electronics

MDPI

Article

Analysis of an SDN-Based Cooperative Caching Network with Heterogeneous Contents

Qi Li [1], Xiaoxiang Wang [1,*], Dongyu Wang [1], Yibo Zhang [1], Yanwen Lan [1] and Qiang Liu [2] and Lei Song [3]

[1] The Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China; liqi1287345989@bupt.edu.cn (Q.L.); dy_wang@bupt.edu.cn (D.W.); yibo@bupt.edu.cn (Y.Z.); yanwen@bupt.edu.cn (Y.L.)

[2] Beijing University of Posts and Telecommunications, Beijing 100876, China; qiangliu@bupt.edu.cn

[3] Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China; sleilei_ren@163.com

* Correspondence: cpwang@bupt.edu.cn

Received: 15 October 2019; Accepted: 4 December 2019; Published: 6 December 2019

Abstract: The ubiquity of data-enabled mobile devices and wireless-enabled data applications has fostered the rapid development of wireless content caching, which is an efficient approach to mitigating cellular traffic pressure. Considering the content characteristics and real caching circumstances, a software-defined network (SDN)-based cooperative caching system is presented. First, we define a new file block library with heterogeneous content attributes [file popularity, mobile user (MU) preference, file size]. An SDN-based three-tier caching network is presented in which the base station supplies control coverage for the entire macrocell and cache helpers (CHs), MUs with cache capacities offer data coverage. Using the 'most popular content' and 'largest diversity content', a distributed cooperative caching strategy is proposed in which the caches of the MUs store the most popular contents of the file block library to mitigate the effect of MU mobility, and those of the CHs store the remaining contents in a probabilistic caching manner to enrich the content diversity and reduce the MU caching pressure. The request meet probability (RMPro) is subsequently proposed, and the optimal caching distribution of the contents in the probabilistic caching strategy is obtained via optimization. Finally, using the result of RMPro optimization, we also analyze the content retrieval delays that occur when a typical MU requests a file block or a whole file. Simulation results demonstrate that the proposed caching system can achieve quasi-optimal revenue performance compared with other contrasting schemes.

Keywords: wireless content caching; heterogeneous contents; cooperative caching strategy

1. Introductione

As the numbers of mobile users (MUs) have increased in recent years, mobile multimedia data transmitted by wireless networks have also increased exponentially. Cisco's most recent report estimates that global multimedia data are predicted to increase nearly eightfold between 2015 and 2020. Strong evidence exists that this growth will continue with a staggering annual rate of increase in the oncoming 5G era. Wireless multimedia data are primarily delivered through cellular networks (e.g., 3G, 4G and 5G). Recent studies have shown that fast-growing data traffic will soon become a significant burden on the cellular network [1].

Techniques, e.g., millimeter wave communication (using new spectral resources), massive multiple-input multiple-output (improving spectral efficiency), and network densification (improving spatial spectral efficiency and network coverage) methods, have been developed to relieve the traffic burden. Despite benefits from these techniques, the deployment costs of radio frequency chains or high-speed backhaul installations are prohibitively high, and backhaul availability and capacity also create performance bottlenecks [2]. Driven by the fact that only a small portion of the wireless multimedia data is frequently accessed by majority of MUs, wireless content caching, i.e., prefetching popular contents during off-peak times at the edge of wireless networks (base stations (BSs), cache helpers (CHs), or MUs) is a promising approach to unleashing the ultimate potential of the cellular network, alleviating network congestion and improving the quality of experience (QoE) for the MU.

1.1. Related Work

The role of wireless content caching in the oncoming 5G network is demonstrated in [3,4], and its architecture is presented in [5] where the mobile edge applications can be conveniently provided by chaining the service functions with the aid of mobile edge computing and virtualized resources. As academic attention on caching technology has continuously increased, interesting studies have appeared on diverse caching networks and the corresponding caching strategies such as the femto-cellular network [6–12], device-to-device (D2D) network [2,13–17], and heterogeneous network [18]. We discuss these works in additional detail.

In a user-centric femt-cellular network, multiple CHs with cache capacities serve a typical user in a joint transmission manner, and each CH caches only one of the most popular files. The optimal cache distribution is obtained through optimization of the file transmission success probability [8]. In [9], in which the contents are cached into small-cell BSs, the authors designed distributed caching optimization algorithms via belief propagation to minimize the downloading latency with the aid of a factor graph. Cluster-centric small cell networks and a combined caching scheme were proposed [11] in which a protion of the cache space in each small-BS (SBS) cluster was reserved for storing the most popular contents in every SBS, and the remaining space was used to cooperatively cache different partitions of the less popular contents in different SBSs. Based on those studies, the authors offered analysis on the successful content delivery probability.

Cooperation among D2D transmitters was introduced, and two novel hybrid caching strategies, i.e., single-point caching combined with two-point cooperative caching with joint transmission or multi-stream transmission, were proposed, aimed at conserving the energy cost of content deliverers by modeling the locations of D2D transmitters as a Gauss–Poisson process to accurately capture the clustering behaviors [2]. A spatial model for a D2D network was developed in [13] in which the MU locations were modeled as a Poisson cluster process. The authors derived the distributions of distances from a typical device to both intra/inter-cluster devices and analyzed the coverage probability of a typical D2D receiver and the area spectral efficiency of the entire network.

In a three-tier heterogeneous network in which ratio access network caching and D2D caching coexist, a traditional caching strategy known as ‘caching the most popular contents everywhere’ was adopted. The authors developed the corresponding content access protocol and analyzed the average ergodic rate, outage probability, throughput and delay based on the multiclass processor-sharing queue model and continuous-time Markov process [18].

Various deficiencies exist in these above works, i.e., the assumption of node connectivity is critical, the network topology is too idealistic to fully capture the randomness and complexity of MU locations, the cache resource is not fully used, or the caching strategy is too simple to reflect the true feature of caching network. Therefore, we propose a three-tier caching network that reflects the characteristics of the real caching system in which CHs and MUs all have cache capacities and the channel between them is formulated.

The content heterogeneity is also ignored in the literature, where the same size and popularity are assumed for all the contents and the popularity simply follows the same Zipf distribution for every MU. However, three important content attributes, namely, life, popularity, and size, are proposed in [19], and the birth-death process, Zipf distribution, and exponential distribution are adopted for these formulations, respectively. The work in [20] developed a heterogeneous request model that incorporates MU preference for different genres. The authors deduced the content request probability under each genre in the context of social wireless networks. Based on the above related works, we present a heterogeneous content model [file popularity, MU preference, file size] to capture the traits of our file library.

To make full use of the storage capacity of the edge network, the fog ratio access network (F-RAN) is proposed as the evolution of a heterogeneous cloud wireless access network for local content distribution [21–24], providing the possibility to integrate virtualized servers into networks and brings cloud service closer to end device. The edge caching optimization problem of F-RAN in [21] is formulated to find the optimal policy by maximizing the overall cache hit rate. Ref. [23] proposes that uncertainties related to task demands and the different computing capacities of fog nodes inquire an effective load balancing algorithm, and the load balancing problem has been addressed by reinforcement learning under the constraint of achieving the minimum latency. In the paper [24], authors investigate a proactive probabilistic caching optimization in F-RAN and derive the analytical results of successful transmission probability to assess the performance. But these works only focus on the caching capacity of BS or CH and ignore the MU's, resulting in the wasting of resource and poor QoS. Therefore, we take the caching capacity of MUs and CHs into consideration comprehensively and design a two-tier caching strategy.

In addition, stochastic geometry is recognized as a useful tool for identifying the relationships among the network model, caching strategy and network performance. In certain papers [18,20,25], the MU locations were modeled as mutually independent Poisson Point Processes (PPPs), and the authors developed expressions for the distribution of signal to interference plus noise ratio (SINR), the throughput of caching system, and all types of performance evaluation indices based on this approach.

1.2. Contributions and Outcomes

We consider a distributed caching strategy for a software-defined network (SDN) based three-tier network under more practical considerations. Our contributions consist of three components:

- With consideration of the content heterogeneity, we model the content attributes as [file popularity, MU preference, file size], where file popularity denotes the global probability that the file will be requested in the network, the MU preference for file category is the local probability for which category the MU prefers, and the file size represents the scale heterogeneity. Based on these definitions, we define a new local file block library and calculate the local popularity of each file block.
- We proposed an SDN-based three-tier network. In the base station tier (BS-Tier), the BS centered at the macrocell supplies the entire cell with control coverage. In the cache helper tier (CH-tier) and mobile user tier (MU-tier) with cache capacity, CHs distributed in the center of hexagonal grids and randomly distributed MUs supply data coverage together. Taking the 'most popular content' (MPC) and 'largest diversity content' (LDC) into consideration synthetically, a distributed cooperative caching strategy is proposed in which the caches of the MUs store the most popular blocks in the file block library to mitigate the effect of MU mobility and those of the CHs store the remainder in a probabilistic caching manner to enrich the diversity of stored contents and reduce the MU caching pressure.
- Based on the above caching system, we derive the request meet probability (RMPro) and obtain the optimal caching distribution of the contents in a probabilistic caching manner via RMPro optimization. Using the optimization result, we also analyze the content retrieval delays that occur when a typical MU requests a file block or a whole file. A key intermediate step in the analysis is the derivation of the SINR distribution.

The remainder of this paper is organized as follows. We present the system model including the heterogeneous content model, caching network and so on in Section 2. In Section 3 we define RMPro as the performance metric and formulate the main system problem. In Section 4, the content retrieval delays when MU requests a file block or a whole file are analyzed. Simulation results are presented in Section 5 and Section 6 concludes the paper.

2. System Model

In this section, we will introduce a heterogeneous content model and an SDN-based cooperative caching network as well as the corresponding caching strategy.

2.1. Heterogeneous Content Model

The content heterogeneity, which is neglected by most of the literature, includes different MU preferences for files in diverse categories, differences of file size, etc. In this part, let [file popularity, MU preference, file size] denote the content characteristics and a heterogeneous file library containing L files is presented.

(1) **File popularity**: File popularity denotes the global probability that a file will be randomly requested in the network, which follows the Zipf distribution [2]. In the file library $\{f_i, 1 \le i \le L\}$, files are sorted in descending popularity order, and the popularity of i-th file f_i is

$$p_{f_i} = \frac{i^{-\gamma_f}}{\sum\limits_{j=1}^{L} j^{-\gamma_f}}, 1 \le i \le L \tag{1}$$

where γ_f is the Zipf distribution coefficient of the global file popularity, and the larger γ_f is, the higher the popularities that the major popular files capture.

(2) **MU preference**: Different files in the library belong to diverse categories with nonoverlap. The MU preference for a file category is the local probability of which category the MU prefers, mainly determined by a typical factor, i.e., geographic position. This scenario means that MUs with closer locations tend to have similar preferences, and we define the position set where the MUs have the same MU preference as a geographic cluster. We assume that there are K categories $\{c_t, 1 \le t \le K\}$ in the above file library, and a rank-ordered list of K categories represents a local MU preference. The local MU preferences in different geographic clusters are also modeled using the Zipf distribution with different distribution coefficients [20]. The local popularity of t-th category in $\{c_t, 1 \le t \le K\}$ is

$$p_{c_t} = \frac{t^{-\gamma_c}}{\sum\limits_{j=1}^{K} j^{-\gamma_c}}, 1 \le t \le K \tag{2}$$

where γ_c is the coefficient of the local category popularity, and MUs in different geographic clusters have unequal γ_c.

The probability of file f_i among all other files in the category c_t is

$$\mathbb{P}(f_i | f_i inc_t) = \frac{p_{f_i}}{\sum\limits_{f_l \in c_t} p_{f_l}}, 1 \le i \le L, 1 \le t \le K \tag{3}$$

The local file popularity, i.e., the probability that MUs in a specific geographic cluster request f_i of category c_t, is

$$\mathbb{P}(f_i, f_i inc_t) = \mathbb{P}(f_i | f_i inc_t) \times p_{c_t}, 1 \le i \le L, 1 \le t \le K \tag{4}$$

Incorporating Equations (1)–(3) into Equation (4), we have

$$\mathbb{P}(f_i, f_i inc_t) = \frac{i^{-\gamma_f} t^{-\gamma_c}}{\sum\limits_{f_l \in c_t} l^{-\gamma_f} \sum\limits_{k=1}^{K} k^{-\gamma_c}}, 1 \le i \le L, 1 \le t \le K \tag{5}$$

(3) **File size**: File size is another important aspect of content heterogeneity. Because the majority of files have a limited scale and only a small fraction is of large size, the exponential distribution is adopted to approximate the file scale distribution, and we have

$$\mathbb{P}(s_i = x) = \gamma_s e^{-\gamma_s x}, 1 \le i \le L \tag{6}$$

which represents the probability that the size s_i of f_i is equal to x, and γ_s is the exponential distribution coefficient of the file scale distribution.

To simplify, we segment all files into file blocks of the same size M. We denote all file blocks of $f_i, 1 \le i \le L$ as the set $\{f_{i,j}, 1 \le j \le \frac{s_i}{M}\}$, where $f_{i,j}$ is the j-th file block of f_i. Considering that strong correlation exists between the file blocks of the same file, when an MU requests the file block $f_{i,j}$, a high probability exists that the MU will request $f_{i,j+1}$, similar to a Markov process. Thus, we assume that the file popularity is also the popularity of the first file block in this file and denote the transition probability (the probability that an MU requests file block $f_{i,j+1}$ when it has requested $f_{i,j}$) between two adjacent file blocks as σ. The probability that MUs in a certain geographic cluster request the file block $f_{i,j}$, which belongs to category c_t, is

$$\mathbb{P}(f_{i,j}, f_i inc_t) = \sigma^{j-1} \frac{i^{-\gamma_f} t^{-\gamma_c}}{\sum\limits_{f_l \in c_t} l^{-\gamma_f} \sum\limits_{k=1}^{K} k^{-\gamma_c}}, 1 \le i \le L, 1 \le t \le K, 1 \le j \le \frac{s_i}{M} \tag{7}$$

Without loss of generality, we arbitrarily choose a geographic cluster referred to as the representative geographic cluster, which means that γ_c is a constant value. All of the corresponding file blocks $\{f_{i,j}, 1 \le j \le \frac{s_i}{M}, 1 \le i \le L\}$ form a local file block library $\{fb_r, 1 \le r \le L_m\}, L_m = \sum\limits_{j=1}^{L} \frac{s_j}{M}$ in which the file blocks are also sorted in descending popularity order, and the sum of all file block popularities are normalized. Thus, the r-th file block fb_r (which is assumed as $f_{i,j}$, and the index from fb_r to $f_{i,j}$ is applied through a hash table; detail not explained in this work) in the local file block library is requested by the probability of

$$p_{fb_r} = norm(p(f_{i,j}, f_i inc_t)) = norm(\sigma^{j-1} \frac{i^{-\gamma_f} t^{-\gamma_c}}{\sum\limits_{f_l \in c_t} l^{-\gamma_f} \sum\limits_{k=1}^{K} k^{-\gamma_c}})$$
$$1 \le r \le L_m, 1 \le i \le L, 1 \le t \le K, 1 \le j \le \frac{s_i}{M} \tag{8}$$

where $norm()$ is the process of normalization based on the sum of probabilities.

Figure 1 shows the local file block popularity under different values of γ_f, γ_c or σ when $\gamma_s = 4$. When $\gamma_f = 0.8$ and $\gamma_c = 0.8$, the distribution of $\sigma = 0.8$ is more even than that of $\sigma = 0.5$, which is also more flat than that of $\sigma = 0.3$. When $\gamma_c = 0.8$ and $\sigma = 0.5$, the distribution of $\gamma_f = 0.8$ is more even than that of $\gamma_f = 1.2$. When $\gamma_f = 0.8$ and $\sigma = 0.5$, the distribution of $\gamma_c = 0.8$ is slightly smoother than that of $\gamma_c = 1.2$. However, all of these distributions have long tails.

Remark 1. *Based on the heterogeneous content model [file popularity, MU preference, file size], where the Zipf distribution and exponential distribution are adopted for the first two formulations and the last, respectively, the local file block popularity in the local file block library for MUs in the representative geographic cluster is* *a distribution with a long tail. The larger γ_f or γ_c, the higher the local popularities that the major popular file blocks have, whereas a larger σ makes the distribution more even.*

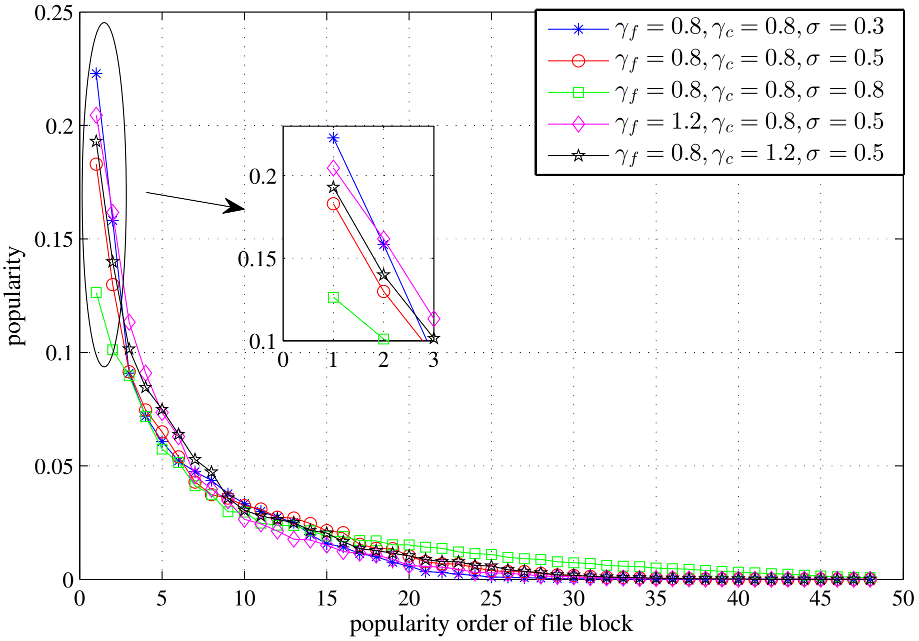


Figure 1. The local file block popularity when γ_f, γ_c and σ get different values and $\gamma_s = 4$.

2.2. Caching Network Architecture

As shown in Figure 2, the caching network consists of three tiers: BS-Tier, CH-Tier and MU-Tier.

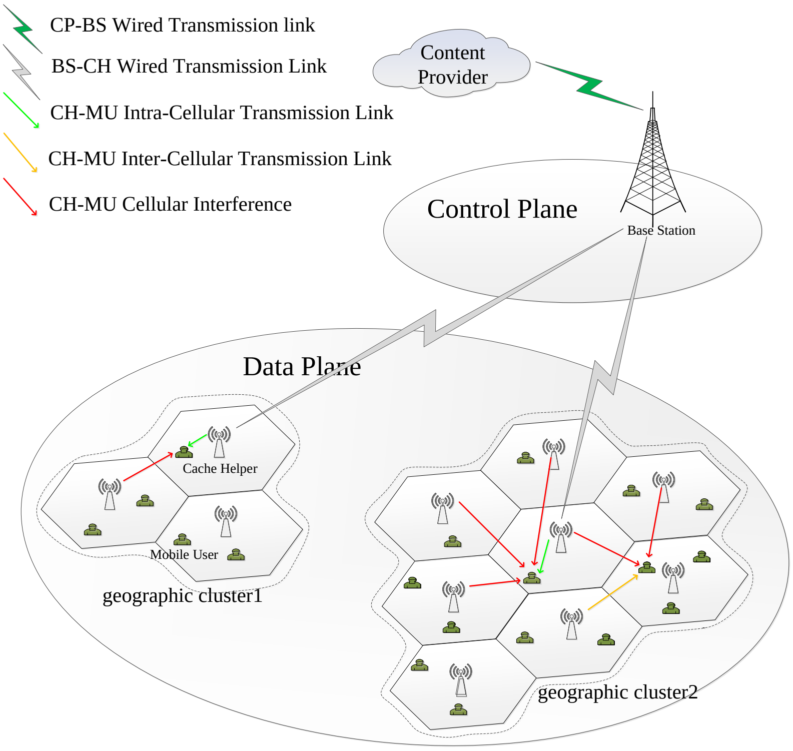

CP-BS Wired Transmission link
BS-CH Wired Transmission Link
CH-MU Intra-Cellular Transmission Link
CH-MU Inter-Cellular Transmission Link
CH-MU Cellular Interference
Content Provider
Control Plane
Base Station
Data Plane
Cache Helper
Mobile User
geographic cluster1
geographic cluster2

Figure 2. The figure of caching network architecture.

In the BS-Tier, the BS can obtain all contents from the content provider (CP) directly through the CP-BS wired transmission link. The macrocell covered by the BS is zoned in two independent manners, i.e., nonoverlapping geographic cluster and hexagonal grid. The geographic cluster is formed because the MUs in close geographic positions have the same MU preference for file category. For example, MUs in geographic cluster1 of Figure 2 might belong to the same financial company, and they usually want to request common financial contents. The hexagonal grid is also a microcell in which CH is situated in the center. According to the CH positions, each hexagonal grid belongs to a corresponding geographic cluster.

In the CH-Tier, each CH possesses a cache of size $N_H \times M$ and a single antenna (the CH is a unit with caching capacity deployed on the SBS (serving microcell), so CH shares the same antenna with the microcell), meaning that a CH can store N_H file blocks at most and transmit only one file block simultaneously. CHs can communicate with the corresponding BS through the BS-CH wired transmission link and with the MUs in its communication range through the CH-MU intra/inter-cellular transmission link.

Considering that the communication range of CH is usually larger than the microcell, we denote the radius of the microcell as d_H and the communication range of CH as R_H in Figure 3, and the radius d_H is obtained by the enclosing circle of a hexagon [26]. Therefors, we have

$$R_H \geq \sqrt{7} d_H \tag{9}$$

To take full advantage of the CH′ communication capability and avoid excessive interference, let $R_H = \sqrt{7} d_H$. The MUs in a certain CH's microcell are known as member-MUs, the CH is the host-CH and there are six neighbor-CHs surrounding every MU. Thus, each CH can directly serve their member-MUs, as well as nonmember-MUs in its communication range when necessary.

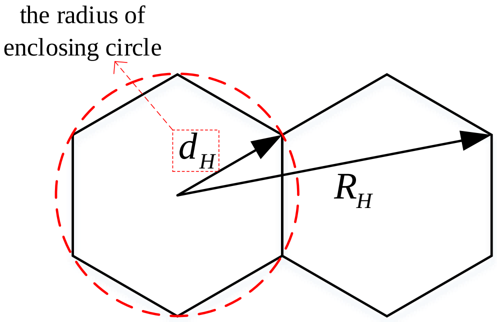


Figure 3. The figure of enclosing circle of a hexagon grid.

In the MU-Tier, all MUs with caches of the same size $N_U \times M$ are distributed in a 2D homogeneous PPP of density λ_U. Therefore, an MU can store N_U file blocks at most. The probability that there are N MUs in a circle area of radius R is

$$\mathbb{P}(N, R, \lambda_U) = \frac{(\pi R^2 \lambda_U)^N}{N!} e^{-\pi R^2 \lambda_U} \tag{10}$$

In view of SDN that decouples the control plane from the data plane, the control coverage supplied by BS with global knowledge of the network state can be further decoupled from the data coverage supplied by CHs and MUs. More specifically, the BS supplies control coverage for the entire macrocell such as the management of the cache placement process and the maintenance of a caching table (knowing which CH or MU stores which contents), etc. CHs are deployed to supply data coverage for the MUs in its communication range, and randomly distribute MUs supply data service to themselves.

Without loss of generality, we focus our analysis on a typical MU chosen arbitrarily in a randomly chosen microcell (representative microcell) inside the representative geographic cluster.

2.3. Cooperative Cache Placement Strategy

In the SDN-based three-tier caching network, caching of contents at the MUs and CHs is an effective way to release wireless traffic pressure and offer better MU QoE.

The cache criteria MPC and LCD must be comprehensively considered for the distributed cache placement strategy. A two-tier caching strategy is adopted. In the MU-Tier, all MUs located in the representative geographic cluster store the N_U most popular blocks of the local file block library, which meets the expectation of MPC. Because all MUs cache the same most popular contents, the mobility of MUs in the same geographical cluster has no impact on the caching performance of the MU-Tier. And the scene of high mobility in 5G area is beyond our consideration. In the CH-Tier, the remaining $L_m - N_U$ file blocks are independently cached in the CHs in a probabilistic caching manner, which enriches the diversity of the stored contents and reduces the MU caching pressure by cooperative caching.

We denote the caching distribution in the probabilistic caching as $Q = \{q_i, N_U + 1 \le i \le L_m\}$, where q_i is the caching probability of i-th file block fb_i and $\sum\limits_{q_i \in Q} q_i = N_H$. In view of the local file block popularity distribution with a long tail and $\lim\limits_{r \to L_m} p_{fb_r} \to 0$, the Zipf distribution has a similar feature and is considered as a heuristic choice for matching this caching distribution Q [14]. To make the best of every CH cache, we define an Ext-Zipf distribution and assume $Q\tilde{}$Ext-Zipf(θ, N_H), the probability sum of which is equal to N_H, where

$$q_i = \frac{N_H (i - N_U)^{-\theta}}{\sum\limits_{j=1}^{L_m - N_U} j^{-\theta}}, N_U + 1 \le i \le L_m \tag{11}$$

When the typical MU requests a file block, it can be obtained in the following four ways shown in Figure 4.

- Local-Cache-provide: If the requested file block belongs to the first N_U in the local file block library, the typical MU can obtain it directly from its local cache without any delay.
- Host-CH-provide: If the popularity order of the requested file block ranges from $N_U + 1$ to L_m in the local file block library, it might be stored in the host-CH or neighbor-CHs. If it is stored in the host-CH, the block can be obtained from the host-CH through the CH-MU intra-cellular transmission link with a delay time. Considering that it is impossible that all CHs are always idle and waiting to send requested contents, let μ $(0 \le \mu \le 1)$ denote the proportion of active CHs. Because the communication range R_H is larger than the microcell radius d_H, there is a probability μ of CH-MU cellular interference from one of the six neighbor-CHs.
- Neighbor-CH-provide: The popularity order of the requested file block also ranges from $N_U + 1$ to L_m, but it is cached in the neighbor-CHs instead, and thus, it can be obtained from certain neighbor-CHs through the CH-MU inter-cellular transmission link with delay. The probability $\mu(1 - q_i)$ exists that CH-MU cellular interference comes from one of the remaining neighbor-CHs.
- CP-BS-provide: When the content request cannot be met in the above three ways, the BS should obtain the contents from CP and transmit it to the host-CH via a wired link with a constant delay δ. The host-CH retransmits the contents to typical MU.

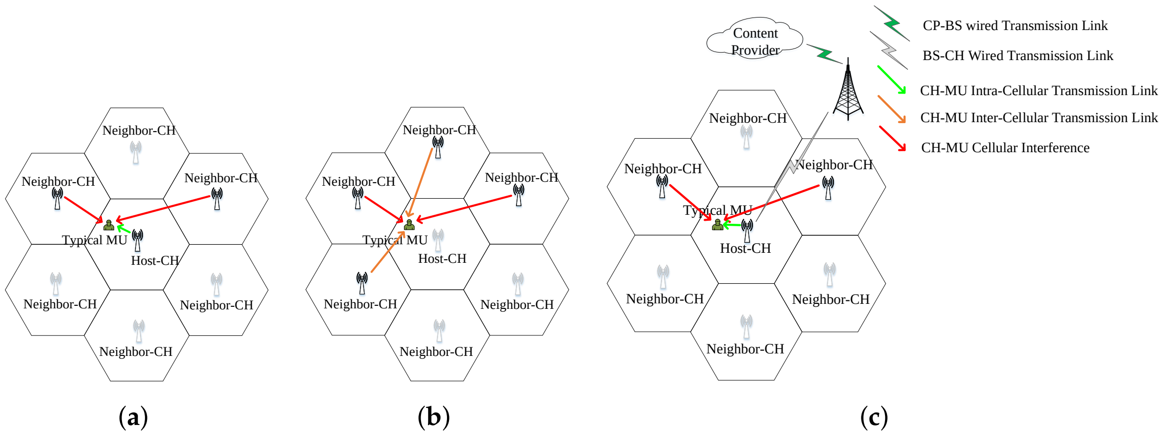

Content Provider
CP-BS wired Transmission Link
BS-CH Wired Transmission Link
CH-MU Intra-Cellular Transmission Link
CH-MU Inter-Cellular Transmission Link
CH-MU Cellular Interference
Neighbor-CH
Typical MU
Host-CH
(a)
(b)
(c)

Figure 4. Illustration of ways to obtain the requested file block. (**a**) Host-CH-provide. (**b**) Neighbor-CH-provide. (**c**) CP-BS-provide.

The BS controls the execution sequence of the above four methods, and the priorities are ordered as follows: Local-Cache-provide > Host-CH-provide > Neighbor-CH-provide > CP-BS-provide.

2.4. Delivery Model

Due to the stationarity of the cache network, we assume that the typical MU is located at the origin. We denote the locations of its host-CH and six neighbor-CHs as y_0 and $\{y_i, 1 \leq i \leq 6\}$, respectively. The distance between the typical MU and its host-CH or a neighbor-CH is $r_i = ||y_i||, 0 \leq i \leq 6$.

Lemma 1. *The probability density function (PDF) of the distance r_0 between the typical MU and host-CH is:*

$$f_{r_0}(x) = \frac{2\pi\lambda_U x e^{-\pi\lambda_U x^2}}{1 - e^{-\pi\lambda_U d_H{}^2}}, 0 \leq x \leq d_H \tag{12}$$

The PDF of the distance $r_i, 1 \leq i \leq 6$ between the typical MU and a certain neighbor-CH is:

$$f_{r_i}(x) = \frac{2\pi\lambda_U x e^{-\pi\lambda_U x^2}}{e^{-\frac{3}{4}\pi\lambda_U d_H{}^2} - e^{-7\pi\lambda_U d_H{}^2}}, \frac{\sqrt{3}}{2} d_H \leq x \leq \sqrt{7} d_H \tag{13}$$

Proof. See Appendix A. □

In Local-Cache-provide, the requested contents can be obtained quickly and accurately. In the latter three methods, where the transmission of requested file blocks is partially wireless and time-consuming, the corresponding SINR experienced by the typical MU should exceed a certain predetermined SINR threshold Γ for successful demodulation and decoding.

In Host-CH-provide, the serving transmitter is the host-CH, and interference comes from certain of six neighbor-CHs. The SINR at the typical MU is

$$SINR_{HCH} = \frac{S_0}{\sum\limits_{i=1}^{6} A_i I_i + \sigma^2} = \frac{P_H h_0 r_0{}^{-\beta}}{\sum\limits_{i=1}^{6} A_i P_H h_i r_i{}^{-\beta} + \sigma^2} \overset{(a)}{=} \frac{h_0 r_0{}^{-\beta}}{\sum\limits_{i=1}^{6} A_i h_i r_i{}^{-\beta}} \tag{14}$$

where P_H is the transmission power of CH, β is the pathloss coefficient, σ^2 denotes the additive noise power, the Rayleigh fading coefficient $h_i \sim \exp(1), 0 \leq i \leq 6$, A_i˜Bernoulli(μ) and r_0 (or $r_i, 1 \leq i \leq 6$) is the distance between the typical MU and host-CH (or a neighbor-CH), the PDF of which follows formula (12) (or (13)). Because we only consider the interference-limited network, we have step (a).

In Neighbor-CH-provide, the transmitters are certain neighbor-CHs that jointly transmit, and interference comes from other neighbor-CHs. The SINR experienced by the typical MU is

$$SINR_{NCH} = \frac{\sum\limits_{i=1}^{6} B_i S_i}{\sum\limits_{j=1}^{6} C_j I_j + \sigma^2} = \frac{\sum\limits_{i=1}^{6} B_i h_i {r_i}^{-\beta}}{\sum\limits_{j=1}^{6} C_j h_j {r_j}^{-\beta}} \tag{15}$$

where B_i˜Bernoulli$((1-\mu)q_i)$, C_i˜Bernoulli$(\mu(1-q_i))$, and $(1-\mu)q_i$ is the probability that the requested file block is transmitted by a certain neighbor-CH.

In CP-BS-provide, we analyze a macrocell in isolation, and the SINR is the same as that in Host-CH-provide.

$$SINR_{BS} = SINR_{HCH} \tag{16}$$

3. Performance Metrics and Problem Formulation

In this section, a performance metric for the caching system is first derived. Second, a system problem is formulated based on the metric, and an optimal scheme is proposed.

3.1. Request Meet Probability

The request meet probability (RMPro) takes the successful transmission rate of the contents into consideration based on the cache hit probability [20]. RMPro is formally defined as the probability that the typical user can successfully obtain the requested file block. We denote RMPro and the probability that the requested file block is obtained by Local-Cache-provide, Host-CH-provide, Neighbor-CH-provide, or CP-BS-provide as $\mathbb{P}_{RMPro}$, $\mathbb{P}_{MU}$, $\mathbb{P}_{HCH}$, $\mathbb{P}_{NCH}$ and $\mathbb{P}_{BS}$ respectively, and we have

$$\mathbb{P}_{RMPro} = \mathbb{P}_{MU} + \mathbb{P}_{HCH} + \mathbb{P}_{NCH} + \mathbb{P}_{BS} \tag{17}$$

(1) When the requested file block is obtained by Local-Cache-provide, we have

$$\mathbb{P}_{MU} = \sum_{i=1}^{N_U} p_{fb_i} \tag{18}$$

(2) When the requested file block is obtained by Host-CH-provide,

$$\mathbb{P}_{HCH} = \sum_{i=N_U+1}^{L_m} p_{fb_i} q_i \mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i) \tag{19}$$

where $\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i)$ is the successful transmission rate when the host-CH supplies file block $fb_i, N_U + 1 \leq i \leq L_m$ for the typical MU.

Lemma 2. *When the host-CH supplies the requested file block $fb_i, N_U + 1 \leq i \leq L_m$ for the typical MU, the successful transmission rate is*

$$\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i) = \eta \sum_{j=1}^{N} \omega_j \varphi(x_j) \tag{20}$$

where

$$\eta = \frac{d_H}{2(e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2})^6 (1 - e^{-\pi\lambda_U {d_H}^2})}$$

$$\varphi(x) = 2\pi\lambda_U(\frac{d_H}{2}x + \frac{d_H}{2})e^{-\pi\lambda_U(\frac{d_H}{2}x+\frac{d_H}{2})^2}$$
$$\times [e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2} - 2\pi\mu\lambda_U \int\limits_{\frac{\sqrt{3}}{2}d_H}^{\sqrt{7}d_H} \frac{r_1 e^{-\pi\lambda_U {r_1}^2}}{1 + {r_1}^\beta/(\Gamma(\frac{d_H}{2}x + \frac{d_H}{2})^\beta)} dr_1]^6 \tag{21}$$

In the above equations, N is the number of quadrature nodes, ω_j is the weight coefficient, and x_j is the j-th root of the Legendre polynomials.

Proof. See Appendix B. □

Remark 2. *The successful transmission rate $\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i)$ is independent of the transmitted file block fb_i, and thus, the shorthand of $\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i)$ can be written as $\mathbb{P}(SINR_{HCH} \geq \Gamma)$. Because the transmitter is doubtlessly the host-CH in Host-CH-provide, $\mathbb{P}(SINR_{HCH} \geq \Gamma)$ is a constant value with predetermined system parameters, which do not depend on the caching probability. The interference comes from neighbor-CHs, and thus, the larger the proportion of active CHs μ, the more interference accumulates, and $\mathbb{P}(SINR_{HCH} \geq \Gamma)$ decreases. A larger value of SINR threshold Γ also produces a smaller $\mathbb{P}(SINR_{HCH} \geq \Gamma)$.*

(3) When the requested file block is obtained by Neighbor-CH-provide,

$$\mathbb{P}_{NCH} = \sum_{i=N_U+1}^{L_m} p_{fb_i}(1 - q_i)\left[1 - (1 - q_i)^6\right]\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i) \tag{22}$$

where $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ is the successful transmission rate when certain of the six neighbor-CHs supply the file block $fb_i, N_U + 1 \leq i \leq L_m$ for the typical MU and the host-CH does not work.

Lemma 3. *In the Neighbor-CH-provide, the successful transmission rate $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ of file block $fb_i, N_U + 1 \leq i \leq L_m$ is*

$$\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i) = \begin{cases} 1, N_U + 1 \leq i \leq \varepsilon \\ 0, \varepsilon < i \leq L_m \end{cases} \tag{23}$$

where $\varepsilon = \min(\left\lfloor (\frac{\mu\Gamma \sum_{j=1}^{L_m - N_U} j^{-\theta}}{(1-\mu+\mu\Gamma)N_H})^{-\frac{1}{\theta}} \right\rfloor + N_U, L_m)$, and $\lfloor * \rfloor$ is an operation whose result is the maximal integer below $*$.

Proof. See Appendix C. □

Remark 3. *The successful transmission rate $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ is highly dependent on which file block fb_i is transmitted. The indicator ε represents which file blocks can be received successfully by Neighbor-CH-provide, and thus, we can make full sense of $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ by analysis of the indicator ε. The signal and interference all come from neighbor-CHs and thus, the larger the proportion of active CHs μ, the less signal is superposed and the more interference accumulates such that ε decreases. A larger threshold Γ also results in a smaller ε.*

(4) When the requested file block is obtained by CP-BS-provide, the successful transmission rate is the same as that of Host-CH-provide.

$$\mathbb{P}_{BS} = \sum_{i=N_U+1}^{L_m} p_{fb_i}(1 - q_i)^7\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i) \tag{24}$$

3.2. Problem Formulation

The aim of the SDN-based cooperative caching system is to maximize RMPro,

$$\begin{aligned} &\max_{\theta} \mathbb{P}_{RMPro} \\ &s.t. \theta > 0 \end{aligned} \tag{25}$$

To find the maximal RMPro, we take the derivative of $\mathbb{P}_{RMPro}$ with respect to θ. Although finding θ_{Opt} analytically in a closed form does not seem possible, numerical analyses are probable with low effort. We adopt a binary search on function (25) of one variable θ ranging from 0 to 1.

4. Analysis of the Content Retrieval Delay

The content retrieval delay is another important performance characteristic of the caching system and is defined as the average delay experienced by typical MU when retrieving random requested content from any available source. In this section, we analyze the content retrieval delays that occur when the typical MU asks for a file block or a whole file.

4.1. Content Retrieval Delay Of Request For A File Block

From the view of the typical MU, the content retrieval delay of a request for a file block is its average transmission time. Depending on the source, four cases exist.

- When the requested file block is stored in the local cache of typical MU, it is assumed that the transmission rate is infinite, and the delay time tends to be 0, i.e.,

$$T_{MU} = 0 \tag{26}$$

- When the requested file block is obtained in the manner of Host-CH-provide, the delay time is

$$T_{HCH} = \frac{M}{R_{HCH}} = \frac{M}{W \log(\mathbb{E}(SINR_{HCH}) + 1)} \tag{27}$$

  where R_{HCH} is the average transmission rate when the host-CH is the transmitter and certain neighbor-CHs act as interference-makers, W is the transmission bandwidth of CH and $\mathbb{E}(*)$ is the mean of $*$.

Lemma 4. *The average transmission rate of a file block fb_i, $N_U + 1 \leq i \leq L_m$ by Host-CH-provide is*

$$R_{HCH} = W \log(\frac{(e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2})}{6\mu(1 - e^{-\pi\lambda_U {d_H}^2})} \frac{\gamma(\frac{2-\beta}{2}, \pi\lambda_U {d_H}^2)}{\gamma(\frac{2-\beta}{2}, 7\pi\lambda_U {d_H}^2) - \gamma(\frac{2-\beta}{2}, \frac{3}{4}\pi\lambda_U {d_H}^2)} + 1) \tag{28}$$

which is independent of which file block fb_i is transmitted and is also a constant value with predetermined caching system parameters. In the formula, $\gamma(a,b) = \int_0^b x^{a-1}e^{-x}dx$ is the well-known incomplete gamma function.

Proof. See Appendix D. □

- When the requested file block is transmitted to the typical MU by Neighbor-CH-provide, the delay time is

$$T_{NCH} = \frac{M}{R_{NCH}} = \frac{M}{W \log(\mathbb{E}(SINR_{NCH}) + 1)} \tag{29}$$

  where R_{NCH} is the average transmission rate while the neighbor-CHs are not only transmitters but interference-makers.

Lemma 5. *The average transmission rate of a file block $fb_i, N_U + 1 \leq i \leq L_m$ by Neighbor-CH-provide is*

$$R_{NCH} = W \log(\frac{(1-\mu)q_i}{\mu(1-q_i)} + 1) \tag{30}$$

which is highly dependent on the caching probability of file block $f_i, N_U + 1 \leq i \leq L_m$ with fixed system parameters.

Proof. According to Appendix C, the average SINR at the typical MU by Neighbor-CH-provide is

$$\mathbb{E}(SINR_{NCH}) = \frac{(1-\mu)q_i}{\mu(1-q_i)} \tag{31}$$

Based on the Shannon equation, the average transmission rate is shown in Formula (30). □

- When the requested file block is supplied by CP-BS-provide, the delay time includes two components. One component is the time spent on the transmission from CP to the host-CH via the BS, which is assumed as a constant value δ, and the other is the transmission time from the host-CH to the typical MU, which is the same as T_{HCH}.

$$T_{BS} = \delta + T_{HCH} \tag{32}$$

Combining the four cases, the content retrieval delay of a random request for a file block is

$$\begin{aligned} T_{fb} = &\sum_{i=1}^{N_U} p_{fb_i} T_{MU} + \sum_{i=N_U+1}^{L_m} p_{fb_i} q_i T_{HCH} + \sum_{i=N_U+1}^{L_m} p_{fb_i}(1-q_i)\left[1-(1-q_i)^6\right] T_{NCH} \\ &+ \sum_{i=N_U+1}^{L_m} p_{fb_i}(1-q_i)^7 T_{BS} \end{aligned} \tag{33}$$

4.2. Content Retrieval Delay of Request for a Whole File

When the typical MU requests a file, the retrieval delay is defined as the possible maximal spent time. According to the analysis of the previous section, the content retrieval delay of a request for a whole file is the maximum time spent on the combination of Host-CH-provide and CP-BS-provide or Neighbor-CH-provide and CP-BS-provide.

In the first combination, all file blocks that the requested file includes are assumed to be obtained by Host-CH-provide or CP-BS-provide. The transmission of these file blocks requires the antenna of host-CH so that the blocks must remain in a waiting queue. The caching process for each file block $fb_l, N_U + 1 \leq l \leq L_m$ is independent and follows the caching distribution Q. When the typical MU requests file $f_i, 1 \leq i \leq L$, the transmission time of all file blocks via the combination of Host-CH-provide and CP-BS-provide is

$$T_{HB}(f_i) = \sum_{fb_t \in f_i} \left\{ q_t \sum_{i=1}^{\min\{N_H, |\{fb_t|fb_t \in f_i\}|\}} T_H(fb_t, i) + (1-q_t)^7 \sum_{j=1}^{|\{fb_t|fb_t \in f_i\}|} T_B(fb_t, j) \right\} \tag{34}$$

where $\{fb_t|fb_t \in f_i\}$ denotes all file blocks that f_i includes, $|*|$ is the cardinality of set $*$, $T_H(fb_t, i)$ is the average transmission time of file block fb_t (the i-th in line to be sent by Host-CH-provide), and $T_B(fb_t, j)$ is the time of fb_t (the j-th in line to be sent by CP-BS-provide),

$$T_H(fb_t, i) = T_{HCH} \mathbb{P}(i-1, \{fb_l|fb_l \in f_i, l \neq t\}) \tag{35}$$

$$T_B(fb_t, j) = T_{BS} \mathbb{P}(j-1, \{fb_l|fb_l \in f_i, l \neq t\}) \tag{36}$$

where $\mathbb{P}(i-1, \{fb_l | fb_l \in f_i, l \neq t\})$ is the probability that there are already $i-1$ file blocks of $\{fb_l | fb_l \in f_i, l \neq t\}$ stored in the transmission queue of a certain CH or BS, and $\{fb_l | fb_l \in f_i, l \neq t\}$ is the file block set of f_i except the one fb_t being transmitted.

In the second combination, all file blocks contained are transmitted by Neighbor-CH-provide or CP-BS-provide. The transmission of these file blocks requires different antennas, i.e., the antennas of the neighbor-CHs or the antenna of the host-CH. Therefore, when the typical MU requests a file $f_i, 1 \leq i \leq L$, the delay time is the maximum of the average transmission time of file blocks by Neighbor-CH-provide or CP-BS-provide.

$$T_{NB}(f_i) = \max\left\{ \sum_{fb_t \in f_i} (1-q_t)\left[1-(1-q_t)^6\right] \sum_{i=1}^{\min\{N_H, |\{fb_t | fb_t \in f_i\}|\}} T_N(fb_t, i), \right. \\ \left. \sum_{fb_t \in f_i} (1-q_t)^7 \sum_{j=1}^{|\{fb_t | fb_t \in f_i\}|} T_B(fb_t, j) \right\} \tag{37}$$

where $T_N(fb_t, i)$ is the average transmission time of fb_t (the i-th in line to be sent by Neighbor-CH-provide),

$$T_N(fb_t, i) = T_{NCH}\mathbb{P}(i-1, \{fb_l | fb_l \in f_i, l \neq t\}) \tag{38}$$

Combining (34) and (37), the average delay time of request for a whole file is

$$T_f = \sum_{i=1}^{L} \{\mathbb{P}(f_i, f_i inc_t)\{\max\{T_{HB}(f_i), T_{NB}(f_i)\}\}\} \tag{39}$$

where $\mathbb{P}(f_i, f_i inc_t)$ is the local file popularity from the point of the typical MU.

5. Performance Evaluation

In this section, we evaluate the performance of the proposed schemes by MATLAB. The simulation setup and performance analysis are presented as follows.

5.1. Simulation Setup

According to the references [15,18–20], we set most of parameters shown in Tables 1 and 2. The transition probability between two adjacent file blocks represents the correlation inside a file. And the probability value ranging from 0.3 to 0.8, indicates that the relevance of the contents inside a file varies from weaken to strong. Considering the limited wired transmission link from CP to CH via BS, we assume the general size of file block is 5 Mbit and the transmission rate is 1 Mbps. Therefore, the transmission time from CP to host-CH is 5 s. The parameter setting for a heterogeneous file library is shown in Table 1, where the three main elements of content heterogeneity are set as several values for numerical evaluation. The parameter settings of the cooperative caching network are shown in Table 2, which includes several values of the proportion of the active CHs and SINR threshold.

Two schemes are compared with the proposed caching system: one is the popularity-based caching policy (PBCP) [4] in which each MU or CH stores file blocks according to the popularity distribution, and the other is the uniform caching policy (UNCP) in which each MU or CH stores file blocks from the library uniformly.

Table 1. Parameter setting of the heterogeneous file library.

Parameters	Values
Number of files in the library (L)	10
Number of file categories in the library (K)	3
The Zipf distribution coefficient of global file popularity (γ_f)	0.8, 1.2 (default)
The Zipf distribution coefficient of local file category popularity (γ_c)	0.8 (default), 1.2
The transition probability between two adjacent file blocks (σ)	0.3, 0.5 (default), 0.8
The exponential distribution coefficient of file scale distribution (γ_s)	4
Size of file blocks in the library (M)	5 (Mbit)
Number of file blocks in the library (L_m)	48

Table 2. Parameter setting of the cooperative caching network.

Parameters	Values
Number of file blocks that can be cached in the CH cache (N_H)	5
Number of file blocks that can be cached in the MU cache (N_U)	2
The density of MU PPP distribution (λ_U)	$5000/(500^2\pi)$
The communication range of CH (R_H)	100 (m)
The radius of microcell (d_H)	$100/\sqrt{7}$ (m)
The proportion of active CHs (μ)	1×10^{-4}, 0.2, 0.5, 0.8, 1
The pathloss coefficient (β)	4
The SINR threshold (Γ)	1×10^{-4}, 0.3
The transmission power of CH (P_H)	0.2 (W)
The bandwidth of CH transmission (W)	1 (Mbps)
The transmission time from CP to host-CH via BS (δ)	5 (s)

5.2. Simulation Result And Analysis

In this section, we analyze the influence of the proportion of active CHs μ or the SINR threshold Γ on the caching system performance.

Figure 5 shows the effect of the SINR threshold Γ on the successful transmission rate $\mathbb{P}(SINR_{HCH} \geq \Gamma)$ under different proportions of active CHs μ by Host-CH-provide. The larger the value μ, the more interference accumulates at typical MU, and $\mathbb{P}(SINR_{HCH} \geq \Gamma)$ decreases. A larger value of Γ also results in a smaller $\mathbb{P}(SINR_{HCH} \geq \Gamma)$. The impact of Γ on $\mathbb{P}(SINR_{HCH} \geq \Gamma)$ is nearly linear. Because there is no inference at the typical MU when $\mu = 0$, the received signal of the typical MU can always be successfully demodulated and decoded, and there is no effect of Γ variation on $\mathbb{P}(SINR_{HCH} \geq \Gamma)$, which remains a high value.

In addition, ε is the key indicator that decides the value of the successful transmission rate $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ by Neighbor-CH-provide in formula (23) and represents which file blocks can be received successfully. As shown in Figure 6, the larger the proportion of active CHs μ, the less signal superposition and the more interference accumulates such that ε decreases and fewer blocks can be obtained successfully. A larger threshold Γ also produces a smaller ε. Because there is no inference-maker among the neighbor-CHs when $\mu = 0$, there is also no effect of Γ variation on ε, and ε has a high value, which means that most blocks can be obtained successfully. Because none of the neighbor-CHs can act as the potential transmitter when $\mu = 1$, there is no probability of successful demodulation and decoding of file blocks $fb_i, N_U \leq i \leq L_m$ and $\varepsilon = N_U = 2$, i.e., the number of file blocks locally cached in the MU.

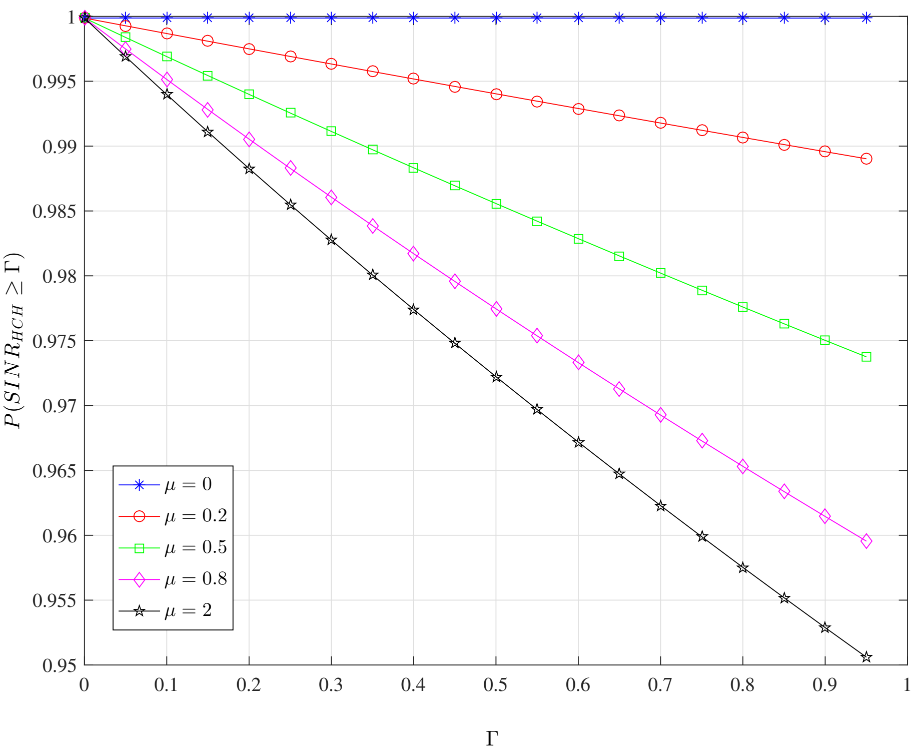


Figure 5. The effect of Γ on $\mathbb{P}(SINR_{HCH} \geq \Gamma)$ under different μ in Lemma 2.

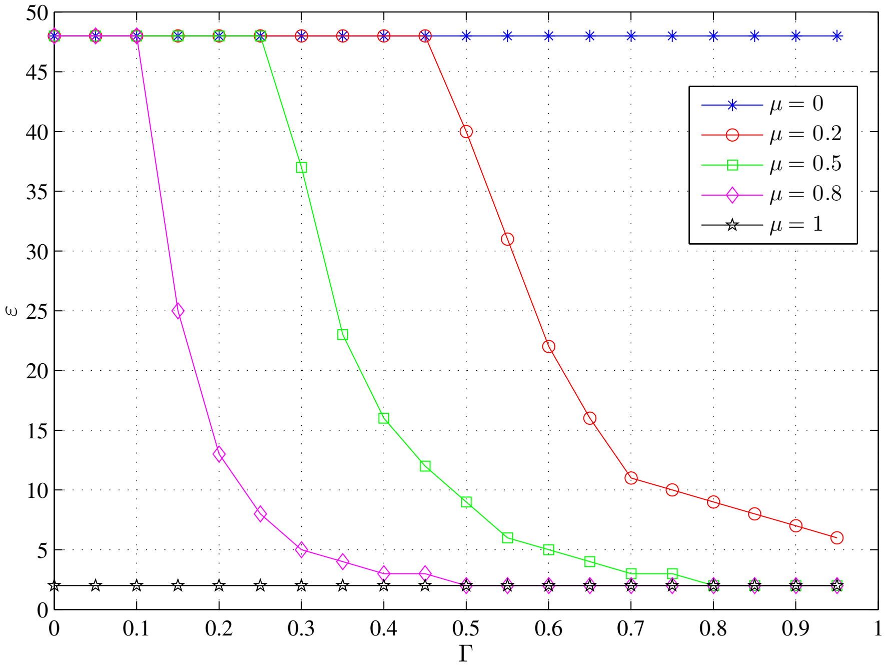


Figure 6. The effect of Γ on ε of $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ under different μ in Lemma 3.

As shown in Figure 7, RMPro decreases as the SINR threshold Γ or proportion of active CHs μ increases. Regardless of how the parameter Γ or μ changes, RMPro always retains a value greater than 0.83. Except for Local-Cache-provide, a larger value of Γ means that successful reception of the requested file blocks is a more challenging process for the last three methods of obtaining contents.

A larger value of μ means that there is more interference from neighbor-CHs at the typical MU, which results in the hardship of demodulation and decoding.

According to Figure 7, we should reasonably increase the distribution density of CHs in deployment to ensure that some CHs are idle and μ remains at a relatively low number. At the same time, the process of demodulation and decoding must be optimized to ensure a high successful transmission rate.

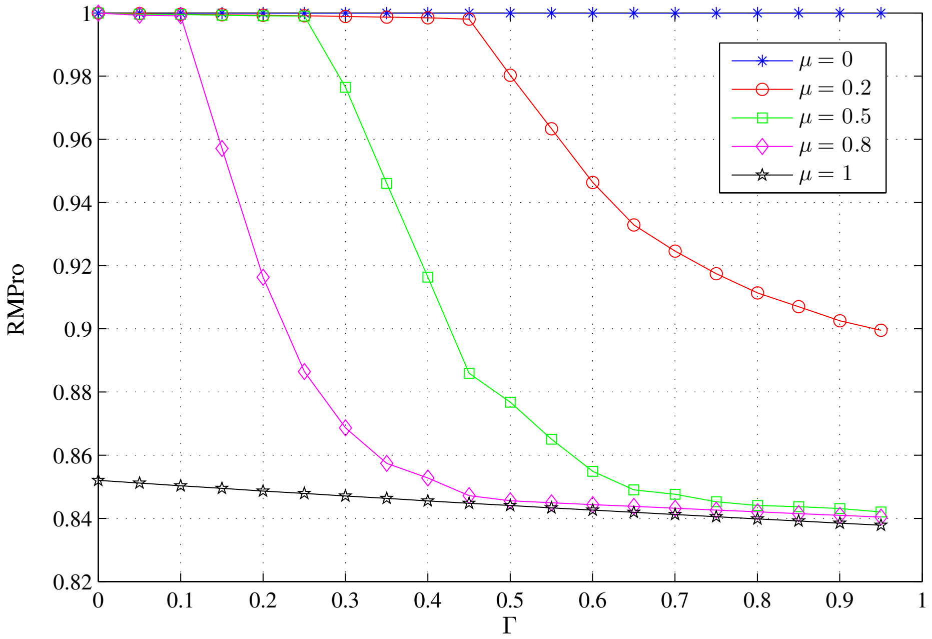


Figure 7. The effect of Γ on the RMPro under different μ.

Figure 8 indicates that the effects of Γ or μ on RMPro of PBCP or UNCP are the same as in the proposed scheme, and RMPro of the proposed scheme is always larger than that of the other two, even by a little. When Γ is a smaller number, it is obvious that RMPro of our proposed scheme is far better than that of the others and is especially better than that of UNCP. When Γ is greater than a certain number, the performance of the three schemes more or less converge.

Regardless of the popularity, all contents are always stored evenly in the UNCP strategy and, thus, the retrieval delay of UNCP must be a high value. In the PBCP strategy, the less popular contents are always transmitted from CP via BS, where the delay caused is also immeasurable, especially when the cache capacity of MU or CH are small. The retrieval delay caused by these two strategies is not comparable to the proposed scheme. Therefore, we only consider the effects of μ or Γ on the delay of our proposed scheme.

Figures 9 and 10 show the impact of the proportion of active CHs μ on the content retrieval delays under different SINR thresholds Γ when requesting a file block or a whole file, respectively, indicating that the content retrieval delay increases with parameter μ. In more detail, when μ remains within a small range of values, the delay time does not change substantially with μ variation, but the delay time increases exponentially when μ varies beyond a certain value. It is noted that parameter Γ has little influence on the content retrieval delay because Γ affects the delay time only by its effect on the successful transmission rate. According to the two figures, we should set μ at a relatively low number to control the content retrieval delay, which is the same as our optimization objective of the caching system.

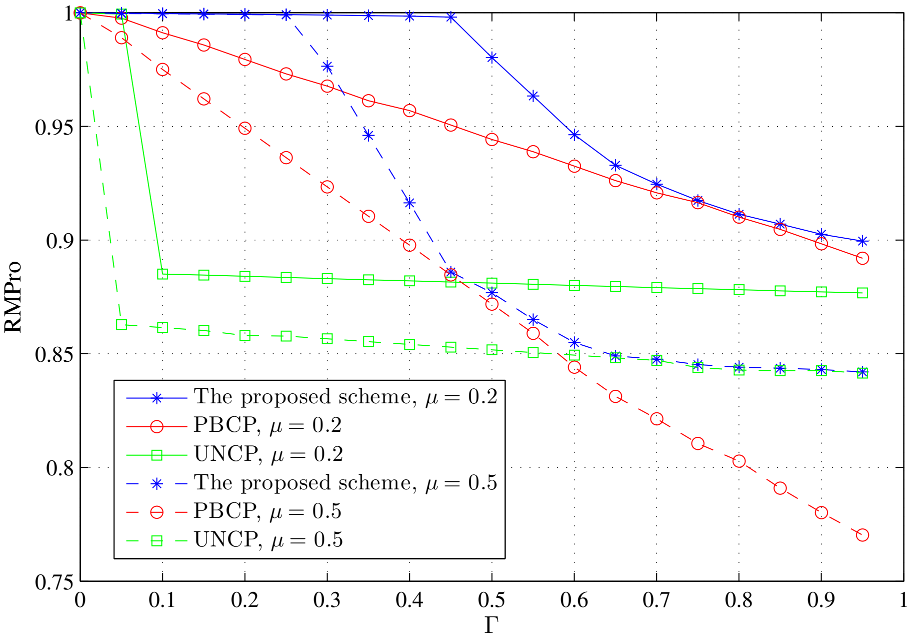


Figure 8. The comparison between PBCP, UNCP and the proposed scheme under different values of Γ or μ.

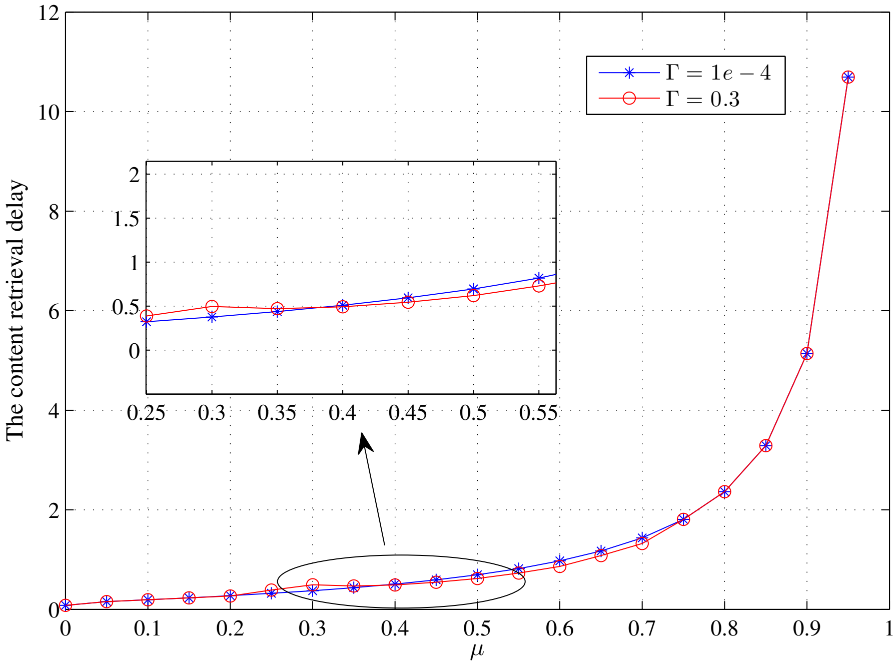


Figure 9. The effect of μ on the content retrieval delay under different Γ when requesting for a file block randomly.

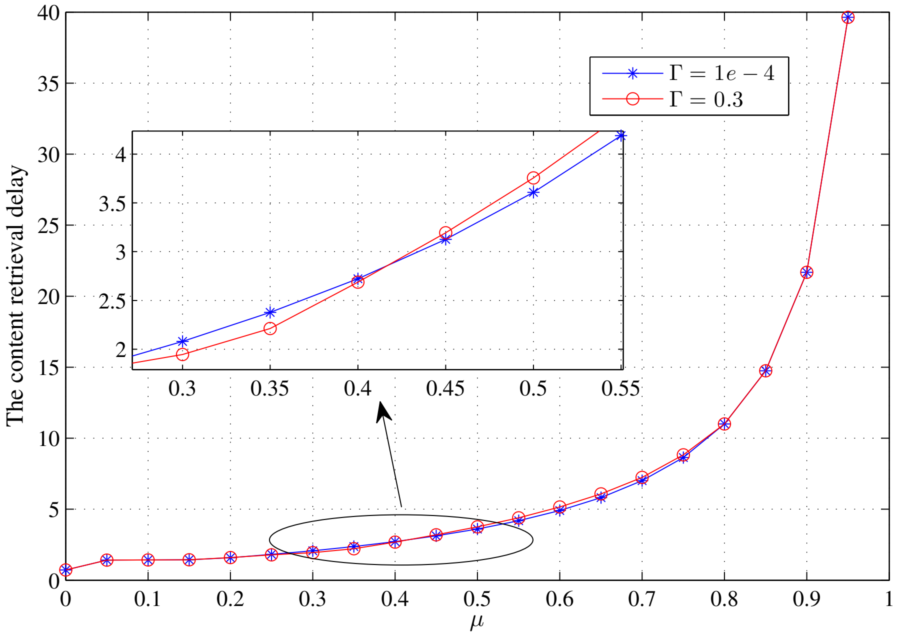


Figure 10. The effect of μ on the content retrieval delay under different Γ when requesting for a whole file randomly.

6. Conclusions

In this paper, we proposed an SDN-based cooperative caching system. First, we modeled the content attributes as [file popularity, MU preference, file size] and defined a local file block library. Second, taking MPC and LDC into consideration, an SDN-based three-tier cooperative caching network was proposed in which the caches of the MUs stored the most popular contents of the file block library to mitigate the effect of MU mobility and those of the CHs cached the remaining contents in a probabilistic caching manner to enrich the diversity of the stored contents and reduce the MU caching pressure. A performance metric RMPro was also derived, and the optimal caching distribution of the probabilistic caching scheme was obtained via optimization. Finally, using the optimization result, we also analyzed the content retrieval delays that occur when the typical MU randomly requested a file block or a whole file. Simulation results indicated that the proposed caching system could achieve quasi-optimal revenue performance compared with other contrasting schemes. We also suggested that the density of CH should increase reasonably in the deployment and that the process of demodulation and decoding must be optimized.

In future work, we will enrich the content heterogeneity, i.e., the life-time of the file, and D2D communication will be considered for the cooperative caching system.

Author Contributions: Conceptualization, Q.L. (Qi Li); Data curation, X.W. and D.W.; Formal analysis, Q.L. (Qi Li), Y.L. and Q.L. (Qiang Liu); Funding acquisition, D.W.; Investigation, Q.L. (Qi Li) and L.S.; Methodology, Q.L. (Qi Li) and Y.Z.; Project administration, D.W.; Supervision, X.W. and D.W.; Validation, Y.L.; Writing—original draft, Q.L. (Qi Li); Writing—review editing, Q.L. (Qi Li)

Funding: This research was funded in part by the National key R&D Program of China No.2018YFC1504502, and in part by the National Natural Science Foundation of China under Grant 61701038.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SDN	Software-Defined Network
MU	Mobile User
CH	Cache Helper
RMPro	Request Meet Probability
BS	Base Station
QoE	Quality of Experience
D2D	Device-to-Device
SBS	Small-BS
F-RAN	Fog Ratio Access Network
PPP	Poisson Point Processe
SINR	Signal to Interference plus Noise Ratio
BS-Tier	Base Station Tier
CH-Tier	Cache Helper Tier
MU-Tier	Mobile User Tier
MPC	Most Popular Content
LDC	Largest Diversity Content
CP	Content Provider
PDF	Probability Density Function
PBCP	Popularity-Based Caching Policy
UNCP	UNiform Caching Policy

Appendix A. Proof of Lemma 1

As shown in Figure A1a, the distance between the typical MU and its host-CH is r_0, meaning that there exists at least one MU in the circular region of a radius r_0 centered at host-CH within the region of the representative microcell. According to formula (10), we have

$$\mathbb{P}(r_0 \leq x) = 1 - \mathbb{P}(0, x, \lambda_U) = 1 - e^{-\pi\lambda_U x^2} \tag{A1}$$

The host-CH only serves MUs in the representative microcell range of $0 \leq r_0 \leq d_H$. By the derivation and process of normalization based on the sum of probabilities, the PDF of r_0 is

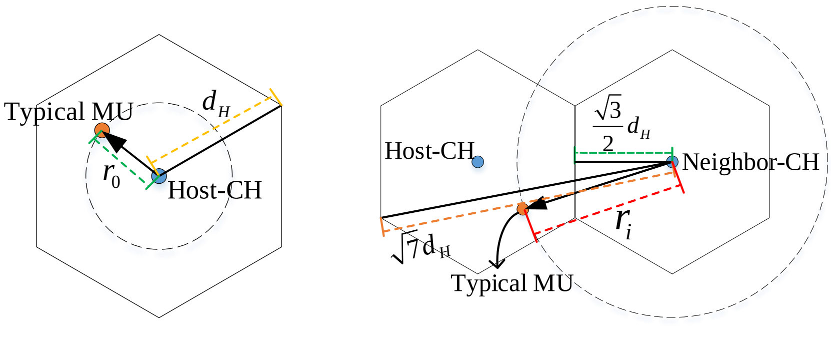
$$f_{r_0}(x) = \frac{2\pi\lambda_U x e^{-\pi\lambda_U x^2}}{1 - e^{-\pi\lambda_U {d_H}^2}}, 0 \leq x \leq d_H \tag{A2}$$

As shown in Figure A1b, the typical MU is located in the common area of the representative microcell and the circular area of radius $r_i, 1 \leq i \leq 6$ centered at a neighbor-CH, which means that there exists at least one MU in the circular area of a radius r_i centered on a neighbor-CH. We also have

$$\mathbb{P}(r_i \leq x) = 1 - \mathbb{P}(0, x, \lambda_U) = 1 - e^{-\pi\lambda_U x^2} \tag{A3}$$

The neighbor-CH serves only the MUs in the annulus area of a radius from $\frac{\sqrt{3}}{2}d_H$ to $\sqrt{7}d_H$ centered on itself. The PDF of $r_i, 1 \leq i \leq 6$ is

$$f_{r_i}(x) = \frac{2\pi\lambda_U x e^{-\pi\lambda_U x^2}}{e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2}}, \frac{\sqrt{3}}{2}d_H \leq x \leq \sqrt{7}d_H \tag{A4}$$



(**a**) (**b**)

Figure A1. Illustration of the distance between typical MU and CH. (**a**) The distance between typical MU and host-CH. (**b**) The distance between typical MU and a neighbor-CH.

Appendix B. Proof of Lemma 2

Before the derivation of $\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i)$, we must first derive the PDF of the signal $S_i, 0 \leq i \leq 6$ (or interference $I_i, 1 \leq i \leq 6$) strength. Because the PDF of S_i is equal to I_i, we only derive S_i condition on $r_i, 0 \leq i \leq 6$,

$$\mathbb{P}(S_i \leq x | r_i) = \mathbb{P}(h_i {r_i}^{-\beta} \leq x | r_i) = 1 - e^{-{r_i}^{\beta} x} \tag{A5}$$

By derivation, the PDF of S_i is

$$f_{S_i}(x|r_i) = {r_i}^{\beta} e^{-{r_i}^{\beta} x}, x \geq 0 \tag{A6}$$

When file block $fb_i, N_U + 1 \leq i \leq L_m$ is transmitted by Host-CH-provide, the probability that $SINR_{HCH}$ exceeds the certain SINR threshold Γ is

$$\begin{aligned}
\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i) &= \mathbb{P}(\frac{h_0 {r_0}^{-\beta}}{\sum\limits_{i=1}^{6} A_i I_i} \geq \Gamma) \\
&= \mathbb{P}(h_0 \geq \Gamma {r_0}^{\beta} \sum_{i=1}^{6} A_i I_i) \\
&= \mathbb{E}_{r_0} \mathbb{E}_{\{A_i I_i, 1 \leq i \leq 6\}} (e^{-\Gamma {r_0}^{\beta} \sum\limits_{i=1}^{6} A_i I_i}) \\
&= \mathbb{E}_{r_0} \prod_{i=1}^{6} \mathbb{E}_{I_i} (\mu e^{-\Gamma {r_0}^{\beta} I_i} + 1 - \mu) \\
&= \mathbb{E}_{\{r_i, 0 \leq i \leq 6\}} \prod_{i=1}^{6} (\int\limits_{0}^{+\infty} (\mu e^{-\Gamma {r_0}^{\beta} I_i} + 1 - \mu) f_{S_i}(I_i | r_i) dI_i) \\
&= \int\limits_{r_0} [\int\limits_{r_1} (1 - \frac{\mu}{1 + {r_1}^{\beta} / (\Gamma {r_0}^{\beta})}) f_{r_1}(r_1) dr_1]^6 f_{r_0}(r_0) dr_0 \\
&\overset{(a)}{=} \int\limits_{0}^{d_H} [\int\limits_{\frac{\sqrt{3}}{2} d_H}^{\sqrt{7} d_H} \psi(r_0, r_1) f_{r_1}(r_1) dr_1]^6 f_{r_0}(r_0) dr_0 \\
&= \frac{d_H}{2} \int\limits_{-1}^{1} [\int\limits_{\frac{\sqrt{3}}{2} d_H}^{\sqrt{7} d_H} \psi(\frac{d_H}{2} x + \frac{d_H}{2}, r_1) f_{r_1}(r_1) dr_1]^6 f_{r_0}(\frac{d_H}{2} x + \frac{d_H}{2}) dx \\
&\overset{(b)}{=} \frac{d_H}{2} \sum_{j=1}^{N} \omega_j [\int\limits_{\frac{\sqrt{3}}{2} d_H}^{\sqrt{7} d_H} \psi(\frac{d_H}{2} x_j + \frac{d_H}{2}, r_1) f_{r_1}(r_1) dr_1]^6 f_{r_0}(\frac{d_H}{2} x_j + \frac{d_H}{2})
\end{aligned} \tag{A7}$$

where $\mathbb{E}_a(b)$ is the mean of b conditioning on a, N is the number of quadrature nodes, and x_j is the j-th root of the Legendre polynomials as follows

$$P_N(x) = \frac{1}{2^N}\sum_{i=0}^{N}\binom{N}{i}^2 (x-1)^{N-i}(x+1)^i \tag{A8}$$

Additionally, ω_j is the weight coefficient and can be obtained by the following [27].

$$\omega_j = \frac{2(1-{x_j}^2)}{(N+1)^2(P_{N+1}(x_j))^2} \tag{A9}$$

In step(a), we denote $\psi(r_0, r_1) = 1 - \frac{\mu}{1+{r_1}^{\beta}/(\Gamma {r_0}^{\beta})}$. Because $\lim\limits_{r_0\to\infty}[\int\limits_{\frac{\sqrt{3}}{2}d_H}^{\sqrt{7}d_H} \psi(r_0, r_1) f_{r_1}(r_1) dr_1]^6 f_{r_0}(r_0) = 0$, the Gauss-Legendre quadrature method can converge to the correct value much faster and with fewer nodes than the Gauss-Laguerre quadrature method [28]. Based on the process of Gauss-Legendre quadrature, we have step(b). Incorporating (12), (13) into (A7), we have

$$\begin{aligned}\mathbb{P}(SINR_{HCH} \geq \Gamma | fb_i) = &\frac{d_H}{2(e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2})^6 (1-e^{-\pi\lambda_U {d_H}^2})} \sum_{j=1}^{N} \omega_j [e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2} \\ &- 2\pi\mu\lambda_U \int\limits_{\frac{\sqrt{3}}{2}d_H}^{\sqrt{7}d_H} \frac{r_1 e^{-\pi\lambda_U {r_1}^2}}{1+{r_1}^{\beta}/(\Gamma(\frac{d_H}{2}x_j + \frac{d_H}{2})^{\beta})} dr_1]^6 2\pi\lambda_U(\frac{d_H}{2}x_j + \frac{d_H}{2}) e^{-\pi\lambda_U(\frac{d_H}{2}x_j + \frac{d_H}{2})^2} \\ &= \eta \sum_{j=1}^{N} \omega_j \varphi(x_j)\end{aligned} \tag{A10}$$

where

$$\eta = \frac{d_H}{2(e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2})^6 (1-e^{-\pi\lambda_U {d_H}^2})}$$

$$\begin{aligned}\varphi(x) = &2\pi\lambda_U(\frac{d_H}{2}x + \frac{d_H}{2}) e^{-\pi\lambda_U(\frac{d_H}{2}x + \frac{d_H}{2})^2} \\ &\times [e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2} - 2\pi\mu\lambda_U \int\limits_{\frac{\sqrt{3}}{2}d_H}^{\sqrt{7}d_H} \frac{r_1 e^{-\pi\lambda_U {r_1}^2}}{1+{r_1}^{\beta}/(\Gamma(\frac{d_H}{2}x + \frac{d_H}{2})^{\beta})} dr_1]^6\end{aligned} \tag{A11}$$

Appendix C. Proof of Lemma 3

When the file block $fb_i, N_U+1 \leq i \leq L_m$ is transmitted by Neighbor-CH-provide, the probability that $SINR_{NCH}$ exceeds the certain SINR threshold Γ is

$$\begin{aligned}
\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i) &\overset{(a)}{=} \mathbb{P}(\mathbb{E}(SINR_{NCH}) \geq \Gamma | fb_i) \\
&= \mathbb{P}(\mathbb{E}(\frac{\sum\limits_{i=1}^{6} B_i S_i}{\sum\limits_{j=1}^{6} C_j I_j}) \geq \Gamma) \\
&\overset{(b)}{=} \mathbb{P}(\frac{\mathbb{E}_{\{B_i S_i, 1 \leq i \leq 6\}}(\sum\limits_{i=1}^{6} B_i S_i)}{\mathbb{E}_{\{C_j I_j, 1 \leq i \leq 6\}}(\sum\limits_{j=1}^{6} C_j I_j)} \geq \Gamma) \\
&= \mathbb{P}(\frac{(1-\mu) q_i}{\mu(1-q_i)} \geq \Gamma) \\
&= \begin{cases} 1, q_i \geq \dfrac{\mu\Gamma}{1-\mu+\mu\Gamma} \\ 0, q_i < \dfrac{\mu\Gamma}{1-\mu+\mu\Gamma} \end{cases}
\end{aligned} \tag{A12}$$

where step (a) is a relaxation process meaning that when the average SINR at typical MU exceeds Γ, the transmission of fb_i by Neighbor-CH-provide is successful. In step (b), we assume the correlation between $S_i, 1 \leq i \leq 6$ is weak.

Incorporating (11) into (A12), the formula $q_i \geq \frac{\mu\Gamma}{1-\mu+\mu\Gamma}$ becomes

$$\frac{N_H (i-N_U)^{-\theta}}{\sum\limits_{j=1}^{L_m-N_U} j^{-\theta}} \geq \frac{\mu\Gamma}{1-\mu+\mu\Gamma}$$

$$i \leq (\frac{\mu\Gamma \sum\limits_{j=1}^{L_m-N_U} j^{-\theta}}{(1-\mu+\mu\Gamma)N_H})^{-\frac{1}{\theta}} + N_U \tag{A13}$$

Let $\varepsilon = \min(\left\lfloor (\frac{\mu\Gamma \sum\limits_{j=1}^{L_m-N_U} j^{-\theta}}{(1-\mu+\mu\Gamma)N_H})^{-\frac{1}{\theta}} \right\rfloor + N_U, L_m)$, where $\lfloor * \rfloor$ is an operation whose result is the maximal integer below $*$. The $\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i)$ is

$$\mathbb{P}(SINR_{NCH} \geq \Gamma | fb_i) = \begin{cases} 1, N_U+1 \leq i \leq \varepsilon \\ 0, \varepsilon < i \leq L_m \end{cases} \tag{A14}$$

Appendix D. Proof of Lemma 4

When the file block $fb_i, N_U + 1 \leq i \leq L_m$ is transmitted by Host-CH-provide, the average SINR experienced by the typical MU is

$$\begin{aligned}
\mathbb{E}(SINR_{HCH}) &= \mathbb{E}(\frac{S_0}{\sum\limits_{i=1}^{6} A_i I_i}) \\
&\overset{(a)}{=} \frac{\mathbb{E}_{s_0}(S_0)}{\mathbb{E}_{\{A_i I_i, 1 \leq i \leq 6\}}(\sum\limits_{i=1}^{6} A_i I_i)} \\
&= \frac{\mathbb{E}_{S_0}(S_0)}{6\mu \mathbb{E}_{\{I_i, 1 \leq i \leq 6\}}(I_1)} \\
&= \frac{\int\limits_{0}^{d_H} \int\limits_{0}^{+\infty} S_0 f_{S_0}(S_0|r_0) dS_0 f_{r_0}(r_0) dr_0}{6\mu \int\limits_{\frac{\sqrt{3}}{2} d_H}^{\sqrt{7} d_H} \int\limits_{0}^{+\infty} I_1 f_{I_1}(I_1|r_1) dI_1 f_{r_1}(r_1) dr_1}
\end{aligned} \tag{A15}$$

In step(a), we assume the correlation between S_0 and $I_i, 1 \leq i \leq 6$ is weak. Incorporating (12), (13), (A6) into (A15) and according to Equation (3.381) in [29], we have

$$\mathbb{E}(SINR_{HCH}) = \frac{(e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2})}{6\mu(1 - e^{-\pi\lambda_U {d_H}^2})} \frac{\gamma(\frac{2-\beta}{2}, \pi\lambda_U {d_H}^2)}{\gamma(\frac{2-\beta}{2}, 7\pi\lambda_U {d_H}^2) - \gamma(\frac{2-\beta}{2}, \frac{3}{4}\pi\lambda_U {d_H}^2)} \tag{A16}$$

where $\gamma(a, b) = \int\limits_{0}^{b} x^{a-1} e^{-x} dx$ is the incomplete gamma function. According to the Shannon equation, the average transmission rate is

$$R_{HCH} = W \log(\frac{(e^{-\frac{3}{4}\pi\lambda_U {d_H}^2} - e^{-7\pi\lambda_U {d_H}^2})}{6\mu(1 - e^{-\pi\lambda_U {d_H}^2})} \frac{\gamma(\frac{2-\beta}{2}, \pi\lambda_U {d_H}^2)}{\gamma(\frac{2-\beta}{2}, 7\pi\lambda_U {d_H}^2) - \gamma(\frac{2-\beta}{2}, \frac{3}{4}\pi\lambda_U {d_H}^2)} + 1) \tag{A17}$$

References

1. CISCO. CISCO Visual Networking Index: Global Mobile Data Traffic Forcast Update, 2015–2020. Available online: https://www.cisco.com/c/dam/m/en_in/innovation/enterprise/assets/mobile-white-paper-c11-520862.pdf (accessed on 14 October 2019).
2. Deng, N.; Haenggi, M. The benefits of hybrid caching in Gauss–Poisson D2D networks. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 1217–1230. [CrossRef]
3. Bąstug, E.; Bennis, M.; Debbah, M. Living on the edge: The role of proactive caching in 5G wireless networks. *IEEE Commun. Mag.* **2014**, *52*, 82–89. [CrossRef]
4. Wang, X.; Chen, M.; Taleb, T.; Ksentini, A.; Leung, V. Cache in the air: Exploiting content caching and delivery techniques for 5G systems. *IEEE Trans. Commun. Mag.* **2014**, *52*, 131–139. [CrossRef]
5. Lei, L.; Xiong, X.; Hou, L.; Zheng, K. Collaborative edge caching through service function chaining: Architecture and challenges. *IEEE Wirel. Commun.* **2018**, *25*, 94–102. [CrossRef]
6. Song, J.; Song, H.; Choi, W. Optimal caching placement of caching system with helpers. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 1825–1830.
7. Blaszczyszyn, B.; Giovanidis, A. Optimal geographic caching in cellular networks. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015.
8. Bastug, E.; Bennis, M.; Debbah, M. Cache-enabled small cell networks: Modeling and tradeoffs. In Proceedings of the 2014 11th International Symposium on Wireless Communications Systems (ISWCS), Barcelona, Spain, 26–29 August 2014; pp. 649–653.

9. Chae, S.H.; Ryu, J.Y.; Quek, T.Q.S.; Choi, W. Cooperative transmission via caching helpers. In Proceedings of the 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, 6–10 December 2015.
10. Li, J.; Chen, Y.; Lin, Z.; Chen, W.; Vucetic, B.; Hanzo, L. Distributed caching for data dissemination in the downlink of heterogeneous networks. *IEEE Trans. Commun.* **2015**, *63*, 3553–3568. [CrossRef]
11. Shanmugam, K.; Golrezaei, N.; Dimakis, A.G.; Molisch, A.F.; Caire, G. Femtocaching: Wireless content delivery through distributed caching helpers. *IEEE Trans. Inform. Theory* **2012**, *59*, 8402–8413. [CrossRef]
12. Chen, Z.; Lee, J.; Quek, T.Q.S.; Kountouris, M. Cooperative caching and transmission design in cluster-centric small cell networks. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 3401–3415. [CrossRef]
13. Afshang, M.; Dhillon, H.S.; Chong, P.H.J. Modeling and performance analysis of clustered device-to-device networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 4957–4972. [CrossRef]
14. Golrezaei, N.; Mansourifard, P.; Molisch, A.F.; Dimakis, A.G. Base-station assisted device-to-device communications for high-throughput wireless video networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 3665–3676. [CrossRef]
15. Chen, Z.; Pappas, N.; Kountouris, M. Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal. *IEEE Commun. Lett.* **2017**, *21*, 584587. [CrossRef]
16. Golrezaei, N.; Dimakis, A.G.; Molisch, A.F. Scaling behavior fo device-to-device communications with distributed caching. *IEEE Trans. Inform. Theory* **2014**, *60*, 4286–4298. [CrossRef]
17. Amer, R.; Butt, M.M.; Bennis, M.; Marchetti, N. Inter-Cluster Cooperation for Wireless D2D Caching Networks. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 6108–6121. [CrossRef]
18. Yang, C.; Yao, Y.; Chen, Z.; Xia, B. Analysis on cache-enabled wireless heterogeneous networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 131–145. [CrossRef]
19. Wang, K.; Chen, Z.; Liu, H. Push-based wireless converged networks for massive multimedia content delivery. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 2894–2905.
20. Wang, R.; Hajiaghajani, F.; Biswas, S. Incentive Based Cooperative Content Caching in Social Wireless Networks. In Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, 8–13 October 2017.
21. Jiang, Y.; Ma, M.; Bennis, M.; Zheng, F.; You, X. User Preference Learning Based Edge Caching for Fog Radio Access Network. *IEEE Trans. Commun.* **2019**, *67*, 1268–1283. [CrossRef]
22. Peng, A.; Jiang, Y.; Bennis, M.; Zheng, F.C.; You, X. Performance analysis and caching design in fog radio access networks.In Proceedings of the IEEE Globecom Workshops, Abu Dhabi, UAE, 9–13 December 2018; pp. 1–6.
23. Baek, J.; Kaddoum, G.; Garg, S.; Kaur, K.; Gravel, V. Managing Fog Networks using Reinforcement Learning Based Load Balancing Algorithm. In Proceedings of the IEEE 2019 IEEE WCNC, Marrakech, Morocco, 15–19 April 2019.
24. Wang, R.; Li, R.; Wang, P.; Liu, E. Analysis and Optimization of Caching in Fog Radio Access Networks. *IEEE Trans. Veh. Technol.* **2019**, *68*, 8279–8283. [CrossRef]
25. Rao, J.; Feng, H.; Yang, C.; Chen, Z.; Xia, B. Optimal caching placement for D2D assisted wireless caching networks. In Proceedings of the IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 22–27 May 2016.
26. Zhuang, Y.; Luo, Y.; Cai, L.; Pan, J. A geometric probability model for capacity analysis and interference estimation in wireless mobile cellular systems. In Proceedings of the Global Telecommunications Conference 2011 (GLOBECOM 2011), Kathmandu, Nepal, 5–9 December 2011.
27. Zhang, Y.; Wang, X.; Wang, D.; Zhang, Y.; Lan, Y. BER Performance of Multicast SCMA Systems. *IEEE Wirel. Commun. Lett.* **2018**, *8*, 1073–1076. [CrossRef]
28. Cohen, H. *Numerical Approximation Methods*; Springer: New York, NY, USA, 2011.
29. Gradshteyn, I.S.; Ryzhik, I.M. *Table of Integrals, Series and Products*, 6th ed.; Academic Press: New York, NY, USA, 2000.