Wireless Caching: Cell-Free versus Small Cells

Shuaifei Chen*, Jiayi Zhang*, Emil Björnson*, Shuai Wang†, Chengwen Xing†, and Bo Ai‡

*School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China.

*Department of Electrical Engineering (ISY), Linköping University, SE 58183 Linköping, Sweden

†School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China.

‡State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China.

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500910

***Abstract*—Caching popular contents at a large number of access points and edge-clouds is a promising solution to alleviate the increasing backhaul congestion in beyond fifth-generation (B5G) networks. By integrating with cell-free massive multiple-input multiple-output (CF mMIMO), wireless caching can harness their combined virtues, i.e., almost uniform service quality, strong macro-diversity, and reduction of the data traffic from the core network. In this paper, we consider an offline cache-aided scenario with two caching strategies to minimize the total energy consumption (TEC), which are evaluated from the cache hit probability (CHP). The TEC minimization is showed to be NP-complete and, hence, dealt with a proposed greedy algorithm. An adaptive power control policy is proposed to reduce the TEC. We compare CF mMIMO with small cells in terms of the successful content delivery probability (SCDP) and TEC, respectively. The numerical results show that CF mMIMO can offer a much more uniform service, significantly higher SCDP, and lower average TEC when compared to than SC.**

I. Introduction

Handling the enormously increasing traffic load with low energy consumption (EC) is a key challenge of future networks. Since the users in a specific area/time are likely to request the same popular contents, wireless caching can reduce the traffic volume and the consequent EC in the core network [1], [2]. Wireless caching requires cooperative content transmission in addition to the content sharing among local caches [3]. Hence, the physical layer performance metrics, like spectral efficiency (SE), needs to be considered when designing a caching strategy [4].

Cache-aided wireless communication has been widely considered in small cell (SC) networks [5]–[7]. In [5], the authors analyzed the optimal cache placement strategy aiming to minimize the expected downloading time. Due to the discrete nature and NP-hardness of the problem (to cache or not), greedy algorithms are proposed. Cooperative hybrid caching was employed in [6], where the authors separated the cache space into two parts: the proportion for duplicating the most popular contents to increase the transmission diversity (named *Most Popular Content (MPC)* policy); the rest in different local caches dedicated to the less popular contents to increase the content diversity (named *Largest Content Diversity (LCD)* policy). In [7], the authors focused on the interference management, which is essential for a wireless cache-aided system. Although the literature shows that the combination of wireless caching and SC works well, the inherent defects of the cellular structure (e.g., inter-cell interference) urge us to look into the distributed network structures.

Cell-free massive MIMO (CF mMIMO) is regarded as one of the key technologies of B5G for its virtues of an almost uniform service quality, strong macro-diversity, and interference management [8]–[11]. In CF mMIMO, a large number of distributed access points (APs) are employed to cooperatively serve a relatively small number of user equipments (UEs) under the coordination of a central processing unit (CPU) [12], which is an edge-cloud computer. This distributed structure makes CF mMIMO naturally fit the wireless caching since having numerous local caches at the APs increases both the transmission and content diversity. From the perspective of CF mMIMO, wireless caching can greatly reduce the traffic load and EC via the fronthaul and backhaul, which is the bottleneck of the practical implementation of CF mMIMO [13]. Although a large body of research have investigated different aspects of CF mMIMO (e.g., massive access [14], channel estimation, and performance analysis under practical assumptions, etc.), most of them only consider the end-to-end physical layer performance. Only recently, [15] combined CF mMIMO with mobile edge computing (MEC) and studied the successful edge computing probability. To the best of our knowledge, however, cache-aided CF mMIMO has not been considered in the literature, which motivates this work. Our major contributions are summarized as follows.

- We design a cache-aided framework for CF mMIMO and SC systems with our proposed total EC (TEC) model, which takes the EC of both the physical layer transmission and content storing into consideration.
- We design an adaptive fractional power control (AFPC) policy, and two offline caching strategies aiming to minimize the TEC in the aforementioned framework, named free hybrid caching (FHC) and fixed-η hybrid caching (ηHC), respectively. These two caching strategies are then evaluated from the cache hit probability (CHP). By exploiting these strategies, we compare CF mMIMO and SC in terms of the SCDP and TEC performance.
- We formulate the TEC minimization as an integer linear program (ILP) and prove its NP-completeness. By modeling the TEC as a submodular function, we develop a greedy algorithm for the cache placement.

Notation: The $n \times n$ identity matrix is $\mathbf{I}_n$. The multivariate circularly symmetric complex Gaussian distribution with correlation matrix $\mathbf{R}$ is denoted $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$. The expected value is denoted as $\mathbb{E}\{\cdot\}$.

978-1-7281-7122-7/21/$31.00 ©2021 IEEE

II. CF mMIMO System Setup

We consider a cache-aided CF mMIMO system comprising L APs, each equipped with N antennas and a local cache. As illustrated in Fig. 1, the APs are connected to a CPU via fronthaul connections organized in a star topology; the CPU is connected via a backhaul connection to the core network. There are K single-antenna UEs in the network, each being cooperatively served by a subset of the APs in the user-centric manner [16]. We adopt the standard block fading model where the channel between UE k and AP l, denoted by $\mathbf{h}_{kl} \in \mathbb{C}^N$, is constant in time-frequency blocks of τ_c channel uses [17]. In each block, the channels are assumed to be subject to correlated Rayleigh fading, i.e.,

$$\mathbf{h}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R}_{kl}) \tag{1}$$

where $\mathbf{R}_{kl} \in \mathbb{C}^{N\times N}$ is the spatial correlation matrix and $\beta_{kl} \triangleq \mathrm{tr}(\frac{\mathbf{R}_{kl}}{N})$ is the large-scale fading coefficient that describes pathloss and shadowing [18]. We assume $\{\mathbf{R}_{kl}\}$ are known at the APs. We let $\mathcal{M}_k \subset \{1,\ldots,L\}$ denote the set of APs serving UE k, and $\mathcal{D}_l \subset \{1,\ldots,K\}$ denote the set of UEs served by AP l.

A. Cache Placement & Content Delivery

We consider a finite content library $\mathcal{F} = \{f_1,\ldots,f_M\}$ of M content files, where f_m is m-th most popular file with a normalized size of 1 Mbit. In a period of time T, each UE makes an independent request for a content with a probability according to a given popularity pattern, e.g., the Zipf distribution, which is widely used for video popularity modeling [5]–[7]. We denote by γ_m the request probability (or, the *popularity*) of content f_m, satisfying $0 \le \gamma_m \le 1$ and $\sum_{m=1}^{M}\gamma_m = 1$. Using the Zipf distribution, the request probability of f_m is [5]

$$\gamma_m = \frac{1/m^{\epsilon}}{\sum_{m'=1}^{M} 1/(m')^{\epsilon}}, \tag{2}$$

where ϵ is the skewness parameter. We let c denote the number of contents that a cache can store, assuming $c < M$ to avoid the trivial case where each cache stores all contents.

Hybrid caching strategies that exploit both MPC and LCD can balance the transmission and content diversity. Hence, we consider a CF mMIMO system with hybrid caching, where UE k requesting content f_m will be served by the a subset of APs that are selected in $\mathcal{M}_k$ based on their channel conditions. An AP in $\mathcal{M}_k$ will directly serve UE k when the desired content is stored in its local cache; otherwise, this AP has to retrieve the content from another AP or from the core network via the CPU, which causes extra EC in the content delivering.

We let $\boldsymbol{\chi} = \{\chi_{mj} \in \{0,1\} : m = 1,\ldots,M,\ j = 1,\ldots,L\}$ denote the placement strategy where $\chi_{mj} = 1$ when f_m is placed in the cache at AP l, and $\chi_{mj} = 0$ otherwise. Moreover, we denote by $\boldsymbol{\delta} = \{\delta^j_{ml} \in \{0,1\} : m = 1,\ldots,M,\ l = 1,\ldots,L,\ j = 0,1,\ldots,L\}$ the delivery strategy where $\delta^j_{ml} = 1$ when f_m is delivered from AP j to AP l, and $\delta^j_{ml} = 0$ otherwise. Especially, $j = l$ and $j = 0$ mean that AP l fetches the desired contents from its own cache and the core network via the CPU, respectively. We assume that the CPU is not equipped with a cache in this paper. Hence, the CPU can only deliver the contents between the APs and the core network. The case that the CPU has a cache is left for future work.

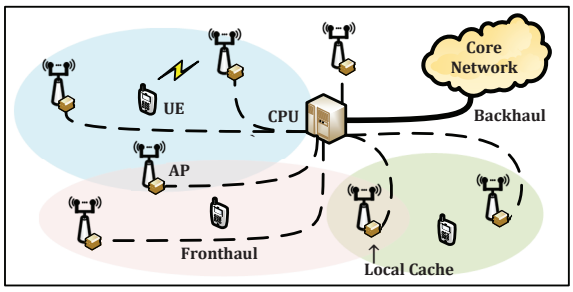

Fig. 1. A cache-aided CF mMIMO network.

We assume that the content popularity and the number of requests at each AP are known. In order to reduce the TEC, two hybrid caching strategies are proposed in this paper.

1) Free Hybrid Caching: In the FHC strategy, an AP is free to fill its cache with any content from the library to minimize the TEC. More precisely, there is no specification of which part of the cache space is supposed to be filled with the duplicated popular contents and which part is meant for the contents only stored in this cache.

In FHC, the delivery strategy $\boldsymbol{\delta}$ is determined by the placement strategy $\boldsymbol{\chi}$ and the AP selection algorithm, while the placement strategy $\boldsymbol{\chi}$ is optimized to minimize the TEC, which is elaborated in Section III.

2) Fix-η Hybrid Caching: In ηHC, the cache space in each AP is partitioned into two parts. Specifically, a fraction of η of the cache in each AP is dedicated to the most popular contents, and the fraction $1-\eta$ is reserved for distinctly storing the less popular contents at different APs to increase the content diversity. Hence, contents $\{f_m : 1 \le m \le \lfloor \eta c \rfloor\}$ are duplicated and cached in every APs in the network. For contents $\{f_m : \lfloor \eta c \rfloor < m \le \lfloor \eta c \rfloor + L(c - \lfloor \eta c \rfloor)\}$, each AP has different $c - \lfloor \eta c \rfloor$ contents at random. The rest contents $\{f_m : m > \lfloor \eta c \rfloor + L(c - \lfloor \eta c \rfloor)\}$ are not cached locally.

In ηHC, delivery strategy $\boldsymbol{\delta}$ the placement strategy $\boldsymbol{\chi}$ are determined by the AP selection algorithm and parameter η.

B. Wireless Content Transmission

Throughout this paper, we consider the downlink (DL) wireless transmission, which consists of τ_p channel uses dedicated for unlink pilots and $\tau_c - \tau_p$ for sending pieces of the content files in the DL.

1) Pilot and Content Transmission: We assume there are τ_p mutually orthogonal τ_p-length pilots, with $K > \tau_p$. We adopt the joint pilot assignment and AP selection scheme proposed in [16], where a UE chooses a *Master AP* with the best channel condition; the Master AP is entrusted to select the proper pilot and serving APs with least interference for this UE. We denote by $t_k \in \{1,\ldots,\tau_p\}$ the index of the pilot assigned to UE k, and $\mathcal{S}_k$ the set of UEs using pilot t_k. When the UEs in $\mathcal{S}_k$

transmit uplink pilot t_k, the signal $\mathbf{y}^{\mathrm{p}}_{t_k l} \in \mathbb{C}^N$ received at AP l is [17, Sec. 3]

$$\mathbf{y}^{\mathrm{p}}_{t_k l} = \sum_{i \in \mathcal{S}_k} \sqrt{\tau_p \rho_{\mathrm{p}}} \mathbf{h}_{il} + \mathbf{n}_{t_k l}, \tag{3}$$

where ρ_{p} denotes the pilot transmit power and $\mathbf{n}_{t_k l} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sigma^2 \mathbf{I}_N\right)$ is the thermal noise. The MMSE estimate of $\mathbf{h}_{il}$ is given by [17, Sec. 3]

$$\hat{\mathbf{h}}_{kl} = \sqrt{\tau_p \rho_{\mathrm{p}}} \mathbf{R}_{kl} \boldsymbol{\Psi}^{-1}_{t_k l} \mathbf{y}^{\mathrm{p}}_{t_k l}, \tag{4}$$

where $\boldsymbol{\Psi}_{t_k l} = \sum_{i \in \mathcal{S}_k} \tau_p \rho_{\mathrm{p}} \mathbf{R}_{il} + \sigma^2 \mathbf{I}_N$ is the correlation matrix of (3). The estimate is distributed as $\hat{\mathbf{h}}_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{B}_{kl}\right)$ with $\mathbf{B}_{kl} = \tau_{\mathrm{p}} \rho_{\mathrm{p}} \mathbf{R}_{kl} \boldsymbol{\Psi}^{-1}_{t_k l} \mathbf{R}_{kl}$.

Let $\mathbf{w}_{kl} = \bar{\mathbf{w}}_{kl} / \sqrt{\mathbb{E}\{\|\bar{\mathbf{w}}_{kl}\|^2\}}$ denote the normalized precoder that AP l selects for transmission to UE k, where $\mathbb{E}\{\|\mathbf{w}_{kl}\|^2\} = 1$. Then the received DL signal at UE k is

$$y^{\mathrm{dl}}_k = \sum_{l=1}^{L} \mathbf{h}^{\mathrm{T}}_{kl} \sum_{i=1}^{K} \sqrt{\rho_{il}} \mathbf{w}^*_{il} s_i + n_k, \tag{5}$$

where $s_k \in \mathbb{C}$ is the independent unit-power content signal intended for UE k, $\rho_{il} \geq 0$ is the transmit power that AP l assigns to UE i, and $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ is the receiver noise. The total transmission power of each AP is upper bounded by the maximum power ρ_{dl}.

2) Spectral Efficiency and Power Control: We employ the *hardening bound* which is widely used in mMIMO [17, Th. 4.6] and CF mMIMO to compute the DL SE.

Lemma 1. *With the MR precoder $\bar{\mathbf{w}}_{kl} = \hat{\mathbf{h}}_{kl}$, the achievable DL SE for UE k of CF mMIMO is given in* (7) *on the bottom of this page.*

Proof: It follows the approach in [17, Cor. 4.5] but with the received signal in (5) and is omitted due to the space limitation. ∎

Since there are SE differences among the UEs, one UEs will finish its transmission earlier due to the same size of the contents, which reduces the interference for the remaining UEs. With that in mind, we propose an AFPC policy where the APs stop serving a UE when this UE gets its desired content, and reallocate the transmit power to serve the remaining UEs. More precisely, AP l keeps updating its power allocation as

$$\rho_{kl} = \begin{cases} \frac{\sqrt{\beta_{kl}}}{\sum_{i \in \mathcal{D}_l / \{k^\star\}} \sqrt{\beta_{il}}} \rho_{\mathrm{dl}} & \text{if } k \in \mathcal{D}_l / \{k^\star\}, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

when the requested content has been transmitted to a generic UE $k^\star$. Consequently, the SEs of the remaining UEs are promoted by using (7) with the updated power allocation.

Since the UEs will have different SEs but request equal-sized content, they will finish their transmissions in a particular order. We let k_{f} be the index of the k_{f}th UE that finishes receiving its 1 Mbit content, for $k_{\mathrm{f}} = 1, \ldots, K$. We denote by $t^{\mathrm{tx}}_{k_{\mathrm{f}}} = \sum_{i=1}^{k_{\mathrm{f}}} t_i$ the transmission time used for the k_{f}th UE where the time slot t_i can be sequentially computed by

$$t_i = \frac{1 - B \cdot \sum_{n=1}^{i-1} t_n \mathsf{SE}^{(n)}_i}{B \cdot \mathsf{SE}^{(i)}_i}, \tag{8}$$

where $\mathsf{SE}^{(n)}_i$ is the updated SE of the ith UE during time slot t_n and B is the system bandwidth. The contents should be transmitted to the UEs on time, i.e., $t^{\mathrm{tx}}_{k_{\mathrm{f}}} \leq T$, for $k_{\mathrm{f}} = 1, \ldots, K$, otherwise, the transmission fails.

3) Energy Consumption Model: The TEC $\mathsf{E}_{\mathrm{total}}$ is assumed to have two major contributors: the one caused by the physical layer transmission and the other caused by caching itself, as

$$\mathsf{E}_{\mathrm{total}} = \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{j=0}^{L} \gamma_{ml} \mathsf{SV}^j_{ml} \delta^j_{ml} + \sum_{m=1}^{M} \sum_{j=1}^{L} \mathsf{ST}_{mj} \chi_{mj} T, \tag{9}$$

where γ_{ml} is the number of requests for content f_m generated by the UEs served by AP l in time period T. ST_{mj} is the cost caused by storing content file f_m at AP j, and SV^j_{ml} is the cost associated when AP l retrieves content f_m from cache j to serve the requesters. For $\forall m, l$, SV^j_{ml} can be further described as

$$\mathsf{SV}^j_{ml} = \begin{cases} \mathsf{E}^{\mathrm{tx}}_l & j = l \\ \mathsf{E}^{\mathrm{tx}}_l + \mathsf{E}_{\mathrm{fh}} + \mathsf{E}_{\mathrm{fh}} / (\gamma_{ml} r^{\mathrm{fh}}_m) & j \neq l, j \neq 0 \\ \mathsf{E}^{\mathrm{tx}}_l + \mathsf{E}_{\mathrm{fh}} + \mathsf{E}_{\mathrm{bh}} / (\gamma_{ml} r^{\mathrm{bh}}_m) & j = 0 \end{cases}, \tag{10}$$

where $r^{\mathrm{fh}}_m = \sum_{l=0}^{L} \delta^j_{ml}$ and $r^{\mathrm{bh}}_m = \sum_{l=1}^{L} \delta^0_{ml}$ are the numbers associated with when an AP retrieves content f_m from another AP via the fronthaul connection and the number associated with when an AP retrieves content f_m from the core network via the backhaul connection, respectively. The rationale behind these scalings at the denominator is that the CPU does not have to fetch the same content file several times from an AP or the core network within time period T. In (10), $\mathsf{E}^{\mathrm{tx}}_l$ is the EC caused by transmitting content f_m from AP l to its serving UEs via wireless connections. Referring to [13], $\mathsf{E}^{\mathrm{tx}}_{ml}$ can be modeled as

$$\mathsf{E}^{\mathrm{tx}}_l = t^{\mathrm{tx}}_{K^l_{\mathrm{f}}} \left(\frac{1}{\alpha_l} \sum_{i=1}^{K} \rho_{il} + \mathsf{P}^{\mathrm{tc}}_l \right), \tag{11}$$

where K^l_{f} is index of the last UE in $\mathcal{D}_l$ finishes the transmission, $0 < \alpha_l \leq 1$ is the power amplifier efficiency, and $\mathsf{P}^{\mathrm{tc}}_l$ is the internal power required for the circuit components (e.g.,

$$\mathsf{SE}_k = \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \frac{\left| \sum_{l=1}^{L} \sqrt{\rho_{kl} \mathrm{tr}(\mathbf{B}_{kl})} \right|^2}{\sum_{i=1}^{K} \sum_{l=1}^{L} \frac{\rho_{il} \mathrm{tr}(\mathbf{B}_{il} \mathbf{R}_{kl})}{\mathrm{tr}(\mathbf{B}_{il})} + \sum_{i \in \mathcal{S}_k} \left| \sum_{l=1}^{L} \frac{\sqrt{\rho_{il}} \mathrm{tr}\left(\mathbf{B}_{kl} \mathbf{R}^{-1}_{kl} \mathbf{R}_{il}\right)}{\sqrt{\mathrm{tr}(\mathbf{B}_{il})}} \right|^2 - \left| \sum_{l=1}^{L} \sqrt{\rho_{kl} \mathrm{tr}(\mathbf{B}_{kl})} \right|^2 + \sigma^2} \right) \tag{7}$$

converters, mixers, and filters) related to AP l [19].

Moreover, E_{fh} is the EC caused by delivering an 1-Mbit content via the 1-hop fronthaul connection between the CPU and an AP [20]. E_{bh} is the EC for delivering an 1-Mbit content from the core network to the CPU via the backhaul connection.

III. Problem Formulation and Optimization

In this section, we consider the TEC minimization problem when using FHC. We first formulate this problem as an optimization problem and prove that it is NP-Complete. Then, we model this optimization as a submodular function optimization. Finally, we develop a greedy algorithm for the cache placement.

A. Total Energy Consumption Minimization

We aim to minimize the TEC caused by both serving and caching the content files. With the definition in (9), we formulate the minimization problem as an ILP:

$$\underset{\boldsymbol{\chi}}{\text{minimize}} \sum_{m=1}^{M}\sum_{l=1}^{L}\sum_{j=0}^{L}\gamma_{ml}\mathsf{SV}_{ml}^{j}\delta_{ml}^{j}+\sum_{m=1}^{M}\sum_{j=1}^{L}\mathsf{ST}_{mj}\chi_{mj}T \tag{12}$$

$$\text{s.t.} \quad \sum\nolimits_{j=0}^{L}\delta_{ml}^{j}\geq\mathbb{I}_{\{\gamma_{ml}>0\}},\forall m,l \tag{13}$$

$$\sum\nolimits_{m=1}^{M}\chi_{mj}\leq c,\forall j\neq 0 \tag{14}$$

$$\chi_{mj}\in\{0,1\},\forall m,\forall j\neq 0. \tag{15}$$

In the objective function of the above problem, the first term is the total serving EC, the second term is the total caching EC. The first set of constraints in (13) ensures that any request for contents can be served, the second set of constraints in (14) ensures that the total amount of contents stored in a cache will not exceed its capacity, the third set of constraints in (15) indicated the discrete nature of the optimization variables.

B. NP-Completeness Proof

The ILP in (12) is proved to be NP-complete by the following lemma. Before that, we first give the definition of the weighted set cover problem.

Definition 1 (Weighted Set Cover Problem)**.** *Given a set $\mathcal{A}=\{a_l : l=1,\dots L\}$ and a set $\mathcal{B}=\{b_j : j=1,\dots,J\}$, where element b_j of $\mathcal{B}$ is a subset of $\mathcal{A}$ with a cost $w_j\geq 0$. The objective is to find a set of subsets of $\mathcal{B}$, whose union is $\mathcal{A}$ and minimize the total cost.*

Lemma 2. *The considered optimization problem in* (12) *is NP-complete.*

Proof: We exploit a similar technique as the one used in [21] to rewrite the weighted set cover problem, which is known as an NP-complete problem, to an instance of our considered problem in (12).

We rewrite the weighted set cover problem to our considered problem as follows. 1) We set the size of the content library in our problem to 1, i.e., $M=1$. 2) We map each element a_l in the weighed set cover problem to an AP requesting the content in our problem. 3) We map each subset b_j to cache j, where $a_l\in b_j$ means that AP l can fetch the content from cache j. 4) We set the caching cost $\mathsf{ST}_{mj}=w_j$. 5) We set the serving cost $\mathsf{SV}_{ml}^{j}=0$ when $a_l\in b_j$, and $\mathsf{SV}_{ml}^{j}=w_j$ otherwise. It can be seen that a solution to this instance of our problem is also a solution to the weighted set cover problem. ∎

C. Submodular Function Optimization

The optimization problem in (12) can be modeled as the minimization of a submodular function, which allows us to employ a low-complexity greedy algorithm to solve the problem. We first give the definition of a submodular function.

Definition 2 (Submodular Function)**.** *Given a finite ground set $\mathcal{S}$ and denote by $g_{\mathcal{A}}(i)=g(\mathcal{A}+i)-g(\mathcal{A})$ the marginal value of element $i\in\mathcal{S}$ with respect to $A\subseteq\mathcal{S}$, a set function $g:2^{\mathcal{S}}\rightarrow\mathbb{R}$ is submodular if for all $\mathcal{A}\subseteq\mathcal{B}\subseteq\mathcal{S}$ and for all $i\in\mathcal{S}\backslash\mathcal{B}$, we have*

$$g_{\mathcal{A}}(i)\geq g_{\mathcal{B}}(i). \tag{16}$$

Intuitively, (16) captures the concept of diminishing returns: as the set becomes larger, the benefit of adding a new element to the set will decrease.

Lemma 3. *The considered objective function in* (12) *is a submodular function.*

Proof: We take a similar approach as in [5] to illustrate that the marginal value of adding a new content to any cache j decreases as the placement set $\overline{\boldsymbol{\chi}}=\{\chi_{mj}:\chi_{mj}=1,\chi_{mj}\in\boldsymbol{\chi}\}$ becomes larger. We denote by $g_{\overline{\boldsymbol{\chi}}}(\chi_{mj})=\mathsf{E}_{\mathrm{total}}(\overline{\boldsymbol{\chi}})-\mathsf{E}_{\mathrm{total}}(\overline{\boldsymbol{\chi}}\cup\chi_{mj})$ the marginal value of adding a new placement χ_{mj} to $\overline{\boldsymbol{\chi}}$, which is defined as the reduction in the TEC by placing a new content f_m at cache j.

We consider two placement set $\overline{\boldsymbol{\chi}}$ and $\widehat{\boldsymbol{\chi}}$ where $\overline{\boldsymbol{\chi}}\subset\widehat{\boldsymbol{\chi}}$. Consider a new placement $\chi_{mj}\notin\widehat{\boldsymbol{\chi}}$. For both $\overline{\boldsymbol{\chi}}$ and $\widehat{\boldsymbol{\chi}}$, adding χ_{mj} will cause the same new storing cost since content f_m is placed at the same local cache j and reduce the serving cost since the CPU would fetch f_m from local cache j instead of the core network on the other. However, the reduction in the serving cost of adding χ_{mj} to $\overline{\boldsymbol{\chi}}$ is greater than or equal to that to $\widehat{\boldsymbol{\chi}}$ since $\overline{\boldsymbol{\chi}}\subset\widehat{\boldsymbol{\chi}}$ and thus the cost of serving the request of content f_m with placement $\overline{\boldsymbol{\chi}}$ cannot be less than $\widehat{\boldsymbol{\chi}}$. To sum up, $g_{\overline{\boldsymbol{\chi}}}(\chi_{mj})\geq g_{\widehat{\boldsymbol{\chi}}}(\chi_{mj})$. ∎

We develop a simple greedy algorithm which iteratively refines the placement set $\overline{\boldsymbol{\chi}}$. The greedy algorithm works as follows. We find the placement χ_{mj} with the highest marginal value $g_{\overline{\boldsymbol{\chi}}}(\chi_{mj})$ to placement set $\overline{\boldsymbol{\chi}}$. The operation is repeated until no χ_{mj} has a positive marginal value, which indicates adding a new content to the local caches will not reduce but increase the TEC. The greedy algorithm is summarized in Algorithm 1.

Lemma 4. *The proposed greedy algorithm will achieve an approximate value for the optimization problem like* (12) *within a factor* $1/2$ *of the optimum.*

Proof: It follows the classical results in [22]. ∎

Algorithm 1: Greedy Caching for (12)

Input: $\{\gamma_{ml}\}, \{\mathsf{SE}_k\}, \{\rho_{il}\}, \{\alpha_l\}, \{\mathsf{P}_l^{\mathrm{tc}}\}, \mathsf{E}_{\mathrm{fh}}, \mathsf{E}_{\mathrm{bh}}, c$

Output: $\overline{\boldsymbol{\chi}}$

Initiation: $\overline{\boldsymbol{\chi}} = \emptyset$;

repeat

  Compute $\mathsf{E}_{\mathrm{total}}(\overline{\boldsymbol{\chi}})$ using (9);

  Find placement

$$\chi_{mj} = \arg \max_{\chi_{mj} \notin \overline{\boldsymbol{\chi}}} g_{\overline{\boldsymbol{\chi}}}(\chi_{mj}) \tag{17}$$

  with $g_{\overline{\boldsymbol{\chi}}}(\chi_{mj}) > 0$.

until $g_{\overline{\boldsymbol{\chi}}}(\chi_{mj}) \leq 0$, *or* $\sum_{m=1}^{M} \chi_{mj} = c$;

IV. Small Cell System

To compare with a conventional SC network, we assume that each UE is served by only one selected AP. More precisely, the channel estimation and DL beamforming for a UE is locally done at the selected AP without any transmission cooperation via the CPU. Nevertheless, the content cooperation remains, i.e., the content files can still be passed from an AP to another AP via the CPU.

Since the UEs do not have the knowledge of the instantaneous channel estimates in the DL SC case, an achievable DL SE should be computed using the same approach as in CF mMIMO case.

Corollary 1. *The achievable DL SE for UE k of SC shares the same expression as* (7) *of CF mMIMO in Lemma 1, the only difference is that UE k is only served by AP*

$$\ell = \arg \max_{l} \ \beta_{kl}. \tag{18}$$

V. Numerical Results

In this section, we quantitatively evaluate the performance of CF mMIMO in the considered cache-aided DL system with our proposed AFPC policy, FHC and ηHC strategies, and compare it to that of the SC systems. We consider a setup with $K = 20$ independently and uniformly distributed in a 1×1 km^2 square coverage area. The APs are randomly deployed and each of them is equipped with half-wavelength-spaced uniform linear arrays with $N = 4$ antennas. The wrap-around technique is used to approximate an infinitely large network. We employ the well-known 3GPP Urban Microcell model to compute the large-scale propagation conditions, such as pathloss and shadow fading. Unless otherwise specified, other system parameters are referred to that in [6], [12], [13], which are given in Table I.

We first evaluate our proposed cache strategies with the CHP in Fig. 2 with $M = 210$, which is defined as the probability of finding a requested content stored in the local caches. The first observation is that the CHP with ηHC decreases as the parameter η increases since a smaller η means more different contents stored in the local caches, which promotes the CHP. Moreover, the increased L allows the local caches to store more different contents when using ηHC, and hence also

TABLE I
System Parameters.

Parameters	Values
B	20 MHz
τ_c, τ_p, M, c,	200, 10, 50, 10
ϵ, η, α_l, $\forall l$	0.5, 0.8, 0.4
ρ_{p}, ρ_{dl}, $\mathsf{P}_l^{\mathrm{tc}}$, $\forall l$	100 mW, 1000 mW, 200 mW
E_{fh}	0.25×10^{-3} Joule/Mbit ($\mathsf{E}_{\mathrm{bh}} = 15 \cdot \mathsf{E}_{\mathrm{fh}}$)
ST_{mj}, $\forall m, j$	6.25×10^{-6} Watt/Mbit

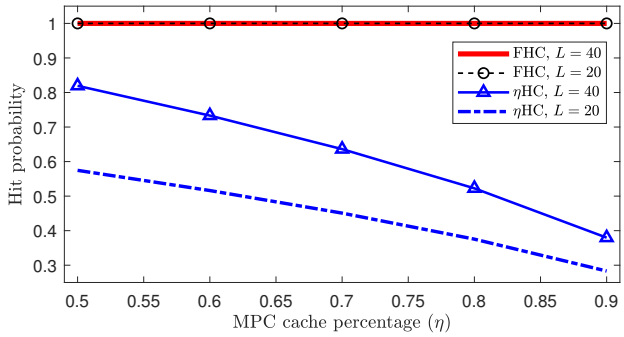

Fig. 2. Hit probability versus MPC percentage η with different caching strategy and AP number, L.

increases the CHP. Unlike ηHC, FHC offers 100% CHP in every setups of the considered η and L since Algorithm 1 makes each requested content be locally cached at the APs to minimize the TEC.

Fig. 3 demonstrates the successful content delivery probability (SCDP) of the considered system, which represents the ability to transmit the desired contents to the requesting UEs on time. Equivalently, the SCDP is defined as $\mathbb{P}\{\min_{k\in1,\ldots,K} \mathsf{SE}_k \geq \underline{\mathsf{SE}}\}$ where the target SE $\underline{\mathsf{SE}} = \frac{1}{BT}$ is an equivalent argument of the time period T. We first observe that SC loses in SCDP since there are risks that a UE has really poor channel with its only serving AP and hence fails in the content transmission. Then, we notice that the larger L promotes the SCDP in both CF mMIMO and SC due to the reduced average distance between the UEs and the APs. When comparing Fig. 3(a) and Fig. 3(b), it can be seen that our proposed AFPC policy can significantly improve the SCDP since the AFPC adjusts the transmit power allocation once the content has been delivered to benefit from the fact that the inter-user interference has disappeared.

By averaging the different random setups with random AP and UE locations, in Fig. 4, we shows the average TEC when the impacts of caching strategy and power control policy are considered with $T = 100$ ms. We first notice that CF mMIMO outperforms SC with lower TEC in every cases. The reason is that there are risks in SC that many UEs get their desired contents which are retrieved from other APs instead of their own APs, causing huge delivering EC on fronthauls. Another observation is that FHC has lower average TEC than ηHC due to the fact that all requested contents can be pre-cached using FHC, which greatly reduces the TEC. Beyond that, we notice that the TEC is much reduced when using AFPC

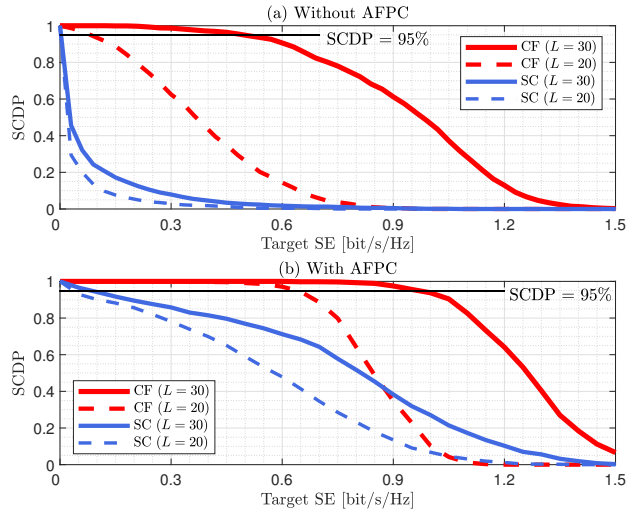

Fig. 3. Comparison between CF mMIMO and SC on the SCDP with different AP number, L and power control policy.

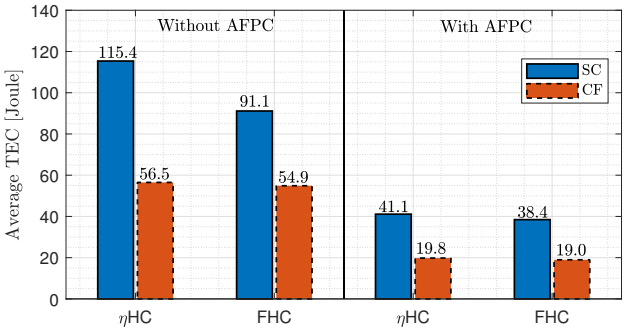

Fig. 4. Comparison between CF mMIMO and SC on the average TEC with different caching strategy and power control policy.

policy in every cases. The reason is that AFPC can reduce the transmission time of each AP by canceling interference and efficiently reallocating the DL transmit power; which reduces the EC of content transmission at the APs (see (11)). From Figs. 3 and 4, it is clear that CF mMIMO offers more uniform service to the UEs than SC.

VI. Conclusion

We investigated the CHP, SCDP and TEC performance of a cache-aided CF mMIMO system. An adaptive power control policy and two offline caching strategies were proposed. For FHC, we showed that the TEC minimization has a submodular property, which allows us to develop a greedy algorithm for the cache placement. We compared CF mMIMO and SC from both power control policy and caching strategies, respectively. Numerical results showed that CF mMIMO can outperform SC in terms of both SCDP and average TEC. To be specific, given a SCDP $= 95\%$, the target SE of CF mMIMO is almost two orders of magnitude higher than that of SC, and the average TEC of CF mMIMO is about half of that of SC. Finally, the advantages of the proposed AFPC and FHC were demonstrated in terms of the SCDP, TEC, and CHP, respectively.

References

[1] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, Mar. 2018.

[2] B. Ai, R. He, H. Zhang, M. Yang, Z. Ma, G. Sun, and Z. Zhong, "Feeder communication for integrated networks," *IEEE Wireless Commun.*, vol. 27, no. 6, pp. 20–27, Jun. 2020.

[3] W. Han, A. Liu, and V. K. Lau, "PHY-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[4] J. Zhang, H. Du, P. Zhang, J. Cheng, and L. Yang, "Performance analysis of 5G mobile relay systems for high-speed trains," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2760–2772, Dec. 2020.

[5] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[6] Z. Chen, J. Lee, T. Q. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[7] X. Xu and M. Tao, "Modeling, analysis, and optimization of caching in multi-antenna small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5454–5469, Nov. 2019.

[8] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[9] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.

[10] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, Jul. 2017.

[11] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878–99 888, Jul. 2019.

[12] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2019.

[13] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[14] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, Early Access, 2020.

[15] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2884–2899, Apr. 2020.

[16] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[17] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[18] J. Zhang, L. Dai, Z. He, B. Ai, and O. A. Dobre, "Mixed-ADC/DAC multipair massive MIMO relaying systems: Performance analysis and power optimization," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 140–153, Jan. 2019.

[19] J. Zhang, L. Dai, Z. He, S. Jin, and X. Li, "Performance analysis of mixed-ADC massive MIMO systems over Rician fading channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1327–1338, Jun. 2017.

[20] V. Sivaraman, A. Vishwanath, Z. Zhao, and C. Russell, "Profiling per-packet and per-byte energy consumption in the netfpga gigabit router," in *Proc. IEEE INFOCOM Workshops*, 2011, pp. 331–336.

[21] A. Gharaibeh, A. Khreishah, B. Ji, and M. Ayyash, "A provably efficient online collaborative caching algorithm for multicell-coordinated systems," *IEEE Trans. Mobile Comput.*, vol. 15, no. 8, pp. 1863–1876, Aug. 2015.

[22] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions," *Math. Program.*, vol. 14, no. 1, pp. 265–294, 1978.