Content-Aware Proactive Caching and Energy-Efficient Design in Clustered Small Cell Networks

Xiang Yu, Huiting Luo(✉), Long Teng, and Ting Liu

School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, People's Republic of China
luohuiting107@163.com

Abstract. This paper considers clustered small cell networks (SCNs) with combined design of cooperative caching and energy-efficient policy in the Coordinated Multi-Point (CoMP)-enabled cellular network. Small base stations (SBSs) with cache storage are grouped into associative clusters which can communicate with each other. This paper focus on movie on-demand streaming from Internet-based servers and proposed combined caching mode, where every SBS utilizes parts of cache space to cache the most popular contents (MPC), while the remaining is used for cooperatively caching different partitions of the less popular contents (LPC). Instead of the known content popularity, we constructs a content-aware weighted feature matrix (CWFM) in terms of spatiotemporal variation. Based on estimated content popularity and transmission design, we propose a caching scheme that makes a caching decision to maximize the energy efficiency (EE). To tackle this problem, A two-step stepwise optimization method is adopted. First, EE conditioning is optimized with a approach of linear programming and variable recovery. Then, the optimal proportion of cache space for MPC is analyzed by comparing the energy-efficient gain from the MPC with the energy-efficient loss from the discarded contents. Extensive simulation results confirm that our algorithm outperforms state-of-the-art algorithms based on MovieLens data set.

Keywords: Clustered SCNs · Popularity prediction · CWFM · EE · Proactive caching

1 Introduction

In recent years, the emergence of smart mobile devices and multimedia capabilities has resulted in the explosive increase of high data rate applications over wireless networks. According to a recent report from Cisco [1], global mobile data traffic will increase sevenfold between 2017 and 2022 with mobile video traffic accounting for the majority. However, current networks are unable to keep up

H. Gao et al. (Eds.): ChinaCom 2020, LNICST 352, pp. 288–308, 2021.
https://doi.org/10.1007/978-3-030-67720-6_20

with the massive growth of mobile data services in the 5G era. As for wireless networks already investigated in [2], by caching content at wireless edge, edge nodes can deliver the cached content to users directly instead of retrieving it in data center, which can significantly offloading the traffic flowing to the network. The sinking of caching and computing capability can greatly alleviate backhaul traffic and improve the cooperation efficiency of local base stations (BSs). Additionally, it also exhibits strong potential to reduce power consumption and improve energy efficiency (EE) of small cell networks (SCNs) [3].

Proactive caching in SCNs allow BSs to proactively prefetch contents from data center through backhaul links during off-peak hours and deliver the cached content to the user during peak hours. Depending on the availability and placement of the requested content, caching mode can be typically classified into two categories, namely uncoded caching [4] and coded caching [5]. Uncoded caching aims at caching complete content in each BS. Coded caching enables each BS to cooperatively caching different partitions contents. [6] demonstrated that the combination of coded caching and uncoded caching can effectively adapt to poor channels. Most the previous works on proactive caching at the SCN have been developing new methods for accurately predicting content popularity and caching the most popular contents with cooperation. [7] used a generalized Zipf distribution to model content popularity and ignored the influence by spatiotemporal variation. [8] estimated content popularity by using the request statistic of content, namely Least Recently Used (LRU) method. The multi-player multi-armed bandit (MPMAB) learning scheme studied in [9] set tradeoff between exploitation and exploration. The essence is learning in long period learning process to maximize the local content popularity or caching the unrequested content that may be popular. However, on the accuracy of predicting the content popularity, the three methods behaved poorly for ignoring learning the internal relation of user's favor and desired content. [10] utilized feature to preliminarily express content attribute, but it never fully set up the relation between user's favor and content popularity.

In heterogeneous networks, coexistence between SBSs and conventional macro BSs causes additional intercell interference when spectrum resources are shared. Coordinated Multi-Point (CoMP) technique was proposed to limit the intercell interference and cell-edge throughput by allowing geographically separated SBSs to deliver information to users cooperatively [11]. Joint transmission with CoMP in the downlink of heterogeneous cellular networks with randomly located SBSs was studied in [12], where expressions for coverage probability and diversity gain were derived for typical user by using tools from stochastic geometry. Further, recent studies in wireless edge caching with CoMP exhibited new perspectives on the benefits of caching to improve network performance. On cache-level cooperation in CoMP SCNs, the cache space of multiple SBSs in a cluster is utilized as a entity. Parts of cache space is used to cache the most popular contents (MPC) while the remaining selectively cache the less popular contents (LPC) to improve the content diversity. Considering cooperative transmission via caching manager, [13] analyzed the average cache service probability

with comprehensive consideration of joint transmission and parallel transmission, and show an optimal inherent tradeoff between transmission diversity and content diversity in cluster-centric network. Nevertheless, none of the existing works provide efficient solutions for the cache utilization policy in cooperative clustered SCNs based on unknown content popularity in prior. By using a more accurate content popularity prediction method and caching cooperatively in the clustered SCNs, the system energy efficiency can be significantly improved.

With the development of the network, SCN with intensive deployment of multiple SBS can serve more users and provide users with higher QoS. Meanwhile, multiple SBSs results in higher energy consumption. Due to the changes in network scenarios, more aspects need to be considered in the design of energy strategy, such as the unknown popularity, collaboration between multiple SBSs, and different caching methods. This paper proposes clustered SCNs with combined design of cooperative caching and energy-efficient policy based on estimated content popularity with spatiotemporal variation in order to maximize EE, which has not been considered in cluster SCNs. The SBSs are grouped into associated clusters, and the SBSs in different cluster can communicate with each other to enhance the performance of cellular network. The overall cache space within a cluster is arranged by central controller so as to either distribute the same MPC in every SBS or cache different partitions of the LPC in different SBSs. In terms of accuracy of predicting content popularity, we use the content features in popularity prediction, which can connect user's favor with content's attributes. Based on estimated popularity, the controllers in each cluster cooperatively assign cache space for the MPC while the remaining cache space for the LPC to achieve largest content diversity. Within a certain cluster, when the requested content is cached in SBSs, depending on whether the content is cached using MPC or LPC strategy, we use two transmission schemes accordingly, namely joint transmission and parallel transmission. When content is cached in MPC, it is delivered by federated transport, and when content is cached in LPC, it will be delivered using parallel transport. We model the average energy efficiency as optimization objective in the clustered SCNs, namely ratio of total transmission rate to total power. For this complex problem in clustered SCNs, we adopt a two-step stepwise optimization method, and quickly search for an inherent tradeoff between cooperation transmission and content diversity with our proposed scheme. We then maximize the average energy efficiency with optimal content placement in clusters for LPC and optimal proportion of cache space for MPC.

This paper is organized as follows. We present the network model and cooperation schemes in Sect. 2. In Sect. 3, we construct a content-aware weighted feature matrix (CWFM) to predict content popularity in terms of spatiotemporal variation. In Sect. 4, we define the average energy efficiency as the main performance metric and give its formulation. Furthermore, we analytically prove the optimization proportion of cache space for MPC in each cluster using our proposed scheme. Simulation results are presented in Sect. 5 and Sect. 6 concludes the paper.

2 System Model and Cooperation Schemes

In this work, we consider cache-enabled clustered SCNs where SBSs in each cluster are distributed according to a two-dimensional homogeneous Poisson Point Process (PPP) [14] and the distribution function is $\Phi_b = \{b_i \in \mathbb{R}^2, \forall i \in \mathbb{N}^+\}$ with intensity λ_b. The set of all SBSs is denoted by $\mathcal{B} = \{1, 2, \cdots, b, \cdots, B\}$. As shown in Fig. 1, geographically adjacent SBSs are grouped into associative clusters where collaboratively deliver contents requested and improve wireless transmission performance. The cache manager (CM) is connected to a data center via a high-speed dedicated link and some clusters are selected to connected CM for efficient operations. The number of connected clusters is determined by parameter ϵ, which is defined as the number of the proportion of the number of selected clusters to the number of all clusters. In order to facilitate the management, there is a cluster controller SBS connected to the other SBSs while the other SBSs are not directly connected. Some controller SBSs are directly connected with some are not. The result of clustering is that all SBSs are clustered into J clusters, and the set of cluster is denoted by $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$, where J is corresponded to the number of small cell. For convenience, the cache device in SBS has the same storage capacity, and the total cache capacity in each cluster is considered as an entity.

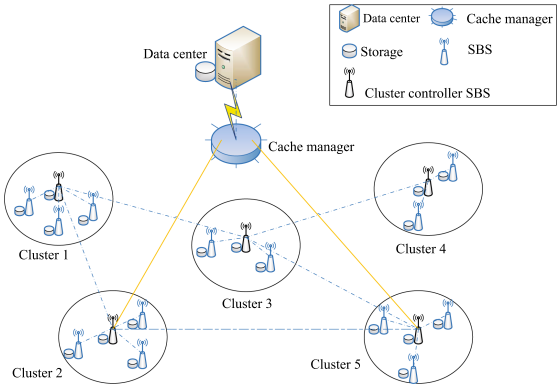

Fig. 1. A cache-enabled clustered SCN.

2.1 Small Cell Clustering

Based on the SCN model as Fig. 1, we consider using a hexagonal grid with inter-cluster center distance $2R_h$, and the area of each cluster is $\mathcal{A} = 2\sqrt{3}R_h^2$ to represent the shape of cell. For a random cluster, the probability mass function that the number of SBSs n inside cluster equal to K follows a Poisson distribution with mean $\lambda_b \mathcal{A}$ is denoted by

$$\mathbb{P}(n = K) = e^{-2\sqrt{3}\lambda_b R_h^2} \frac{(2\sqrt{3}\lambda_b R_h^2)^K}{K!}, \tag{1}$$

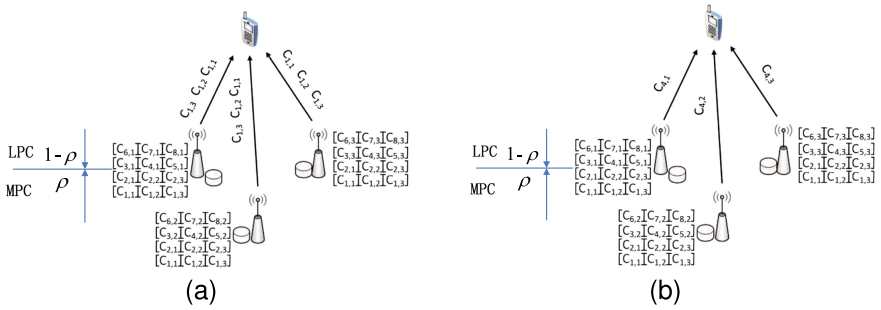

Fig. 2. (a) JT transmission procedure. (b) PT transmission procedure.

there are K SBSs in a certain cluster, the SBS distribution follows a Binomial Point Process. The distance distribution between randomly distributed SBSs and the distance from a SBS to an arbitrary in hexagonal cell have been elaborated in [15], where demonstrated that using a circle with the same area to provide the best approximation of the distribution model, and the radius is $R = R_h\sqrt{\frac{2\sqrt{3}}{\pi}}$. The set of SBSs in j-th cluster is $\mathcal{C}_j = \{b_i \in \Phi_b \bigcap \mathcal{M}(z_j, R)\}$, where $\mathcal{M}(z_j, R)$ denotes the ball centered at z_j with radius R. The j-th cluster with K SBSs is denoted as $\mathcal{C}_j = \{C_{j,1}, \cdots, C_{j,k}, \cdots, C_{j,K}\}$, due to the limitation of cache capacity, each SBS can store up to D contents, then the total available storage capacity of $\mathcal{C}_j$ is u_j, as $KD = u_j$.

The network is operated in time-slotted manner and the time slot is denoted by $\mathcal{T} = \{0, 1, \cdots, t, \cdots, T\}$. The total contents in data center in time slot t are indexed by $\mathcal{F}_t = \{1, 2, \cdots, f, \cdots, F_t\}$, which varies over time for new contents uploaded. In the cluster with K SBSs, each content is divided into K equal-size partitions [16]. At the beginning of each time slot, the CM makes a decision on cooperatively refreshing the cache entities to cache new popular contents. In each cluster, each user makes independent request for contents in each time slot. To be specific, we consider MovieLens, a web-based recommender system based on the rating of movie viewed by users [17] as the dataset. In each time slot, the request statue among SBSs is different.

2.2 Cooperative Transmission

In this work, We assume single antenna at both SBSs and user device. Orthogonal multiple access method is used to tackle simultaneous arrival of content requests. In the SCN, we assume that each content contains S bits, the successful delivery of a content is defined by the event that S bits are successfully delivered using bandwidth W and time T_ϕ. With CoMP, it provides two different transmission modes, namely joint transmission(JT) and parallel transmission(PT), as described in Fig. 2.

Joint Transmission. If the requested content f is in the MPC range, each SBSs in the cluster store entire content. Thus, f is jointly transmitted by K SBSs to user. The purpose is to increase the received signal to interference plus noise ratio (SINR) to enhance the content delivery reliability. It is denoted as JT cooperation scheme with each SBS of a cluster sends S bits to the user using the same bandwidth. The received signals from K SBSs are superimposed and can be regarded as a single stream. The successful content delivery probability (SCDP) [13] is defined as

$$P_K^{\mathrm{JT}} = \mathbb{P}\left[WT_\phi \log_2(1+\mathrm{SINR}) > S \mid K\right]. \tag{2}$$

$$P_K^{\mathrm{JT}} = \mathbb{P}\left[\mathrm{SINR} > 2^{\frac{R_d}{W}} - 1 \mid K\right], \tag{3}$$

where $R_d = \frac{S}{T_\phi}$ as the expected delivery rate for successful content delivery.

Parallel Transmission. If the requested content f is in the LPC range, each K cooperative SBSs in the cluster stores disjoint partitions. Thus, the different partitions need to be transmitted to the user simultaneously by K streams. It is denoted as PT cooperation scheme. We adopt PT with successive decoding based spectrum sharing case with K SBSs simultaneously send $\frac{S}{K}$ bits to the user by sharing the same W bandwidth. The successive decoding with interference cancellation (SIC) is utilized to decode the signal according to the received signal power order [18]. The SCDP is defined as

$$p_K^{\mathrm{PT}} = \mathbb{P}\left[\bigcap_{i\in\mathcal{C}_j} WT_\phi \log_2(1+\mathrm{SINR}_i) > \frac{S}{K} \mid K\right], \tag{4}$$

$$p_K^{\mathrm{PT}} = \mathbb{P}\left[\bigcap_{i\in\mathcal{C}_j} \mathrm{SINR}_i > 2^{\frac{R_d}{KW}} - 1 \mid K\right]. \tag{5}$$

where SINR_i is the SINR from the i-th SBS in cluster $\mathcal{C}_j$ of requested content.

Transmission for Sharing Case and Missing Case. In addition to the above two transmission modes, we also consider transmission for sharing and transmission for missing case. When connecting user in $\mathcal{C}_j$ requests content f which cached in $\mathcal{C}_i$, the cluster controller SBS of $\mathcal{C}_i$ retrieve the content f according to the caching mode, and then the cluster controller SBS of $\mathcal{C}_j$ fetches f from the cluster controller SBS of $\mathcal{C}_i$, and shares the decomposed contents to the remaining SBSs within $\mathcal{C}_j$ with PT mode. Meanwhile, if connecting user in $\mathcal{C}_j$ requests content f which is not cached in the local clusters. In this missing case, the cluster controller SBS in $\mathcal{C}_j$ fetches content f from the data center through backhaul links and transmits content f to user with PT mode.

3 CWFM-Based Content Popularity Prediction

Before we design the caching strategy, we need to determine what contents will be cached, which is to predict the popularity. With the edge network has the ability to control, compute and cache, [19] proposed to cache the content purposefully according to the user preferences, the simulation results indicate that a small number of features can also show obvious effects in improving the cache hits. Further, we expand the thought and φ features are extracted to construct a connection between contents and user's favor. The extracted features meet low redundancy between features, and strong correlation between the features and user. The data set MovieLens [17] is utilized and the user's movie scoring process in a timestamp is equivalent to the user's content request process. We use mutual information to measure the correlation and the redundancy, and the selected feature set is defined as $\mathbf{F}^*$.

For the content popularity in different time slot, φ features is constructed into a matrix with M rows and N columns and matrix $\boldsymbol{A}_f$ is used to represent the attribute of content f, where element A_f^{mn} equels 1 if content f has the corresponding feature and 0 otherwise [20]. As the importance of each feature is non-uniform, we determine a weight to each feature and construct a weighted feature matrix (WFM) in terms of spatiotemporal variation. The WFM in $\mathcal{C}_j$ is defined as $\boldsymbol{P}_j(t)$ in time slot t. At the initial time slot, the feature weight is given by the accumulation of the number of historical requests, shown as $p_j^{mn}(0) \triangleq \sum_{f\in\mathcal{F}_0} Q_f^j(\tau)A_f^{mn}$, where $Q_f^j(\tau)$ and $\mathcal{F}_0$ are the request number of content f in $\mathcal{C}_j$ after the training time τ and the set of initial contents respectively. Therefore, the initial WFM can be written as

$$\mathbf{P}_j(0) \triangleq \sum_{f\in\mathcal{F}_0} Q_f^j(\tau)\cdot\boldsymbol{A}_f. \tag{6}$$

The popularity of content f in $\mathcal{C}_j$ in time slot t is given as

$$g_{j,f}(t) = (\mathbf{1}^N)^T(\boldsymbol{P}_j(t)\otimes\boldsymbol{A}_f\mathbf{1}^N), f\in\mathcal{F}_t, \tag{7}$$

where $\otimes$ represents the Hadamard multiplication and $\mathbf{1}^N$ is the all one column vector. Thus, it is possible to establish an internal relation between contents in terms of popularity, also a relation between content popularity and user preferences.

It is reasonable to combine the historical feature weight and the content request number in the previous time slot to balance the feature weight in the next time slot. Hence, we learn user's status online and dynamically adjusting weight of feature. If one content is requested, the weight of corresponding feature increases. To measure the increment, we introduce an growth factor $\eta_f^j(t) = \sigma^{d_{j,f}(t)/\sum_f d_{j,f}(t)}$, where $d_{j,f}(t)$ is the requested number of content f in $\mathcal{C}_j$ in time slot t and σ is determined by the specific request status in SCNs, which is no less than 1. It is observed that larger σ makes the prediction of user's favor more inclined to the features in the previous time slot. Different contents

in $\mathcal{F}_t$ may have the same features. For simplicity and effectiveness, we adopt statistical method of superimposing the feature weights and averaging them. According to the known $\mathbf{P}_j(t)$ and the request information of $\mathcal{C}_j$ in time slot t, the superimposed CWFM of $\mathcal{C}_j$ at beginning of time slot $t+1$ is

$$\overline{\mathbf{P}_j(t+1)} = \sum_{f \in \mathcal{F}_t} \eta_f^j(t)\, \boldsymbol{A}_f \otimes \mathbf{P}_j(t). \tag{8}$$

Therefore, the feature matrix of $\mathcal{C}_j$ at the beginning of time slot $t+1$ can be written as

$$\boldsymbol{P}_j(t+1) = \overline{\mathbf{P}_j(t+1)} \oslash \sum_{f \in \mathcal{F}_t} \boldsymbol{A}_f, \tag{9}$$

where $\oslash$ represents the Hadamard division. The feature weight of each content can be refreshed with the passage of time slot , and the prediction result of the popularity of content f in $\mathcal{C}_j$ within time slot $t+1$ is

$$g_{j,f}(t+1) = (\mathbf{1}^N)^T(\boldsymbol{P}_j(t+1) \otimes \boldsymbol{A}_f \mathbf{1}^N), f \in \mathcal{F}_{t+1}. \tag{10}$$

As a result, by refreshing and iterations, CWFM can achieve the accurate prediction of content popularity, which contains the newly uploaded contents in each time slot.

4 Proposed Energy Efficiency-Based Caching

In clustered SCNs, EE is the most important metric since CoMP can significantly improves SINR and reduces delay, while multiple SBSs cause higher energy consumption. In this paper, we define EE as the ratio of actual delivery rate to energy consumption.

For the requested content f, the delivery policy in time slot t can be expressed as $\mathbf{y}(t) = \{y_{j,f,1}(t), y_{j,f,2}(t), y_{i,j,f}(t), y_{c,j,f}(t), \forall i, j \in \mathcal{J}, i \neq j, \forall f \in \mathcal{F}_t\}$, where $y_{j,f,1}(t)$ represents if content f is fully cached in $\mathcal{C}_j$ and $y_{j,f,2}(t)$ represents if content f is partly cached in $\mathcal{C}_j$. $y_{i,j,f}(t)$ indicates whether $\mathcal{C}_j$ fetches the content f from $\mathcal{C}_i$, $y_{c,j,f}(t)$ indicates whether $\mathcal{C}_j$ fetches the content f from data center. They all take values from $\{0, 1\}$. Note that content f is fetched from $\mathcal{C}_i$ only when $\mathcal{C}_i$ caches it, that is

$$y_{i,j,f}(t) \leqslant y_{i,f,1}(t) + y_{i,f,2}(t). \tag{11}$$

Only one case can be used to deliver the requested content f in each time slot, thus

$$y_{j,f,1}(t) + y_{j,f,2}(t) + y_{i,j,f}(t) + y_{c,j,f}(t) \leqslant 1. \tag{12}$$

As the limitation of cache capacity, it needs to be satisfied that

$$\sum_{f \in \mathcal{F}_t} K y_{j,f,1}(t) + y_{j,f,2}(t) \leqslant u_j. \tag{13}$$

Hence, in time slot t, the EE can be formulated as

$$EE(t) = \sum_{f\in\mathcal{F}_t}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{J}\setminus j} d_{j,f}(t)\frac{[V_1 y_{j,f,1}(t) + V_2 y_{j,f,2}(t) + V_3 y_{i,j,f}(t) + V_4 y_{c,j,f}(t)]}{[P_1 y_{j,f,1}(t) + P_2 y_{j,f,2}(t) + P_3 y_{i,j,f}(t) + P_4 y_{c,j,f}(t)]}, \tag{14}$$

where V_1, V_2, V_3 and V_4 respectively denote the actual deliver rate of the above four deliver categories and P_1, P_2, P_3, P_4 respectively denote the energy consumption. In clustered SCNs, the energy efficiency-optimized caching strategy is actually maximize the CoMP gain and the content cache gain [21].

4.1 Successful Content Delivery Probability Analysis

As EE is the ratio of actual delivery rate to energy consumption. The actual delivery rate is determined by the actual transmission rate (ATR) and SCDP. Hence, the SCDP in different cases is discussed in this section. The SCDP in JT mode and PT mode have been shown as (3) and (5).

For a typical cluster-center user locates at z_j and requests content with JT, the cooperating SBSs in $\mathcal{C}_j$ transmit the same symbol s with the equal transmit power P_e. Considering the channel loss due to distance, we use a standard distance-dependent power law pathloss attenuation as $d^{-\kappa}$, d is the distance and κ is the pathloss exponent. The channel output at the user is

$$Z = \sum_{b\in\mathcal{C}_j}\sqrt{P_e}d_b^{-\frac{\kappa}{2}}h_b s + \sum_{l\in\mathcal{B}\setminus\mathcal{C}_j}\sqrt{P_e}d_l^{-\frac{\kappa}{2}}h_l s_l + N_o, \tag{15}$$

where h_b and h_l represent the small-scale Rayleigh fading from the b-th and the l-th SBS to the user respectively, which follows $h_b, h_l \backsim \mathcal{CN}(0,1)$, d_b, d_l respectively represent the distance from the b-th and the l-th SBS. s_l represents the transmitted symbol of the l-th SBS out of the cluster, s is the joint transmitted symbol in $\mathcal{C}_j$, N_o donates the background thermal noise. Considering the interference-limited network and neglecting N_o, the signal to interference ratio (SIR) is

$$\mathrm{SIR}_{JT} = \frac{|\sum_{b\in\mathcal{C}_j} h_b d_b^{-\frac{\kappa}{2}}|^2}{\sum_{l\in\mathcal{B}\setminus\mathcal{C}_j}|h_l|^2 d_l^{-\kappa}}. \tag{16}$$

The expected SIR is $\theta_1 = 2^{\frac{R_d}{W}} - 1$, and the SCDP in JT case can be expressed as

$$p_K^{JT}(\theta_1) \simeq \int_0^R \cdots \int_0^R \mathcal{L}_{I\setminus R}\left(\frac{\theta_1}{\sum_{k=1}^K x_k^{-\kappa}}\right)\prod_{k=1}^K \frac{2x_k}{R^2}dx_1\cdots dx_K, \tag{17}$$

where $\mathcal{L}_{I\setminus x}$ is the Laplace transform of the interference coming from SBSs located outside of $\mathcal{M}(0,x)$ and has been proven in [13] as

$$\mathcal{L}_{I\setminus x} = \exp\left(-\pi\lambda s^{2/\kappa}\int_{\frac{x^2}{s^{2/\kappa}}}^{\infty}\frac{1}{1+\omega^{2/\kappa}}\right). \tag{18}$$

When user locates at z_j and in the case of PT, the cooperating K SBSs in $\mathcal{C}_j$ transmit different parts of symbol $[s_1, \cdots, s_k, \cdots, s_K]$ with equal transmit power P_e to the user. The channel output at the user is

$$Z = \sum_{b \in \mathcal{C}_j} \sqrt{P_e} d_b^{-\frac{\kappa}{2}} h_b s_b + \sum_{l \in \mathcal{B} \setminus \mathcal{C}_j} \sqrt{P_e} d_l^{-\frac{\kappa}{2}} h_l s_l + N_o. \tag{19}$$

In order to decode multiple streams simultaneously, we adopt received power ordering to decode different data stream by distance value studied in [22]. The distance vector can be expressed as $\mathbf{d} = [d_1^*, \cdots, d_k^*, \cdots, d_K^*]$, where d_k^* denotes the distance from z_j to the k-th nearest SBS. When decoding the information from the k-th SBS, all signals come from closer $k-1$ SBSs should have been successfully decoded and canceled. Hence, the SIR of the k-th stream is given as

$$\mathrm{SIR}_k \simeq \frac{|h_k|^2 (d_k^*)^{-\kappa}}{\sum_{l \in \mathcal{B} \setminus \mathcal{M}(0, d_k^*)} |h_l|^2 d_l^{-\kappa}}. \tag{20}$$

Finally, the remaining interference only comes from out-cluster SBSs and the distance is greater than R, thus the SIR of the final decoded stream is

$$\mathrm{SIR}_k \simeq \frac{|h_K|^2 (d_K^*)^{-\kappa}}{\sum_{l \in \mathcal{B} \setminus \mathcal{M}(0, d_K^*)} |h_l|^2 d_l^{-\kappa}}. \tag{21}$$

The expected SIR is $\theta_2 = 2^{\frac{R_d}{KW}} - 1$, and the SCDP in PT case can be expressed as

$$p_K^{PT}(\theta_2) \simeq \int_{0 < x_k < R} \frac{2Kx_K}{R^2} \mathcal{L}_{I \setminus R}(\theta_2 x_K^\kappa) \prod_{k=1}^{K-1} \frac{2kx_k}{R^2} \mathcal{L}_{I \setminus x_k}(\theta_2 x_k^\kappa) dx_1 \cdots dx_K. \tag{22}$$

When user locates at z_j and transmits for sharing case, the cluster controller in $\mathcal{C}_j$ fetches f from $\mathcal{C}_i$, $i \in \mathcal{J} \setminus j$, and then transmits with PT mode. Since the inter-cluster data transmission consumes time resources, the maximum transmission time of the wireless side is reduced. We define the delay of fetching requested content f from $\mathcal{C}_i$ as $T_{i,j,f}$ and the wireless side maximum transmission time is changed as $\sigma_1 T_\phi$, where $\sigma_1 = 1 - \frac{T_{i,j,f}}{T_\phi}$. Thus, the expected SIR is $\theta_3 = 2^{\frac{R_d}{\sigma_1 W}} - 1$ and the SCDP is $p_K^{PT}(\theta_3)$.

When user locates at z_j and transmits for missing case, the cluster controller in $\mathcal{C}_j$ fetches f from data center and then transmits with PT mode. Since the backhaul delay is relatively large, the maximum transmission time on the wireless side is significantly reduced. We define the delay of fetching requested content f from data center as $T_{c,j,f}$ and the wireless side maximum transmission time is changed as $\sigma_2 T_\phi$, where $\sigma_2 = 1 - \frac{T_{c,j,f}}{T_\phi}$. Hence, the expected SIR is $\theta_4 = 2^{\frac{R_d}{\sigma_2 W}} - 1$ and the SCDP is $p_K^{PT}(\theta_4)$.

The actual delivery rate can be defined as the product of SCDP and ATR. For the expected delivery rate R_d, the ATR has a partial probability above

R_d, and a partial probability below as the channel is unstable. Since the total transmission content size remains unchanged, as long as the average transmission rate is greater than R_d, the ATR can be combined by different transmission rate. In addition, we analyze all possible combinations of transmission rates, and the resulting average transmission rate is called the threshold of JT mode or PT mode. If there is a probability that the ATR is greater than R_d, then there must be a minimum transmission rate v^* less than R_d in the different transmission rates combination. When R_d is less than the threshold, we have v^*=0, and when R_d is greater than the threshold, in order to meet the condition that the actual delivery rate is greater than or equals to the R_d, v^*>0 is necessary, thus the combination of transmission rates interval is $[v^*, \infty]$. The actual delivery rate can be expressed as

$$\sum_{v=v^*}^{\infty} \Delta v \Delta p_K (2^{\frac{v}{KW}} - 1). \tag{23}$$

Wherein, when v^*=0, the result of (23) is the threshold value which is greater than or equals to $R_d p_K^{JT}(2^{\frac{v^*}{KW}} - 1)$, the actual delivery rate can be given as

$$\left\{ R_d p_K (2^{\frac{v^*}{KW}} - 1) | R_d p_K (2^{\frac{v^*}{KW}} - 1) \leq \sum_{v=v^*}^{\infty} \Delta v \Delta p_K (2^{\frac{v}{Kw}} - 1) \right\}. \tag{24}$$

Hence, the effective delivery rates of above cooperative transmission cases are $R_d p_K^{JT}(\theta_1^*)$, $R_d p_K^{PT}(\theta_2^*)$, $R_d p_K^{PT}(\theta_3^*)$, $R_d p_K^{PT}(\theta_4^*)$, respectively, where $\theta_1^* = 2^{\frac{v_1}{W}}$-1, $\theta_2^* = 2^{\frac{v_2}{KW}}$-1, $\theta_3^* = 2^{\frac{v_3}{\sigma_1 W}}$-1 and $\theta_4^* = 2^{\frac{v_4}{\sigma_2 W}}$-1, v_1, v_2, v_3 and v_4 meet the demand of (24).

4.2 EE-Based Proactive Caching Cooperatively

Energy consumption mainly consists of the wireless side emission energy consumption and the wired side fiber transmission energy consumption. For JT case, it contains only emission energy consumption which is given as KP_e while transmitting unit size content. For PT case, it also contains only emission energy consumption. However, this case has lower transmission rate and the duration of each SBS is similar to the JT case. Thus, the overall energy consumption is also KP_e. The energy consumption of the wired side fiber transmission can be decomposed into the unit size content transmission energy consumption $P_{i,j}$ of the cluster controllers between $\mathcal{C}_i$ and $\mathcal{C}_j$ and the unit size decomposition part transmission energy consumption P_{km} of the cluster controller to other SBSs. In sharing case, when content f is cached as MPC in $\mathcal{C}_i$, the energy consumption is $(K-1)P_{km}+P_{i,j}+KP_e$, when it is cached as LPC in $\mathcal{C}_i$, the energy consumption is $2(K-1)P_{km}+P_{i,j}+KP_e$. To distinguish the two types, we use $y_{i,j,f,1}(t)$ and $y_{i,j,f,2}(t)$ respectively to indicate whether it occurs in time slot t and we have

$$y_{i,j,f,1}(t) + y_{i,j,f,2}(t) = y_{i,j,f}(t). \tag{25}$$

Similarly, in transmission for missing case, the energy consumption is given as $(K-1)P_{km} + P_{c,j} + KP_e$, where $P_{c,j}$ denotes the energy consumption which data center transmits unit size content to the cluster controller of $\mathcal{C}_j$.

In missing case, the backhaul energy consumption is significantly larger than the other three cases, and the effective delivery rate of the backhaul is significantly smaller. If the local clusters cache the requested content, it should be likely fetched from local clusters rather than the data center. Hence, when the content caching policy is determined, the transmission policy is also determined accordingly, and the total effective delivery rate in the time slot t is formulated as

$$\begin{aligned} V_{tot}(t) = &\sum_{f\in\mathcal{F}_t}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{J}\setminus j} d_{j,f}(t)\left[p_K^{JT}(\theta_1)R_d y_{j,f,1}(t) + p_K^{PT}(\theta_2)R_d y_{j,f,2}(t)\right] \\ &+ \left[p_K^{PT}(\theta_3)R_d y_{i,j,f,1}(t) + p_K^{PT}(\theta_3)R_d y_{i,j,f,2}(t) + p_K^{PT}(\theta_4)R_d y_{c,j,f}(t)\right]. \end{aligned} \tag{26}$$

The total energy consumption is formulated as

$$\begin{aligned} P_{tot}(t) = &\sum_{f\in\mathcal{F}_t}\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{J}\setminus j} d_{j,f}(t)\left\{KP_e y_{j,f,1}(t) + KP_e y_{j,f,2}(t) + [(K-1)P_{km} + P_{c,j} + KP_e]y_{c,j,f}(t)\right\} \\ &+ \left\{[(K-1)P_{km} + P_{i,j} + KP_e]y_{i,j,f,1}(t) + [2(K-1)P_{km} + P_{i,j} + KP_e]y_{i,j,f,2}(t)\right\}. \end{aligned} \tag{27}$$

Thus, the maximum EE in time slot t is written as

$$\max_{\mathbf{y}(t)} \quad \frac{V_{tot}(t)}{P_{tot}(t)}. \tag{28}$$

Note that the optimization goal is determined by the 5-dimensional Boolean variable. The problem is actual where to place contents in time slot t, and its cache mode to maximum the EE. However, even if the variables are related, the conditional constraints can not simplify the EE formulation into a function form determined by a single variable.

To optimize the EE, if the requested content gets hit within the cluster, transmission time for wireless is a long period, the effective delivery rate is large, and the energy consumption is small. If the requested content gets hit in the neighboring cluster, the transmission time is reduced, the effective delivery rate becomes smaller, and the energy consumption becomes larger. If the requested content misses, the remaining transmission time is least, the effective delivery rate is smallest, and the energy consumption is largest. Hence, the cache hits become critical. Meanwhile, compared with the LPC, MPC cached fully, although the effective delivery rate increases, the cache space occupied is K times larger than the former. That means one of the MPC is cached, $K-1$ LPC should be discarded. The effective delivery rate for the MPC should be approximately K times than the LPC. Therefore, we should first ensure that the requested content hits as many local clusters as possible, and then gradually analyze and convert the content in LPC range into MPC. Comparing with the non-convex problem which directly solves the optimal value, we adopt a two-step stepwise optimization method. Firstly, all the contents are cached as LPC

and transferred by PT mode, which reduces the dimensionality of the variables and the complexity of the computation compared to a direct solution. After that, some of the LPC will be converted to MPC for caching using EE as an indicator. The whole algorithm processing process does not change the cache constraints, but only the variable processing process is solved in two steps, so this method is easy to implement and ensures reliability.

To maximize cache hits, the cache capacity of local cluster is a fixed value, and the optimization objective becomes the maximum effective delivery rate in the PT mode and minimizes the energy consumption. As analyzed above, both in PT mode completely depend on the consumption of fiber transmission and proportional to the time consumption and the energy consumption. Therefore, the optimization objective turns to minimize fiber transmission consumption within time slot t. In this work, we define the transmission consumption is the square root of the product of time consumption and energy consumption. The content caching policy at time slot t is denoted as $\mathbf{y}_r(t) = \{y_{i,j,f,2}(t), \forall i, j \in \mathcal{J}, \forall f \in \mathcal{F}_t\}$. If user associated with $\mathcal{C}_j$ fetches content f from $\mathcal{C}_i$, the consumption is $\sum_{i,j\in\mathcal{J}} \sum_{f\in\mathcal{F}_t} e_{i,j} d_{j,f}(t) y_{i,j,f,2}(t) s_f$, where $e_{i,j}$ represents the consumption that $\mathcal{C}_i$ delivers a unit size content to $\mathcal{C}_j$ and s_f is the size of content f. Note that if $i = j$ represents that user fetches content f from the cluster associated with currently. In addition, if user associated with $\mathcal{C}_j$ fetches content f from data center, the consumption is $\sum_{j\in\mathcal{J}} \sum_{f\in\mathcal{F}_t} e_{c,j} d_{j,f}(t) y_{c,j,f}(t) s_f$, where $e_{c,j}$ represents the consumption that data center delivers a unit size content to $\mathcal{C}_j$. $d_{j,f}(t)$ is the number requests for content f, which is inextricably related with the popularity and can be replaced by $\alpha g_{j,f}(t)$. Thus, our objective is minimizing the consumption as

$$\min_{\mathbf{y}_r(t)} \sum_{i,j\in\mathcal{J}} \sum_{f\in\mathcal{F}_t} \alpha g_{j,f}(t)(e_{i,j} y_{i,j,f,2}(t) + e_{c,j} y_{c,j,f}(t)) s_f. \tag{29}$$

Based on $\sum_{j\in\mathcal{J}} y_{i,j,f,2}(t) + y_{c,j,f}(t) = 1$, we can eliminate $y_{c,j,f}(t)$, the formulation turns to be

$$\max_{\mathbf{y}_r(t)} \quad \sum_{f\in\mathcal{F}_t} \sum_{j\in\mathcal{J}} g_{j,f}(t) s_f \sum_{i\in\mathcal{J}} y_{i,j,f,2}(t)(e_{c,j} - e_{i,j}). \tag{30}$$

As $\mathcal{C}_j$ can retrieve content f from $\mathcal{C}_i$ only when $\mathcal{C}_i$ has cached it before. That is the problem to maximize the objective is actual where to place f, the specific cluster or data center. Hence, $y_{i,j,f,2}(t)$ can be eliminated, the placement policy is given as

$$\max_{\mathbf{y}_r(t)} \quad [y_{1,f,2}(t), \cdots, y_{\mathcal{J},f,2}(t)] \cdot \mathbf{E}_f \cdot [g_{1,f}(t), \cdots, g_{\mathcal{J},f}(t)]. \tag{31}$$

where $\mathbf{E}_f$ is a matrix with a dimension of $J \times J$, and the objective can be written as

$$\max_{\mathbf{y}_r(t)} \quad \mathbf{Y}(t) \cdot \mathbf{E} \cdot \mathbf{G}(t), \tag{32}$$

$\mathbf{Y}(t) = [y_{1,1,2}(t), \cdots, y_{J,1,2}(t); \cdots, y_{j,f,2}(t), \cdots, y_{J,F_t,2}(t)]^T$, $\mathbf{E}$ is a diagonal matrix as $\mathbf{E} = [\mathbf{E}_1, \cdots, \mathbf{E}_f, \cdots, \mathbf{E}_{F_t}]$, $\mathbf{G}(t) = [g_{1,1}(t), \cdots, g_{J,1}(t); \cdots, g_{j,f}(t), \cdots, g_{J,F_t}(t)]$. Thanks to constraints on Boolean variables $y_{j,f,2}(t)$, content size and link consumption, the optimization objective is non-convex. To make it convex, we relax Boolean variable $y_{j,f,2}(t) \in \{0, 1\}$ into [0,1] and tackle the problem with linear programming and variable recovery. Therefore, a optimal solution can be achieved within a certain precision range, and then we can obtain a solution denoted as $\mathbf{Y}^*(t)$. Note that $y^*_{j,f,2}(t)$ in $\mathbf{Y}^*(t)$ represents the contribution degree of content f to $\mathcal{C}_j$. Since the dimension of $\mathbf{Y}^*(t)$ is 1, a quick sort algorithm is adopted to map $\mathbf{Y}^*(t)$ into $\mathbf{Y}_v(t) = [y_1(t), \cdots, y_v(t), \cdots, y_{J \cdot F_t}(t)]$ by size of variable value, where $y^*_{j,f,2}(t)$ mapped into $y_v(t)$ means that $y^*_{j,f,2}(t)$ is the v-th largest item in $\mathbf{Y}^*(t)$. The recovery is performed step by step and the recovery policy for v-th variable is

$$y_{j,f,2}(t) = \begin{cases} 1, \ s_f \leq u_j - \hat{u}_j, \ y_{j,f,2}(t) \neq 1, \ \forall j \in \mathcal{J}, \\ 0, \ \text{otherwise} \end{cases} \tag{33}$$

where $\hat{u}_b$ is the occupied caching space of $\mathcal{C}_j$. Finally, maximum objective value is obtained when CM completes the recovery process, and then substituting the value into (28), we can get the $EE_{ini}(t)$.

4.3 Optimal Design for Maximizing EE

The caching policy can be further improved by discarding some of the LPC and converting the more popular ones into MPC, and using the EE gain of the JT mode to improve the overall EE of the system. The contents cached in $\mathcal{C}_j$ are sorted in descending order of contribution as $\left[1^*_1, \cdots, F^*_j; \cdots, 1^*_J, \cdots, F^*_J\right]$. For $\mathcal{C}_j$, f^*_j is converted into the MPC and the increase of effective delivery rate is $d_{j,f^*_j}(t)\left[R_d p_K^{JT}(\theta_1) - R_d p_K^{PT}(\theta_2)\right]$, while energy consumption remains constant. For $\mathcal{C}_i$, the decrease of energy consumption is $\sum_{i \in \mathcal{J} \setminus j} d_{j,f^*_j}(t)(K-1)P_{km}$, while effective delivery rate remains constant. In addition, converting f^*_j to the MPC means that $K-1$ LPC are discarded. For this part of contents, in $\mathcal{C}_j$, the decrease of effective delivery rate is

$$\sum_{l^*_j = F^*_j - K + 2}^{F^*_j} d_{j,l^*_j}(t)\left[R_d p_K^{PT}(\theta_4) - R_d p_K^{PT}(\theta_2)\right], \tag{34}$$

where l^*_j is the discarded less popular content. The increase of energy consumption is

$$\sum_{l^*_j = F^*_j - K + 2}^{F^*_j} d_{j,l^*_j}(t)\left[P_{c,j} + (K-1)P_{km}\right]. \tag{35}$$

In $\mathcal{C}_i$, the decrease of effective delivery rate is

$$\sum_{i\in\mathcal{J}\setminus j}\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{i,l_j^*}(t)\left[R_d p_K^{PT}(\theta_4)-R_d p_K^{PT}(\theta_3)\right], \tag{36}$$

while the increase of energy consumption is

$$\sum_{i\in\mathcal{J}\setminus j}\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{i,l_j^*}(t)\left[P_{c,j}-P_{j,i}-(K-1)P_{km}\right]. \tag{37}$$

Hence, as content f_j^* is converted to the MPC, the change of effective delivery rate is

$$\begin{aligned} V_{f_j^*}^1(t) = {} & d_{j,f_j^*}(t)\left[R_d p_K^{JT}(\theta_1)-R_d p_K^{PT}(\theta_2)\right]-\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{j,l_j^*}(t)\left[R_d p_K^{PT}(\theta_4)-R_d p_K^{PT}(\theta_2)\right] \\ & -\sum_{i\in\mathcal{J}\setminus j}\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{i,l_j^*}(t)\left[R_d p_K^{PT}(\theta_4)-R_d p_K^{PT}(\theta_3)\right]. \end{aligned} \tag{38}$$

The change of energy consumption is

$$\begin{aligned} P_{f_j^*}^1(t) = {} & \sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{j,l_j^*}(t)\left[P_{c,j}+(K-1)P_{km}\right]-\sum_{i\in\mathcal{J}\setminus j} d_{i,f_j^*}(t)(K-1)P_{km} \\ & +\sum_{i\in\mathcal{J}\setminus j}\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{i,l_j^*}(t)\left[P_{c,j}-P_{j,i}-(K-1)P_{km}\right]. \end{aligned} \tag{39}$$

Hence, the EE can be written as

$$EE_{f_j^*}^1(t)=\frac{V_{f_j^*}^1(t)}{P_{f_j^*}^1(t)}. \tag{40}$$

If content f_j^* is not converted to the MPC and $K-1$ contents as LPC are not discarded, the effective delivery rate is given as

$$\begin{aligned} V_{f_j^*}^2(t) = {} & d_{j,f_j^*}(t)R_d p_K^{PT}(\theta_2)+\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{j,l_j^*}(t)R_d p_K^{PT}(\theta_2) \\ & +\sum_{i\in\mathcal{J}\setminus j} d_{i,f_j^*}(t)R_d p_K^{PT}(\theta_3)+\sum_{i\in\mathcal{J}\setminus j}\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{i,l_j^*}(t)R_d p_K^{PT}(\theta_3). \end{aligned} \tag{41}$$

The energy consumption is

$$P^2_{f_j^*}(t) = d_{j,f_j^*}(t)KP_e + \sum_{i\in\mathcal{J}\setminus j} d_{i,f_j^*}(t)\left[KP_e + P_{j,i} + 2(K-1)P_{km}\right] + \sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{j,l_j^*}(t)KP_e + \sum_{i\in\mathcal{J}\setminus j}\sum_{l_j^*=F_j^*-K+2}^{F_j^*} d_{i,l_j^*}(t)\left[KP_e + P_{j,i} + 2(K-1)P_{km}\right]. \tag{42}$$

The EE can be written as

$$EE^2_{f_j^*}(t) = \frac{V^2_{f_j^*}(t)}{P^2_{f_j^*}(t)}. \tag{43}$$

Whether $\mathcal{C}_j$ converts the content f_j^* to the MPC depends on $EE^1_{f_j^*}(t)$ and $EE^2_{f_j^*}(t)$. If $EE^1_{f_j^*}(t) \geq EE^2_{f_j^*}(t)$, content f_j^* is converted to the MPC and $K-1$ contents with lowest contribution are discarded, and $EE_{ini}(t)$ is refreshed. If $EE^1_{f_j^*}(t) < EE^2_{f_j^*}(t)$, content f_j^* and $EE_{ini}(t)$ remain unchanged. Therefore, at most steps of $J \cdot F$, the entire process of contents conversion can be realized and the maximum value of EE is $EE_{ini}(t)$, and the CM guides the SBSs cooperation of all clusters to maximize the EE of the clustered SCNs.

5 Performance Evaluation

In this section, we compare the performance of proposed energy efficiency caching algorithm based on content-awareness (EECABC) with others: LRU-based caching algorithm with proportion of MPC is 0.2 [8], MPMAB-based caching algorithm [9] and MPC-based caching algorithm with proportion of MPC is 0.1 [13]. In the simulation, we use the MovieLens DataSet [17] which included a total of four full datasets of 100 k, 1 m, 10 m, and 20 m, and in terms of time span as well as data integrity considerations, this paper chooses the latest ml-20m dataset as the simulation data. This dataset has 1000209 ratings of 3952, and assume that the users only request their higher rating movies in requesting process. In addition, considering the integrity of the video information, under the condition of ensuring the reliability of the simulation, we divided the long term timestamp into 20 time slots and 2000 movies were selected as historical contents, and 200 movies were selected as the number of new contents uploaded to the data center in subsequent time slots. We assume that the cache space of each cluster is equal to 9000 M and the unit content size is equal to 300 M. Other parameters in the simulation are showed in Table 1 which refers to the parameter ranges of existing available devices and the parameter settings in [13] [20]. We mainly compare the cache efficiency and system EE, where cache efficiency is defined as the SCDP times the amount of cache hit data divided by the total amount of caching data, which is utilized as a measure of cache hits. The simulation runs 10 times to take the average.

Table 1. Simulation parameters

Parameter	Value
Number of clusters J	7
Proportion of the number of selected cluster controller SBSs (ϵ)	0.3
Density of SBSs (λ_b)	$10^{-4}/\mathrm{m}^2$
Radius of cell (R_h)	100 m
Path fading index (κ)	4
SBS transmit power P_e	1 W
Wired transmission consumption from data center to the selected cluster per content	10 W
Wired transmission consumption between cluster per link per content	[2 W, 2.5 W]
Transmission consumption between cluster controller and the other SBS	0.2 W
Available bandwidth	10 MHz
Transmission delay between interconnected cluster controller SBS per content	5 s
Transmission delay between data center and the connected cluster controller SBS per content	15 s
Average number of contents requested per time slots	20

In order to measure the impact of the number of SBSs on the transmission mode, as shown in Fig. 3, we first compare the relationship between the SCDP and ATR of contents in JT and PT mode when the number of SBSs K is 2, 3 and 4, respectively. In JT mode, the more SBSs in a cluster, the higher SCDP is, and in PT mode is on the contrary. This is because in the JT mode, more SBSs cooperation provide stronger received signal and therefore a higher SIR, the SCDP is increased. In PT mode, the SCDP is defined as the product of the success probabilities of multiple data streams. When the number of parallel streams increase, the reliability of each data stream needs to be ensure. The more SBSs provide a lower SIR, thus the SCDP performance decrease. Obviously, regardless of the value of K in JT or PT mode, as the ATR increases, the SCDP gradually decreases. That is because when the ATR increases, the computational requirements for each SBS and the channel quality requirements increase. In addition, the difference in SCDP between the PT mode and JT mode is quite large, especially when the ATR increase to a large value. In the following simulation, the number of SBSs in a cluster is set to be 3. The threshold of delivery rate in the JT mode is calculated to be 22 Mbps and in the JT mode is 7 Mbps, respectively.

Figure 4 shows the performance comparison among LRU-based, MPMAB-based, MPC-based and proposed EECABC proactive caching algorithm in terms

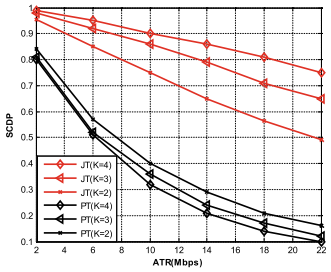

Fig. 3. Comparison among JT and PT in SCDP versus ATR.

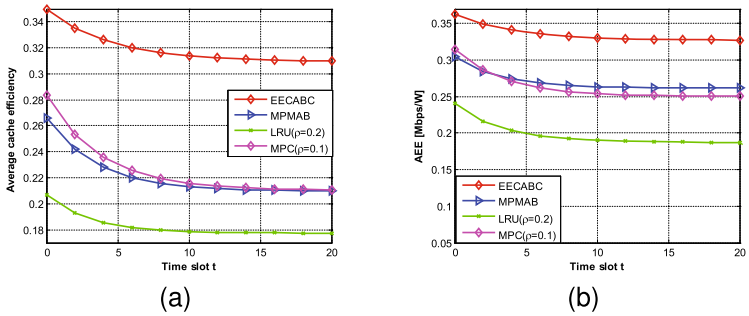

Fig. 4. (a) ACE comparison versus t ($R_d = 10$ Mbps). (b) AEE comparison versus t ($R_d = 10$ Mbps).

of average cache efficiency (ACE) and the average energy efficiency (AEE) as a function of time slot shift with the expected delivery rate R_d is 10 Mbps. As shown in Fig. 4(a), when the time slot changes, the ACE shows a downward trend, and the LRU, MPMAB and MPC cache algorithms decrease obviously. In contrast, the EECABC cache algorithm has a relatively stable rate of decline and eventually tends to be constant. With the passage of time, the new contents are uploaded to the data center gradually. The combination of the historical contents and the new contents improves the difficulty of popularity prediction. Hence, the cache hits is reduced, and the ACE is also reduced. This may prove that the popularity prediction bases on content-awareness has a relatively high prediction accuracy than others and can effectively predict the popularity of new contents. Meanwhile, as shown in Fig. 4(b), the AEE shows a downward trend as well with the time slot changes. The AEE of the EECABC cache algorithm outperforms than the other three algorithms. The combination of the historical contents and the new contents makes the accuracy of content popularity prediction reduce, which has a great impact on the placement of contents. The proposed EECABC cache algorithm achieves a higher prediction accuracy than the others. In addition, it may also prove that in the process of cache algorithm designing, the EECABC algorithm effectively improves the network performance by utilizing the JT mode of the CoMP technology simultaneously.

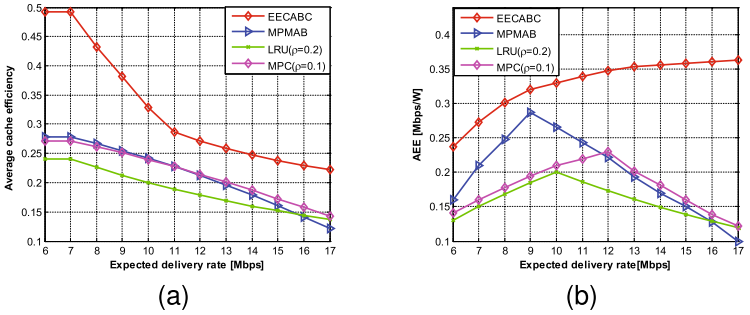

Fig. 5. (a) ACE comparison versus R_d. (b) AEE comparison versus R_d.

Figure 5 shows the performance comparison among LRU-based, MPMAB-based, MPC-based and proposed EECABC proactive caching algorithm in terms of ACE and the AEE as the expected delivery rate R_d is changed. When R_d is in the interval [6,7] which lower than the threshold of 22 Mbps in JT mode and 7 Mbps in PT mode, the SCDP is unchanged, the ACE is also unchanged. When R_d exceeds the threshold of the PT mode, such as in the interval [7,17], the SCDP of the PT mode is rapidly reduced, thus the ACE also decreases. As shown in Fig. 5(a), the ACE of the four cache algorithms show a downward trend when the R_d increases, the proposed EECABC caching algorithm performs better. Meanwhile, when the R_d exceeds 11 Mbps, the downward trend of ACE in proposed EECABC algorithm shows down. This is because our algorithm converts part of LPC into MPC, which can better utilize the advantage of JT mode under high value of R_d.

As AEE is the ratio of actual delivery rate to energy consumption. The actual delivery rate is the product of the SCDP and the expected delivery rate R_d. When R_d increases, the SCDP decreases, thus the AEE fluctuates as the actual delivery rate fluctuates. As shown in Fig. 5(b), the relationship between the AEE and R_d is shown as a downward parabola, and the extreme points are different in different caching algorithms. Comparing the four caching algorithm, our algorithm does not show a rapid decrease after the extreme point, but presents a slow increase and gradually tends to be stable. The reason may be that if the R_d increase gradually, the SCDP of the PT mode is rapidly reduced, due to the relatively high threshold value of JT mode, converting the contents into the MPC can better adapt to the high R_d requirement and we can take advantage of the JT mode. By adaptively adjusting the proportion of the MPC occupied in cache space, our algorithm achieves optimal in EE.

6 Conclusion

In this paper, we use a CWFM-based popularity prediction method with spatiotemporal variation, which can accurately predict the contents popularity, and

then we propose the energy-efficient design in the clustered SCNs, and decomposed the EE problem into two sub-problems. Firstly, we performed variable relaxation, LP and variable recovery under the condition of delivery consumption to maximize the cache efficiency of the local cluster. Then, the EECABC iteratively determine whether the contents with higher popularity was converted into MPC, and discarded the corresponding LPC, thereby maximizing the EE of the system within the global cluster. The simulation results showed that the proposed EECABC outperforms the existing strategies in terms of EE.

References

1. (2019). https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-738429.html
2. Li, Z., Liu, Y., Liu, A., Wang, S., Liu, H.: Minimizing convergecast time and energy consumption in green internet of things. IEEE Trans. Emerg. Topics Comput. **8**(3), 797–813 (2020)
3. Wu, Q., Li, G.Y., Chen, W., Ng, D.W.K., Schober, R.: An overview of sustainable green 5G networks. IEEE Wirel. Commun. **52**, 1897–1903 (2016)
4. Li, X., Wang, X., Leung, V.C.M.: Weighted network traffic offloading in cache-enabled heterogeneous networks. In: IEEE International Conference on Communications (ICC), pp. 1–6, May 2016
5. Piemontese, A., i Amat, A.G.: MDS-coded distributed caching for low delay wireless content delivery. IEEE Trans. Commun. **67**(2), 1600–1612 (2019)
6. Park, S.-H., Simeone, O., Shitz, S.S.: Joint optimization of cloud and edge processing for fog radio access networks. IEEE Trans. Wireless Commun. **15**(11), 7621–7632 (2016)
7. Nguyen, H.T., Tuan, H.D., Duong, T.Q., Poor, H.V., Hwang, W.: Collaborative multicast beamforming for content delivery by cache-enabled ultra dense networks. IEEE Trans. Commun. **67**(5), 3396–3406 (2019)
8. Blasco, P., Güdüz, D.: Learning-based optimization of cache content in a small cell base station. In: Proceedings of IEEE ICC, pp. 1897–1903, June 2014
9. Song, J., Sheng, M., Quek, T.Q.S., Xu, C., Wang, X.: Learning-based content caching and sharing for wireless networks. IEEE Trans. Commun. **65**(10), 4309–4324 (2017)
10. Tanzil, S.S., Hoiles, W., Krishnamurthy, V.: Adaptive scheme for caching YouTube content in a cellular network: a machine learning approach. IEEE Access **5**, 5870–5881 (2017)
11. Chiang, Y., Liao, W., Ji, Y.: RELISH: green multicell clustering in heterogeneous networks with shareable caching. In: IEEE Global Telecommunications Conference (GLOBECOM), pp. 1–7 (2018)
12. Nigam, G., Minero, P., Haenggi, M.: Coordinated multipoint joint transmission in heterogeneous networks. IEEE Trans. Commun. **62**(11), 4134–4146 (2014)
13. Chen, Z., Lee, J., Quek, T.Q., Kountouris, M.: Cooperative caching and transmission design in cluster-centric small cell networks. IEEE Trans. Wireless Commun. **16**(5), 3401–3415 (2017)
14. Zhang, S., He, P., Suto, K., Yang, P., Zhao, L., Shen, X.: Cooperative edge caching in user-centric clustered mobile networks. IEEE Trans. Mobile Comput. **17**(8), 1791–1805 (2018)

15. Zhuang, Y., Luo, Y., Cai, L., Pan, J.: A geometric probability model for capacity analysis and interference estimation in wireless mobile cellular systems. In: IEEE Global Telecommunications Conference (GLOBECOM), December 2011
16. Sourlas, V., Georgatsos, P., Flegkas, P., Tassiulas, L.: Partition-based caching in information-centric networks. In: IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 396–401 (2015)
17. Harper, F.M., Konstan, J.A.: The MovieLens datasets: history and context. IEEE Trans. Commun. **5**(4), 2160–6455 (2015)
18. Sen, S., Santhapuri, N., Choudhury, R.R., Nelakuditi, S.: Successive interference cancellation: carving out MAC layer opportunities. IEEE Trans. Mobile Comput. **12**(2), 346–357 (2013)
19. Müller, S., Atan, O., van der Schaar, M., Klein, A.: Context-aware proactive content caching with service differentiation in wireless networks. IEEE Trans. Wireless Commun. **16**(2), 1024–1036 (2017)
20. Teng, L., Yu, X., Tang, J., Liao, M.: Proactive caching strategy with content-aware weighted feature matrix learning in small cell network. IEEE Commun. Lett. **23**(4), 700–703 (2019)
21. Liu, K., Tao, M.: Exploiting tradeoff between transmission diversity and content diversity in multi-cell edge caching. In: IEEE International Conference on Communications (ICC), pp. 1–6, May 2018
22. Zhang, X., Haenggi, M.: The performance of successive interference cancellation in random wireless networks. IEEE Trans. Inf. Theory **60**(10), 6368–6388 (2014)