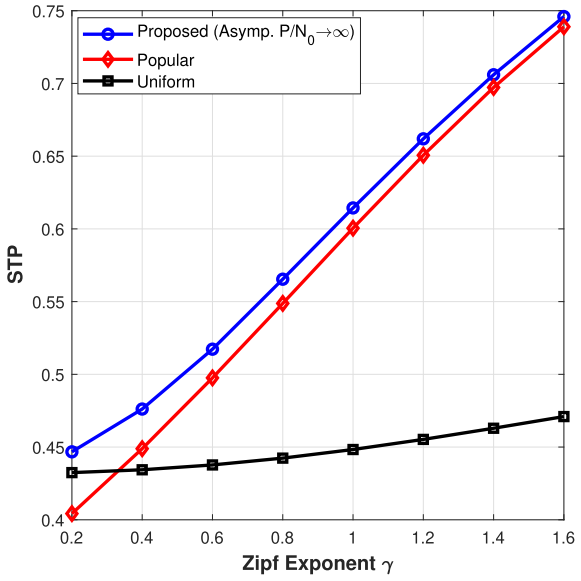

FIGURE 6. STP versus Zipf exponent γ at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

STP with the discovery range for all schemes, which is due to the higher probability of association with a F-AP to serve the requested contents as there are more F-APs residing within the discovery range. Moreover, the figure shows that optimizing that cache placement in the proposed scheme improves the STP by up to 15% over the benchmark schemes when the discovery range is higher than 20 m.

Fig. 6 illustrates the relationship between the STP and Zipf exponent γ. The figure shows an increase in the STP for all schemes with Zipf exponent, which is due to the higher probabilities of requesting the high ranked contents with the increase in Zipf exponent. However, Uniform scheme has a lower increasing trend than the other schemes, since the number of the cached high ranked contents within R is lower as a result of caching the contents evenly at the F-APs. Fig. 6 also shows that the proposed scheme outperforms the benchmark schemes due to optimizing the cache placement.

Fig. 7 plots the STP versus the F-APs' transmission bandwidth W_F. An increase in the STP with the F-APs' transmission bandwidth is observed for all the considered schemes, which is due to the improvement in disseminating the contents by the F-APs with the increase in their transmissions bandwidth. However, Uniform scheme is affected more than Popular scheme by the F-APs' transmission bandwidth as the contents are often served by the direct F-APs, i.e., a single-hop transmission is often required to deliver the requested content. It is also observed that the proposed scheme performs better than the benchmark schemes, which is due to the optimum utilization of the F-APs as direct or transit as a result of optimizing the cache placement.

Fig. 8 illustrates the relationship between the STP and the C-APs' transmission bandwidth W_C. Fig. 8 shows an

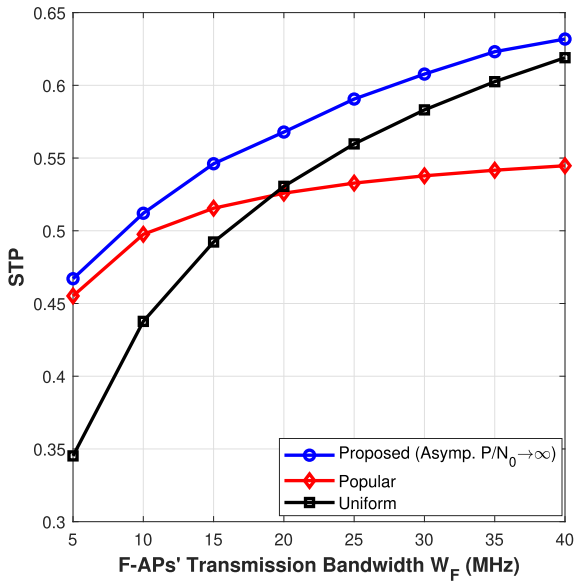

FIGURE 7. STP versus F-APs' transmission bandwidth W_F at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

improvement in the STP with the C-APs' transmission bandwidth for the proposed and Popular schemes, which is due to the higher performance of fetching the contents from C-APs by the transit F-APs. Whereas, Fig. 8 shows that the C-APs' transmission bandwidth has almost no impact on the performance of Uniform scheme, which is due to the extremely low probability of operating the F-APs in the transit mode. Fig. 8 also shows that the proposed caching schemes outperform the benchmark schemes as a result of the

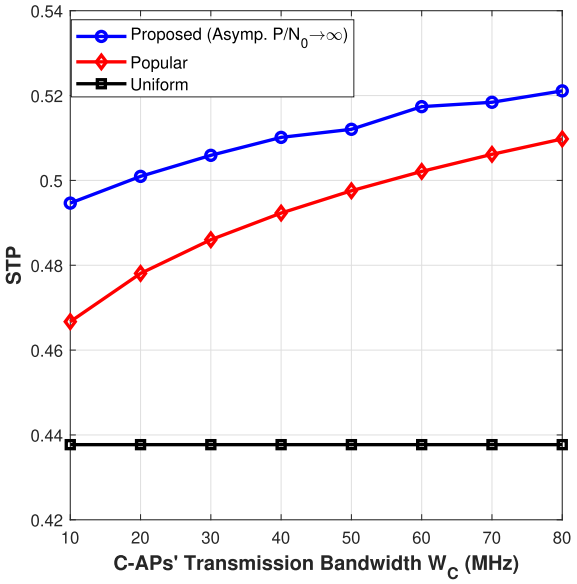

FIGURE 8. STP versus C-APs' transmission bandwidth W_C at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

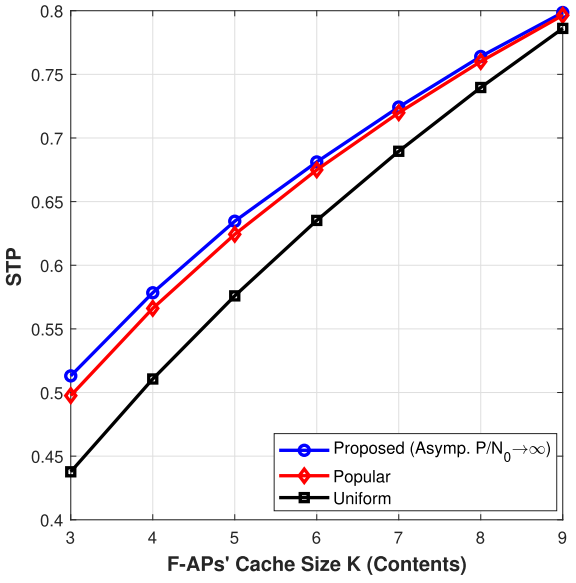

FIGURE 9. STP versus F-APs' cache size K at $M = 10$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

optimizing the operation modes of the F-APs by optimizing the cache placement.

Fig. 9 plots the STP versus the F-APs' cache size K. It can be seen that the proposed scheme performs better than benchmark schemes and the STP increases with K for all schemes. Here, the improvement in the STP is due to caching more contents at each F-AP, which results in an increase in the probability of finding a direct F-AP that caches the requested content in the vicinity of the requester. Also, caching more contents at each F-AP leads to an increase the probability of inactive contents, and thus an increase in the probability of the transit F-APs. It can be also seen that the differences in the STP between all caching schemes reduce with the increase in the cache size. This is mainly due to higher probability that the combinations of the different caching schemes contain higher number of common contents.

Fig. 10 plots the STP versus the total number of contents M. We can see that the STP of all schemes degrades with the total number of contents, which is due to the degradation in the popularity of each content, the lower number of contents that are cached within the discovery range, which results in lower probabilities of the direct and transit F-APs, and the poor performance of disseminating the contents by the C-APs since their transmission bandwidth is evenly divided over a higher number of contents. We can also see that the impact of total number of contents on Uniform scheme is higher than the other schemes, which is due to the extremely low probability of serving the contents by the transit F-APs. Fig. 10 shows that the proposed caching and multicasting scheme achieves higher STPs than the benchmark schemes as a result of optimizing the cache placement, which in turns results in optimizing the operation modes of the F-APs.

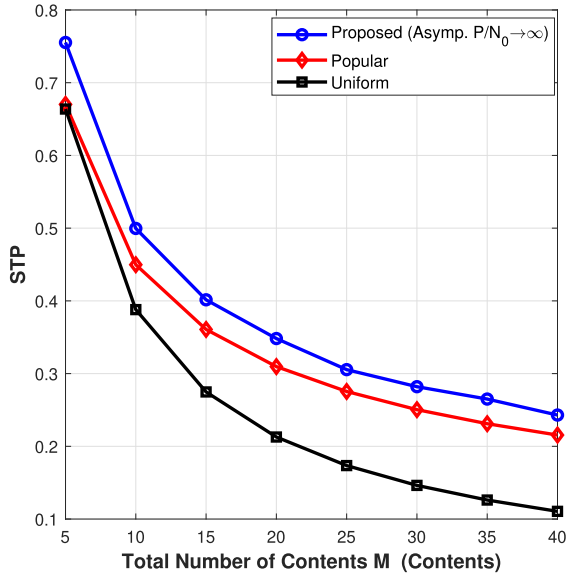

FIGURE 10. STP versus total number of contents M at $K = 2$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

Fig. 11 illustrates the STP versus the end-user density λ_U. The figure demonstrates an exponential decay in the STP for all schemes with the increase of the end-user density, which is due to the higher probability of requesting the contents with the increase in the end-user density, which in turn results in a decrease in the probability of the inactive contents, and thus a decrease in the probability of finding a transit F-AP vicinity of the requester and a lower assigned bandwidth per content due to the adopted multicast scheme. Fig. 11 shows that the proposed scheme outperforms the benchmark schemes by up to 15% at high end-user densities (i.e., $\lambda_U > 0.2$ end-users/m^2).

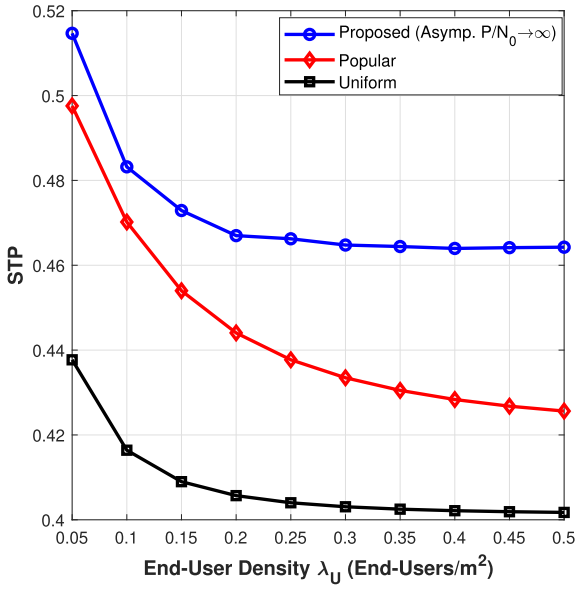

FIGURE 11. STP versus end-user density λ_U at $M = 10$ contents, $K = 3$ contents, $\lambda_F = 0.01$ F-APs/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

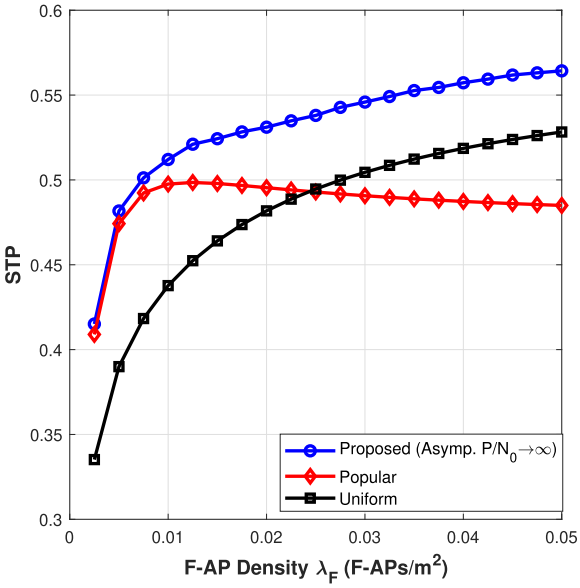

FIGURE 12. STP versus F-AP density λ_F at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_C = 0.001$ C-APs/m^2, $R = 50$ m, $\gamma = 0.6$, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

Fig. 12 plots the STP versus the F-AP density λ_F. The figure shows the proposed scheme outperforms the benchmark schemes and the STP of the proposed and Uniform schemes improves with the F-AP density, which is owing to the decrease in the distance between the end-user that is requesting a content and its serving F-AP and the higher probability of the transit F-APs as there are higher number of F-APs residing within the discovery range of the requester. Whereas, the STP of Popular scheme increases first for the same reasons, then it decreases due the high generated interference from the F-APs that degrades the performance of fetching the contents except the K most popular ones from the C-APs.

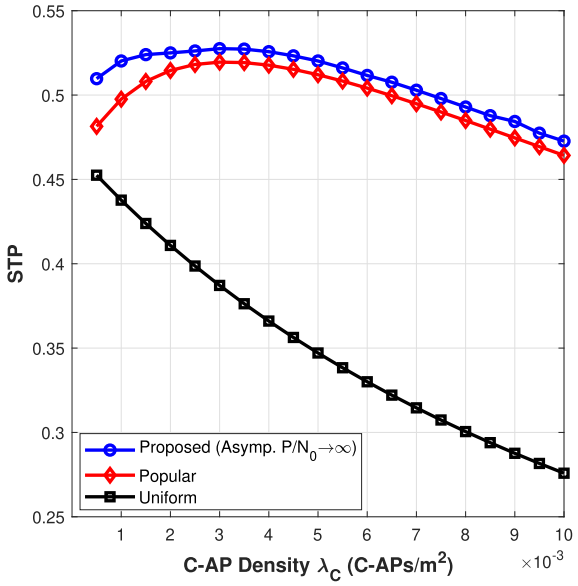

FIGURE 13. STP versus C-AP density λ_C at $M = 10$ contents, $K = 3$ contents, $\lambda_U = 0.05$ end-users/m^2, $\lambda_F = 0.01$ F-APs/m^2, $\gamma = 0.6$, $R = 50$ m, $\alpha = 4$, $W_F = 10$ MHz, $W_C = 50$ MHz, and $\tau = 0.5$ Mbps.

Fig. 13 illustrates the relationship between the STP and C-AP density λ_C. Fig. 13 shows that the STP of the proposed and Popular schemes increases first with C-AP density as a result of the improvement in the performance of fetching the contents from the C-APs by the transit F-APs. Then, it decreases when the gained improvement in the backhaul links cannot compensate the performance degradation on the links between the F-APs and the end-users caused by the higher generated interference from the C-APs. This is also the same reason of the observed decrease in the STP of Uniform scheme as the contents have an extremely low probability to be the transit F-APs. Finally, the figure shows that the proposed scheme performs better than the benchmark caching schemes due to the optimal utilization of the F-APs gained by the optimal cache placement.

V. CONCLUSION

In this paper, a probabilistic caching and multicasting scheme for F-RANs was proposed. The proposed scheme is based on the content combinations, and takes onto account the wireless backhauling of the F-APs. In order to derive the STP, we first derived the probability of a F-AP being a transit or direct F-AP with respect to the requested content and the association probabilities with those types of F-APs. Then, an expression of the STP in the general SNR region has been derived using tools from stochastic geometry and by carefully analyzing the different types of serving F-APs and interferers. The complexity was reduced by deriving a closed-form expression of asymptotic STP in the high SNR region. The optimization problem was formulated to maximize the asymptotic STP of the typical end-user. We developed the novel PCSA to solve the optimization problem and obtain the optimal cache placement. The numerical simulation results showed that PCSA converges faster than the original CSA and outperforms it.

The performance of the asymptotically random caching and multicasting scheme was also analyzed, where the results have shown that the proposed scheme achieves higher STPs than the well-known benchmark caching schemes at all the studied network's parameters.

APPENDIX A PROOF OF (6)

The probability of serving u_0 by a direct F-AP when it requests content m $\mathcal{A}_{m,0}$ is defined as the probability that there is at least one F-AP caches content m within R. Then, $\mathcal{A}_{m,0}$ can be expressed as follows

$$\begin{aligned}\mathcal{A}_{m,0} &\triangleq \mathbb{P}[N(F_m) > 0] \\ &= 1 - \mathbb{P}[N(F_m) = 0] \\ &\overset{(a)}{=} 1 - \exp\left(-\pi T_m \lambda_F R^2\right)\end{aligned} \tag{33}$$

where $N(F_m)$ denotes the number of F-APs caching content m within R, and equality (a) is obtained by the null property of PPPs, i.e., $\mathbb{P}[N(F_m) = 0] = \exp\left(-\pi T_m \lambda_F R^2\right)$.

APPENDIX B PROOF OF (8)

A transit F-AP is a F-AP that does not cache the requested content m and caches inactive contents. Thus, the probability of a F-AP being a transit one with respect to content m can be formulated as

$$\begin{aligned}\Lambda_m &= \mathbb{P}[\text{F-AP does not cache content } m, \\ &\quad \times \text{caches inactive content} \neq m] \\ &\overset{(b)}{=} \mathbb{P}[\text{F-AP does not cache content } m] \\ &\quad \times \mathbb{P}[\text{F-AP caches inactive content} \neq m] \\ &\overset{(c)}{=} (1 - T_m) \frac{\sum_{\mu \in \mathcal{M} \setminus m} T_\mu b_\mu}{K - T_m}\end{aligned} \tag{34}$$

Equality (b) is due to the fact that the two events are independent. Whereas, (c) is obtained by noting that the probability that a F-AP is caching the inactive content $\mu \in \mathcal{M} \setminus m$ is $T_\mu b_\mu$. Thus, $\mathbb{P}[\text{F-AP caches inactive content} \neq m]$ can be calculated by summing over the set $\mathcal{M} \setminus m$, then normalizing it by dividing by the sum of the probabilities of all contents in the set $\mathcal{M} \setminus m$ (i.e., $\sum_{\mu \in \mathcal{M} \setminus m} T_\mu = K - T_m$).

APPENDIX C PROOF OF (9)

The probability of serving u_0 by a transit F-AP when content m is requested $\mathcal{A}_{a,0}$ is defined as the probability of there is no F-AP caches content m and at least one F-AP caches an inactive contents within R. Thus, $\mathcal{A}_{a,0}$ can be expressed as follows

$$\begin{aligned}\mathcal{A}_{a,0} &= \mathbb{P}\left[N(F_m) = 0, N(F_a) > 0\right] \\ &\overset{(d)}{=} \mathbb{P}[N(F_m) = 0]\mathbb{P}[N(F_a) > 0] \\ &= \mathbb{P}[N(F_m) = 0]\left(1 - \mathbb{P}[N(F_a) = 0]\right) \\ &\overset{(e)}{=} \exp\left(-\pi T_m \lambda_F R^2\right)\left(1 - \exp\left(-\pi \Lambda_m \lambda_F R^2\right)\right)\end{aligned} \tag{35}$$

where $N(F_a)$ is number of F-APs that cache inactive contents. Here, equalities (d) and (e) are obtained by noting that $N(F_m)$ and $N(F_a)$ are independent events, and the null property of PPPs, respectively.

APPENDIX D PROOF OF (13)

Given that $F_{m,0}$ caches combination $j \in \mathcal{J}_m$, the content load of $F_{m,0}$ can be expressed as $\mathcal{K}_0 = 1 + \sum_{\mu \in \mathcal{M}_{j,-m}} Y_\mu$, where Y_μ is defined as in section IV. Note that content m is already requested by u_0. Here, the conditional PMF of $\mathcal{K}_0$ follows Poisson binomial distribution [43]. Thus, we have

$$\begin{aligned}&\mathbb{P}\left[\mathcal{K}_0 = \kappa | F_{m,0} \text{ caches comb. } j \in \mathcal{J}_m\right] \\ &= \sum_{\chi \in \{\mathcal{S} \subseteq \mathcal{M}_{j,-m} : |\mathcal{S}| = \kappa - 1\}} \prod_{\mu \in \chi} (1 - b_\mu) \prod_{\mu \in \mathcal{M}_{j,-m}/\chi} b_\mu\end{aligned} \tag{36}$$

where b_μ is given by (7). The conditional probability of caching combination j at $F_{m,0}$ is $\frac{p_j}{T_m}$. Thus, by summing over $\mathcal{J}_m$ (i.e., the set of all possible combinations containing content m) we can prove (13).

APPENDIX E PROOF OF THEOREM 1

Let $I_m \triangleq \sum_{l \in \Phi_m \setminus F_{m,0}} D_{l,0}^{-\alpha} |h_{l,0}|^2$ represent the interference originating from the F-APs storing content m, $I_{-m} \triangleq \sum_{l \in \Phi_{-m}} D_{l,0}^{-\alpha} |h_{l,0}|^2$ denote the interference originating from the F-APs that do not cache content m, and $I_C \triangleq \sum_{l \in \Phi_C} D_{l,0}^{-\alpha} |h_{l,0}|^2$ denote the interference originating from the C-APs. Note that Φ_m and Φ_{-m} are thinned PPPs with densities $T_m \lambda_F$ and $(1 - T_m)\lambda_F$, respectively. Next, we calculate $q_{\kappa,m,0}$ conditioned on $D_{0,0} = d$ as in (37), as shown at the bottom of the next page, where $s = \left(2^{\frac{\kappa \tau}{W_F}} - 1\right) d^\alpha$, (f) is due to $|h|^2 \overset{d}{\sim} \exp(1)$, and (g) flowed from the independence of the PPPs and independence Rayleigh fading channels. Here, $\mathcal{L}_{I_m}(s, d)$, $\mathcal{L}_{I_{-m}}(s, d)$, and $\mathcal{L}_{I_C}(s, d)$ represent the Laplace transform of I_m, I_{-m}, and I_C, respectively.

To proceed, $\mathcal{L}_{I_m}(s, d)$ can be calculated as in (38), as shown at the bottom of the next page, where the probability generating functional of PPP is utilized to obtain (h) [32]. Whereas, equality (i) is obtained by the change of variables in the integral, i.e., $sr^{-1/\alpha}$ to t, then $1/(1 + t^{-\alpha})$ to w. In the same manner, $\mathcal{L}_{I_{-m}}(s, d)$ and $\mathcal{L}_{I_C}(s, d)$ can be obtained as in (39) and (40), as shown at the bottom of the next page, respectively.

Finally, noting that the PDF of $D_{0,0}$ is $f_{D_{0,0}}(d) = 2\pi T_m \lambda_F d \exp\left(-\pi T_m \lambda_F d^2\right)$ where $d \in [0, R]$, due to Φ_m is a homogeneous PPP with density $T_m \lambda_F$. Therefore, the condition on $D_{0,0} = d$ can be removed to calculate $q_{\kappa,m,0}$ as follows

$$q_{\kappa,m,0}(\boldsymbol{p}) = \int_0^R q_{\kappa,m,0,D_{0,0}}(\boldsymbol{p}, d) f_{D_{0,0}}(d) \mathrm{d}d \tag{41}$$

Therefore, the proof is completed.

APPENDIX F PROOF OF THEOREM 2

Let $I_a \triangleq \sum_{l \in \Phi_{a,m} \setminus F_{a,0}} D_{l,0}^{-\alpha} |h_{l,0}|^2$, $I_{-a} \triangleq \sum_{l \in \Phi_{-a,m}} D_{l,0}^{-\alpha} |h_{l,0}|^2$, $I_{C_0} \triangleq \sum_{l \in \Phi_C \setminus C_0} D_{l,0}^{-\alpha} |h_{l,0}|^2$, and $I_F \triangleq \sum_{l \in \Phi_F} D_{l,0}^{-\alpha} |h_{l,0}|^2$. Here, the densities of $\Phi_{a,m}$, and $\Phi_{-a,m}$ are $\Lambda_m \lambda_F$, and $(1 - \Lambda_m)\lambda_F$, respectively.

Following the same steps of Appendix E, $q_{\kappa,a,0,D_{a,0}}(\boldsymbol{p},d)$ and $q_{C,a,D_{C,a}}(\boldsymbol{p},d)$ can be obtained as in (42) and (43), as shown at the bottom of the page, respectively, where $s=\left(2^{\frac{\kappa\tau}{W_F}}-1\right)d^{\alpha}$, and $\tilde{s}=\left(2^{\frac{M\tau}{W_C}}-1\right)d^{\alpha}$. Then, the Laplace transforms of the interference can be calculated as in (44–48), as shown at the bottom of the page.

Next, we remove the conditions on the distance to obtain $q_{\kappa,a,0,}(\boldsymbol{p})$ and $q_{C,a}(\boldsymbol{p})$ as follows

$$q_{\kappa,a,0}(\boldsymbol{p})=\int_0^R q_{\kappa,a,0,D_{a,0}}(\boldsymbol{p},d)f_{D_{a,0}}(d)\mathrm{d}d \tag{49}$$

$$q_{C,a}(\boldsymbol{p})=\int_0^{\infty} q_{C,a,D_{C,a}}(\boldsymbol{p},d)f_{D_{C,a}}(d)\mathrm{d}d \tag{50}$$

$$\begin{aligned}
q_{\kappa,m,0,D_{0,0}}(\boldsymbol{p},d) &\triangleq \mathbb{P}\left[\frac{W_F}{\kappa}\log_2\left(1+SINR_{m,0}\right)\geq\tau|D_{0,0}=d\right]\\
&=E_{I_m,I_{-m},I_C}\left[\mathbb{P}\left[|h_{0,0}|^2\geq s\left(I_m+I_{-m}+I_C+\frac{N_0}{P}\right)\right]\right]\\
&\overset{(f)}{=}E_{I_m,I_{-m},I_C}\left[\exp\left(-s\left(I_m+I_{-m}+I_C+\frac{N_0}{P}\right)\right)\right]\\
&\overset{(g)}{=}\underbrace{E_{I_m}\left[\exp\left(-sI_m\right)\right]}_{\triangleq\mathcal{L}_{I_m}(s,d)}\underbrace{E_{I_{-m}}\left[\exp\left(-sI_{-m}\right)\right]}_{\triangleq\mathcal{L}_{I_{-m}}(s,d)}\underbrace{E_{I_C}\left[\exp\left(-sI_C\right)\right]}_{\triangleq\mathcal{L}_{I_C}(s,d)}\exp\left(-s\frac{N_0}{P}\right)
\end{aligned} \tag{37}$$

$$\begin{aligned}
\mathcal{L}_{I_m}(s,d) &= E\left[\exp\left(-s\sum_{l\in\Phi_m\setminus F_{m,0}}D_{l,0}^{-\alpha}|h_{l,0}|^2\right)\right]\\
&=E\left[\prod_{l\in\Phi_m\setminus F_{m,0}}\exp\left(-sD_{l,0}^{-\alpha}|h_{l,0}|^2\right)\right]\\
&\overset{(h)}{=}\exp\left(-2\pi T_m\lambda_F\int_d^{\infty}\left(1-\frac{1}{1+s\,r^{-\alpha}}\right)r\,dr\right)\\
&\overset{(i)}{=}\exp\left(\frac{-2\pi}{\alpha}T_m\lambda_F s^{\frac{2}{\alpha}}\beta'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},\frac{1}{1+sd^{-\alpha}}\right)\right)\\
&=\exp\left(\frac{-2\pi}{\alpha}T_m\lambda_F\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}d^2\ \beta'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{\frac{-\kappa\tau}{W_F}}\right)\right)
\end{aligned} \tag{38}$$

$$\mathcal{L}_{I_{-m}}(s,d)=\exp\left(\frac{-2\pi}{\alpha}\left(1-T_m\right)\lambda_F\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}d^2\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right) \tag{39}$$

$$\mathcal{L}_{I_C}(s,d)=\exp\left(\frac{-2\pi}{\alpha}\lambda_C\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}d^2\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right) \tag{40}$$

$$q_{\kappa,a,0,D_{a,0}}(\boldsymbol{p},d)=\mathcal{L}_{I_a}(s,d)\mathcal{L}_{I_{-a}}(s,d)\mathcal{L}_{I_C}(s,d)\exp\left(-s\frac{N_0}{P}\right) \tag{42}$$

$$q_{C,a,D_{C,a}}(\boldsymbol{p},d)=\mathcal{L}_{I_{C_0}}(\tilde{s},d)\mathcal{L}_{I_F}(\tilde{s},d)\exp\left(-\tilde{s}\frac{N_0}{P}\right) \tag{43}$$

$$\mathcal{L}_{I_a}(s,d)=\exp\left(\frac{-2\pi}{\alpha}\Lambda_m\lambda_F\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}d^2\,\beta'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{\frac{-\kappa\tau}{W_F}}\right)\right) \tag{44}$$

$$\mathcal{L}_{I_{-a}}(s,d)=\exp\left(\frac{-2\pi}{\alpha}\left(1-\Lambda_m\right)\lambda_F\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}d^2\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right) \tag{45}$$

$$\mathcal{L}_{I_C}(s,d)=\exp\left(\frac{-2\pi}{\alpha}\lambda_C\left(2^{\frac{\kappa\tau}{W_F}}-1\right)^{\frac{2}{\alpha}}d^2\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right) \tag{46}$$

$$\mathcal{L}_{I_{C_0}}(\tilde{s},d)=\exp\left(\frac{-2\pi}{\alpha}\lambda_C\left(2^{\frac{M\tau}{W_C}}-1\right)^{\frac{2}{\alpha}}d^2\beta'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{\frac{-M\tau}{W_C}}\right)\right) \tag{47}$$

$$\mathcal{L}_{I_F}(\tilde{s},d)=\exp\left(\frac{-2\pi}{\alpha}\lambda_F\left(2^{\frac{M\tau}{W_C}}-1\right)^{\frac{2}{\alpha}}d^2\beta\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right) \tag{48}$$

where $f_{D_{a,0}}(d) = 2\pi\Lambda_m\lambda_F\, d\exp\left(-\pi\Lambda_m\lambda_F d^2\right)$ and $f_{D_{C,a}}(d) = 2\pi\lambda_C\, d\exp\left(-\pi\lambda_C d^2\right)$. Thus, we can prove Theorem 2.

REFERENCES

[1] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[2] Z. Li, M. L. Sichitiu, and X. Qiu, "Fog radio access network: A new wireless backhaul architecture for small cell networks," *IEEE Access*, vol. 7, pp. 14150–14161, 2019.

[3] A. Bani-Bakr, K. Dimyati, M. N. Hindia, W. R. Wong, and M. A. Imran, "Feasibility study of 28 GHz and 38 GHz millimeter-wave technologies for fog radio access networks using multi-slope path loss model," *Phys. Commun.*, vol. 47, Aug. 2021, Art. no. 101401. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1874490721001385

[4] A. Bani-Bakr, K. Dimyati, M. N. Hindia, W. R. Wong, A. Al-Omari, Y. A. Sambo, and M. A. Imran, "Optimizing the number of fog nodes for finite fog radio access networks under multi-slope path loss model," *Electronics*, vol. 9, no. 12, p. 2175, Dec. 2020, doi: 10.3390/electronics9122175.

[5] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: Overview and Rel-11 enhancements," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[6] Y. Wang, Y. Zhang, X. Han, P. Wang, C. Xu, J. Horton, and J. Culberson, "Cost-driven data caching in the cloud: An algorithmic approach," in *Proc. IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.

[7] V. Balasubramanian, M. Aloqaily, and M. Reisslein, "FedCo: A federated learning controller for content management in multi-party edge systems," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2021, pp. 1–9.

[8] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commu.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.

[9] M. Emara, H. ElSawy, S. Sorour, S. Al-Ghadhban, M.-S. Alouini, andT. Y. Al-Naffouri, "Flexible design of millimeter-wave cache enabled fog networks," in *Proc. IEEE Globecom Workshops*, Dec. 2018, pp. 1–7.

[10] A. Peng, Y. Jiang, M. Bennis, F.-C. Zheng, and X. You, "Performance analysis and caching design in fog radio access networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[11] Y. Jiang, M. Ma, M. Bennis, F.-C. Zheng, and X. You, "User preference learning-based edge caching for fog radio access network," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1268–1283, Feb. 2019.

[12] R. Wang, R. Li, P. Wang, and E. Liu, "Analysis and optimization of caching in fog radio access networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8279–8283, Aug. 2019.

[13] Y. Jiang, Y. Hu, M. Bennis, F.-C. Zheng, and X. You, "A mean field game-based distributed edge caching in fog radio access networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1567–1580, Mar. 2020.

[14] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled IoT using natural actor-critic deep reinforcement learning," *IEEE Internet Things J.* vol. 6, no. 2, pp. 2061–2073, Apr. 2019.

[15] S. Yan, M. Jiao, Y. Zhou, M. Peng, and M. Daneshmand, "Machine-learning approach for user association and content placement in fog radio access networks," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9413–9425, Oct. 2020.

[16] Y. Sun, M. Peng, and S. Mao, "A game-theoretic approach to cache and radio resource management in fog radio access networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10145–10159, Oct. 2019.

[17] Y. Jiang, X. Cui, M. Bennis, F.-C. Zheng, B. Fan, and X. You, "Cooperative caching in fog radio access networks: A graph-based approach," *IET Commun.*, vol. 13, no. 20, pp. 3519–3528, 2019.

[18] Y. Jiang, C. Wan, M. Tao, and F. C. Zheng, "Analysis and optimization of fog radio access networks with hybrid caching: Delay and energy efficiency," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 69–82, Jan. 2020.

[19] B. Guo, X. Zhang, Q. Sheng, and H. Yang, "Dueling deep-q-network based delay-aware cache update policy for mobile users in fog radio access networks," *IEEE Access*, vol. 8, pp. 7131–7141, 2020.

[20] Q. Li, Y. Zhang, Y. Li, Y. Xiao, and X. Ge, "Capacity-aware edge caching in fog computing networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 9244–9248, Aug. 2020.

[21] Y. Jiang, A. Peng, C. Wan, Y. Cui, X. You, F.-C. Zheng, and S. Jin, "Analysis and optimization of cache-enabled fog radio access networks: Successful transmission probability, fractional offloaded traffic and delay," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5219–5231, May 2020.

[22] G. M. S. Rahman, M. Peng, S. Yan, and T. Dang, "Learning based joint cache and power allocation in fog radio access networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4401–4411, Apr. 2020.

[23] A. Bani-Bakr, M. N. Hindia, K. Dimyati, E. Hanafi, and T. F. Tengku Mohmed Noor Izam, "Multi-objective caching optimization for wireless backhauled fog radio access network," *Symmetry*, vol. 13, no. 4, p. 708, Apr. 2021, doi: 10.3390/sym13040708.

[24] A. Bani-Bakr, K. Dimyati, M. N. Hindia, W. R. Wong, and T. F. T. M. N. Izam, "Joint successful transmission probability, delay, and energy efficiency caching optimization in fog radio access network," *Electron.*, vol. 10, no. 15, p. 1847, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/15/1847

[25] J. Kim, D. Yu, S. Moon, and S. Park, "Grouped noma multicast transmission for f-ran with wireless fronthaul and edge caching," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2019, pp. 145–149.

[26] K. Wang, J. Li, Y. Yang, W. Chen, and L. Hanzo, "Content-centric heterogeneous fog networks relying on energy efficiency optimization," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13579–13592, Nov. 2020.

[27] X. Wei, "Joint caching and multicast for wireless fronthaulin fog radio access networks," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[28] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.

[29] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, Jan. 2017.

[30] Y. Cui, Z. Wang, Y. Yang, F. Yang, L. Ding, and L. Qian, "Joint and competitive caching designs in large-scale multi-tier wireless multicasting networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3108–3121, Jul. 2018.

[31] D. Jiang and Y. Cui, "Analysis and optimization of caching and multicasting for multi-quality videos in large-scale wireless networks," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 4913–4927, Jul. 2019.

[32] M. Haenggi and R. K. Ganti, "Interference in large wireless networks," *Found. Trends Netw.*, vol. 3, no. 2, pp. 127–248, 2009.

[33] S. M. Yu and S. Kim, "Downlink capacity and base station density in cellular networks," in *Proc.11th Int. Symp. Workshops Modeling Optim. Mobile, Ad Hoc Wireless Netw. (WiOpt)*, 2013, pp. 119–124.

[34] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.

[35] D. N. Vo, P. Schegner, and W. Ongsakul, "Cuckoo search algorithm for non-convex economic dispatch," *IET Gener., Transmiss. Distrib.*, vol. 7, pp. 645–654, Oct. 2013.

[36] X. S. Yang and S. Deb, "Cuckoo search via lévy flights," in *Proc. World Congr. Nature Biologically Inspired Comput. (NaBIC)*, 2009, pp. 210–214.

[37] M. Khodier, "Comprehensive study of linear antenna array optimisation using the cuckoo search algorithm," *IET Microw., Antennas Propag.*, vol. 13, no. 9, pp. 1325–1333, Jul. 2019.

[38] T. T. Nguyen and D. N. Vo, "The application of one rank cuckoo search algorithm for solving economic load dispatch problems," *Appl. Soft Comput.*, vol. 37, pp. 763–773, Dec. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494615005803

[39] R. N. Mantegna, "Fast, accurate algorithm for numerical simulation of Lévy stable stochastic processes," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 49, pp. 4677–4683, May 1994. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.49.4677

[40] X.-S. Yang, *Nature-Inspired Optimization Algorithms*. Oxford, U.K.: Elsevier, 2014.

[41] S. Tamoor-ul-Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-Aho, "Modeling and analysis of content caching in wireless small cell networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2015, pp. 765–769.

[42] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 1–11, Feb. 2015.

[43] M. Emara, H. ElSawy, S. Sorour, S. Al-Ghadhban, M.-S. Alouini, and T. Y. Al-Naffouri, "Optimal caching in multicast 5G networks with opportunistic spectrum access," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7.

ALAA BANI-BAKR received the B.Sc. and M.Sc. degrees in electrical/telecommunication engineering from Mutah University, Karak, Jordan, in 2002 and 2007, respectively. He is currently pursuing the Ph.D. degree with the University of Malaya, Kuala Lumpur, Malaysia. He was a Lecturer at the Department of Electrical Engineering, Al-Baha University, Saudi Arabia, from 2010 to 2012. His research interests include fog computing, stochastic analysis, and mmWave communication systems.

MHD NOUR HINDIA (Member, IEEE) received the Ph.D. degree from the Faculty of Engineering in Telecommunication, University of Malaya, Kuala Lumpur, Malaysia, in 2015. He is currently involved with research in the field of wireless communications, especially in channel sounding, network planning, converge estimation, handover, scheduling, and quality of service enhancement for 5G networks. He is currently a Postdoctoral Fellow with the Faculty of Engineering in Telecommunication, University of Malaya. Besides that, he is involved with research with the Research Group in Modulation and Coding Scheme for Internet of Things for Future Network. He has authored or coauthored a number of science citation index journals and conference papers. He has participated as a reviewer and a committee member of a number of ISI journals and conferences.

KAHARUDIN DIMYATI (Member, IEEE) received the degree from the University of Malaya, Malaysia, in 1992, and the Ph.D. degree from the University of Wales Swansea, U.K., in 1996. He is currently a Professor with the Department of Electrical Engineering, Faculty of Engineering, University of Malaya. Since joining the university, he is actively involved in teaching, postgraduate supervision, research, and administration. To date, he has supervised 15 Ph.D. students and 32 master's students by research students. He has published over 100 journal articles. He is a Professional Engineer and a Chartered Engineer. He is a member of IET and IEICE.

ZATI BAYANI ZAWAWI received the M.Sc. degree in mobile communication systems from the University of Surrey, U.K., in 2012, and the Ph.D. degree from Imperial College London, U.K., in 2018. Her research interests include wireless communications, wireless power transfer, and wireless powered communications.

TENGKU FAIZ TENGKU MOHMED NOOR IZAM received the M.Sc. degree in electronic engineering from the University of Sheffield, U.K., in 2008, and the Ph.D. degree from the University of Surrey, U.K., in 2016. He is currently a Lecturer with the Department of Electrical Engineering, University of Malaya, Malaysia. His research interests include parasitic antenna, MIMO system with antenna selection, and wireless channel characterisation.