Dynamic Popularity Driven Caching Optimization at Base Station

Kaiqiang Qi, Shengqian Han and Chenyang Yang
School of Electronics and Information Engineering, Beihang University, Beijing 100191, China
Email: {kaiqiangqi, sqhan, cyyang}@buaa.edu.cn

***Abstract*—Caching at base station (BS) has attracted significant research efforts for future wireless networks. Most existing works are based on the assumption of static content catalogue and stationary popularity distribution, which however is far away from the reality as recently reported in the literature. In this paper, we take the popularity dynamics into account, and study the caching policy at BS for a traffic model consisting of two categories of contents respectively with long and short lifespans. By modeling the two categories of contents with Independent Reference Model (IRM) and Shot Noise Model (SNM), we formulate a cache resource allocation problem to maximize the total cache hit ratio for both categories of contents, which gives rise to a hybrid proactive and reactive caching policy. We solve the problem numerically for general case and provide closed-form solutions for several special cases. Numerical and simulation results demonstrate remarkable performance gain of the proposed caching policy over non-hybrid caching polices.**

I. Introduction

Caching at base station (BS) is a promising way to address the explosive growth of mobile data demand in the fifth-generation cellular systems [1,2]. Aimed at maximizing cache hit ratio (CHR), network throughput or energy efficiency, the optimization of caching policies has drawn significant attention in the literature [1–3].

Most existing works are based on the Independent Reference Model (IRM), which assumes static content catalogue and stationary popularity distribution. Although widely used and easy for optimization and analysis, IRM has been shown insufficient to characterize the real-world arrival process of content requests. In [4], a novel traffic model, Shot Noise Model (SNM), was proposed to describe the temporal locality between requests for a content and the dynamics of content catalogue as well as popularity distribution. Under SNM, the requests for a content arrive almost within a so-called lifespan, beyond which the content is rarely requested. There exist appropriate caching policies for the contents with different lifespans. For instance, it is common to adopt Least Recently Used (LRU) policy when caching the contents with short lifespan [4,5], while popular caching policy (i.e., caching the most popular contents) is more frequently used when caching contents with long lifespan [1, 3]. Therefore, if the content lifespan is known *a priori*, one can select the appropriate caching policy for each content.

This work is supported by National Natural Science Foundation of China (NSFC) under Grant 61671036.

The prediction of content lifespan has been studied, e.g., in [6–8]. It was pointed out in [6] that predicting the likelihood of a content's lifespan longer than a threshold is possible. In [7], the lifespan of the contents at Tencent, one of the largest video-on-demand service provider in China, was analyzed, and the results indicate that News or Sports videos are age-sensitive with short lifespan, but Movie and MV videos have long lifespan. In [8], a lifespan model based on the measurement of the characteristics of YouTube videos was proposed, which provides a possible approach to predict the content lifespan.

In this paper, we study the caching optimization for a practical scenario where the contents requested by the users in a cell are from two categories of contents, which can be modeled by IRM and SNM, respectively. A hybrid caching policy with both popular caching policy and LRU is considered at the BS. Then, given the total cache size of the BS, we formulate a caching resource allocation problem for the two categories of contents to optimize the hybrid policy, aimed at maximizing the total CHR. We provide a numerical solution to the problem under the small-cache scenario, which is relevant in wireless caching networks. In order to gain useful insights, we further derive closed-form solutions in three special cases. Numerical and simulation results validate our analysis, and demonstrate the evident gain of the proposed caching policy over LRU and popular caching policy even with inaccurate knowledge of content lifespan.

II. Traffic Model

Consider a hybrid traffic model consisting of two categories of contents with long and short lifespans, which are characterized by IRM and SNM models, respectively.

IRM describes a stationary request process, which assumes that the content catalogue, $N_{\mathrm{f}}^{\mathrm{I}}$, is constant, meanwhile the content popularity is stationary. As widely considered in the literature, e.g., [9], we assume that the request probability of the i-th most popular content, q_i, obeys Zipf distribution with parameter δ, i.e., $q_i = \frac{i^{-\delta}}{\sum_{j=1}^{N_{\mathrm{f}}^{\mathrm{I}}} j^{-\delta}}$, where $0<\delta<1$ holds generally according to numerous experimental studies [9]. Let μ_0 denote the average request arrival rate for IRM.

SNM is a dynamic request model, which can be characterized by content arrival process and content request process. According to [4], the content arrival process is assumed as a homogeneous Poisson process with an average arrival rate λ, while the request process of each content is assumed as

an inhomogeneous Poisson process. Specifically, for the m-th content, its request process is captured by the following four features: 1) arrival time t_m, 2) popularity profile $\Lambda_m(t)$, satisfying $\Lambda_m(t) \geq 0$ and $\int_0^\infty \Lambda_m(\tau)\mathrm{d}\tau = 1$, 3) lifespan T_m, defined as $T_m = \frac{1}{\int_0^\infty \Lambda_m^2(t)\mathrm{d}t}$, and 4) the volume of requests V_m. In [4], V_m is modeled as an independent identically distributed (i.i.d.) random variable following Pareto distribution with parameters of $\beta > 1$ and $V_{\min}$, and the probability density function is $f_V(v) = \frac{\beta V_{\min}^\beta}{v^{\beta+1}}$, $v \geq V_{\min}$. One can find that the SNM model leads to a time-varying request arrival rate for each content, and the instantaneous request arrival rate of the m-th content at time t can be obtained as $V_m\Lambda_m(t - t_m)$.

The analysis in [4] indicates that the shape of popularity profile has little impact on the performance of LRU caching policy and the performance essentially depends only on the average lifespan. Therefore, a simplified SNM model has been employed in the literature, e.g., in [10], where all the contents in the SNM category have the same lifespan T and a rectangular popularity profile is considered with $\Lambda(t) = \frac{1}{T}$, $t \in [0, T]$. Given the simplified SNM model, we can obtain the average number of active contents, i.e., the average number of arrival contents within the lifespan, as $N_{\mathrm{f}}^{\mathrm{S}} = \lambda T$, which can be regarded as the content catalogue of SNM. The impact of such simplification will be evaluated by simulations later.

III. Cache Resource Allocation

Appropriate caching policies for the two categories of contents differ. For IRM category, which can model contents with long lifespan and hence content popularity can be accurately predicted, the proactive popular caching policy is optimal if one user can only associate with one BS, which caches the most popular contents at the BS [1]. For SNM category, which can model contents with short lifespan, the reactive LRU caching policy is often employed, e.g., in [4]. As a result, when a BS has the requests of both categories of contents, the BS should apply both the popular caching policy and LRU policy. This naturally leads to a hybrid proactive and reactive caching policy: allocating a fraction of the cache resource to cache the contents in IRM category with popular caching policy and the rest resource to cache the contents in SNM category with LRU.

In this section, we design the hybrid caching policy by optimizing the fraction of cache resource allocated to each category. We formulate and solve the cache resource allocation problem, aimed at maximizing the total CHR for both categories of contents. To gain useful insights, we further derive closed-form solutions in three special cases. For simplicity, in the sequel we call the contents in IRM and SNM categories IRM contents and SNM contents, respectively.

A. Problem Formulation

Consider a single-cell wireless caching system, where the BS has the cache size of N_{c} contents and multiple uniformly located users request contents from the BS. If the requested content of a user is cached, the BS will fetch the content from its local cache directly and transmit to the user. Otherwise, the BS will fetch the content from core network via backhaul link. Denote the fraction of cache resource allocated to cache SNM contents as η, then the cache sizes for SNM and IRM contents are ηN_{c} and $(1-\eta)N_{\mathrm{c}}$, respectively, where ηN_{c} is an integer.

Let Q^{S} and Q^{I} denote the numbers of received requests for SNM and IRM contents during an evaluation period T_0, respectively. Then, the total CHR of the BS, defined as the ratio of the average number of requests for the cached contents to that for all contents, can be expressed as

$$\begin{aligned}\bar{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta) &= \frac{\mathbb{E}\left[Q^{\mathrm{S}}\right]\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta) + \mathbb{E}\left[Q^{\mathrm{I}}\right]\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta)}{\mathbb{E}\left[Q^{\mathrm{S}} + Q^{\mathrm{I}}\right]} \\ &\triangleq w^{\mathrm{S}}\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta) + w^{\mathrm{I}}\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta),\end{aligned} \tag{1}$$

where $\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta)$ and $\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta)$ are the average CHRs of SNM contents and IRM contents, respectively, $w^{\mathrm{S}} = \frac{\mathbb{E}\left[Q^{\mathrm{S}}\right]}{\mathbb{E}[Q^{\mathrm{S}}+Q^{\mathrm{I}}]}$ is the fraction of requests for SNM contents, and $w^{\mathrm{I}} = \frac{\mathbb{E}\left[Q^{\mathrm{I}}\right]}{\mathbb{E}[Q^{\mathrm{S}}+Q^{\mathrm{I}}]}$ is the fraction of requests for IRM contents. It is clear that $w^{\mathrm{S}} + w^{\mathrm{I}} = 1$.

For SNM contents, the number of requested contents during the evaluation time T_0, denoted by n_0, is a random variable, which obeys Poisson distribution with the average number λT_0. Then, the average number of requests for SNM contents can be derived by taking the expectation over both n_0 and the request volume of each content V_m as $\mathbb{E}[Q^{\mathrm{S}}] = \mathbb{E}\left\{\mathbb{E}[\sum_{m=1}^{n} V_m | n = n_0]\right\} = \mathbb{E}\left\{n_0\mathbb{E}[V_m]\right\} = \lambda T_0\mathbb{E}[V_m]$.

For IRM contents, the total number of requests received by $N_{\mathrm{f}}^{\mathrm{I}}$ contents during T_0, i.e., Q^{I}, has the expectation $\mathbb{E}[Q^{\mathrm{I}}] = \mu_0 T_0 \triangleq \bar{\mu}N_{\mathrm{f}}^{\mathrm{I}}T_0$. Herein, we introduce an auxiliary variable $\bar{\mu} = \frac{\mu_0}{N_{\mathrm{f}}^{\mathrm{I}}}$ to denote the average request arrival rate for each single IRM content, which is used for performance analysis later.

Then, upon substituting $\mathbb{E}[Q^{\mathrm{S}}]$ and $\mathbb{E}[Q^{\mathrm{I}}]$ into w^{S} and recalling that $N_{\mathrm{f}}^{\mathrm{S}} = \lambda T$, we can obtain $w^{\mathrm{S}} = \frac{\mathbb{E}[V_m]N_{\mathrm{f}}^{\mathrm{S}}}{\mathbb{E}[V_m]N_{\mathrm{f}}^{\mathrm{S}} + \bar{\mu}TN_{\mathrm{f}}^{\mathrm{I}}}$.

Further denoting k_μ as the ratio between the average request arrival rate of each IRM content, $\bar{\mu}$, and that of each SNM content, $\frac{\mathbb{E}[V_m]}{T}$, we can rewrite w^{S} as

$$w^{\mathrm{S}} = \frac{N_{\mathrm{f}}^{\mathrm{S}}}{N_{\mathrm{f}}^{\mathrm{S}} + k_\mu N_{\mathrm{f}}^{\mathrm{I}}}. \tag{2}$$

Based on Che's approximation, the average CHR of LRU policy under SNM can be approximated as [4]

$$\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta) \approx 1 - \int_0^\infty \Lambda(\tau)\frac{\phi_V'\left(-\int_0^{T_{\mathrm{c}}}\Lambda(\tau-\theta)\mathrm{d}\theta\right)}{\mathbb{E}[V_m]}\mathrm{d}\tau, \tag{3}$$

where $\phi_V'(x) = \mathbb{E}\left[V_m e^{xV_m}\right]$, T_{c} is the cache eviction time, representing the duration from the time when a content enters the cache until the content is evicted, which is the unique solution of the following equation [4]

$$\eta N_{\mathrm{c}} = \lambda\int_0^\infty 1 - \phi_V\left(-\int_0^{T_{\mathrm{c}}}\Lambda(\tau-\theta)\mathrm{d}\theta\right)\mathrm{d}\tau, \tag{4}$$

where $\phi_V(x) = \mathbb{E}\left[e^{xV_m}\right]$.

For IRM contents, the average CHR of popular caching

policy can be expressed as [9]

$$\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta) = \sum_{i=1}^{(1-\eta)N_{\mathrm{c}}} q_i = \frac{\sum_{i=1}^{(1-\eta)N_{\mathrm{c}}} i^{-\delta}}{\sum_{j=1}^{N_{\mathrm{f}}^{\mathrm{I}}} j^{-\delta}}. \tag{5}$$

By substituting (2), (3) and (5) into (1), we can finally formulate the optimization problem of cache resource allocation to maximize the total CHR at the BS, averaged over all possibly requested contents and stochastic request process, as

$$\begin{aligned} \max_{\eta} \quad & \bar{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta) \\ s.t. \quad & 0 \le \eta \le 1,\ \eta N_{\mathrm{c}} \in \mathbb{Z}. \end{aligned} \tag{6}$$

B. Optimization of Cache Resource Allocation

To solve problem (6), we first find closed-form expressions of $\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta)$ and $\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta)$ from (3) and (5) by introducing approximations.

For SNM contents, we resort to the approximation under the small-cache scenario, which is relevant in wireless caching networks where the cache size of a BS is typically far smaller than the content catalogue, i.e., $N_{\mathrm{c}} \ll \min\left\{N_{\mathrm{f}}^{\mathrm{S}}, N_{\mathrm{f}}^{\mathrm{I}}\right\}$. The approximated CHR is given in the following proposition.

Proposition 1. The average CHR for SNM contents under the small-cache scenario, i.e., $\frac{N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{S}}} \approx 0$, can be approximated as

$$\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta) \approx 1 - \frac{1}{\mathbb{E}[V_m]}\mathbb{E}\left[V_m \int_0^{\infty} \Lambda(\tau) e^{-\frac{V_m \Lambda(\tau)}{\lambda \mathbb{E}[V_m]}\eta N_{\mathrm{c}}} \mathrm{d}\tau\right]. \tag{7}$$

The proof is given in [11] due to lack of space.

For IRM contents, we approximate $\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta)$ by converting the summation in (5) into integration. Specifically, we approximate $\sum_{i=1}^{k} i^{-\delta}$ with $\int_0^k x^{-\delta}\mathrm{d}x = \frac{k^{1-\delta}}{1-\delta}$ for $\delta \in (0,1)$, and obtain

$$\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta) \approx \left[\frac{(1-\eta)N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{I}}}\right]^{1-\delta}. \tag{8}$$

Then, considering $\Lambda(\tau) = \frac{1}{T}, \tau \in [0,T]$, and substituting (7) and (8) into (1), we have $\bar{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta) \approx w^{\mathrm{S}}\left\{1 - \frac{1}{\mathbb{E}[V_m]}\mathbb{E}\left[V_m e^{-\frac{V_m}{N_{\mathrm{f}}^{\mathrm{S}}\mathbb{E}[V_m]}\eta N_{\mathrm{c}}}\right]\right\} + w^{\mathrm{I}}\left[\frac{(1-\eta)N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{I}}}\right]^{1-\delta} \triangleq \hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$.

It can be proved that $\hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$ is a concave function of η by examining its second derivative. Thus, the optimal value of η that maximizes $\hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$ under the integer constraint $\eta N_{\mathrm{c}} \in \mathbb{Z}$ can be found in two steps: first omit the integer constraint and find the optimal solution to the relaxed problem, denoted by η^*, and then compare $\hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$ corresponding to the ceiling and floor of $\eta^* N_{\mathrm{c}}$ and choose the one with larger $\hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$ as the final solution.

The optimal solution without the integer constraint can be derived from the Karush-Kuhn-Tucker conditions [12] of problem (6) by replacing $\bar{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$ with $\hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)$ as $\eta^* = \max\{\eta_0, 0\}$, where η_0 is the solution of the equation $\frac{\mathrm{d}\hat{p}_{\mathrm{h}}^{\mathrm{tot}}(\eta)}{\mathrm{d}\eta} = 0$, which can be expressed as

$$\begin{aligned} &\frac{w^{\mathrm{S}} N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{S}}\mathbb{E}^2[V_m]}\mathbb{E}\left[V_m^2 e^{-\frac{V_m}{N_{\mathrm{f}}^{\mathrm{S}}\mathbb{E}[V_m]}\eta N_{\mathrm{c}}}\right] \\ &\qquad - w^{\mathrm{I}}(1-\delta)\left(\frac{N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{I}}}\right)^{1-\delta}(1-\eta)^{-\delta} = 0. \end{aligned} \tag{9}$$

It is easy to observe that the left-hand side of (9) is a monotonically decreasing function of η. Thus, we can readily find the unique solution of η_0 by, e.g., bisection search.

We also find that $\eta < 1$ must be satisfied in (9) under $0 < \delta < 1$, hence, $\eta_0 = 1$ will never happen, which indicates that only caching SNM contents at BS is not optimal.

C. Special-case Analysis and Insight

In order to gain useful insights, in what follows, we derive closed-form solutions of (9) under the small-cache scenario in three special cases.

We start with a brief discussion on the popularity skewness of SNM contents. Since the request volume of each SNM content follows i.i.d. Pareto distribution with parameter β, it is proved in [13] that the sorted request volumes (i.e., content popularity) in descending order approximately obey Zipf distribution with parameter $\delta^{\mathrm{I}} = \beta^{-1}$ for a large content catalogue. For notational simplicity, in this subsection we consider the case where SNM and IRM contents have the same popularity skewness, i.e., $\delta^{\mathrm{I}} = \delta$. The analysis for arbitrary δ^{I} and δ is similar.

By substituting $\delta = \beta^{-1}$ and (2) into (9), we obtain that

$$\underbrace{\frac{\beta}{k_\mu(\beta-1)\mathbb{E}^2[V_m]}\mathbb{E}\left[V_m^2 e^{-\frac{V_m}{N_{\mathrm{f}}^{\mathrm{S}}\mathbb{E}[V_m]}\eta N_{\mathrm{c}}}\right]}_{g_l(\eta)} = \underbrace{\left[\frac{(1-\eta)N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{I}}}\right]^{-\frac{1}{\beta}}}_{g_r(\eta)}, \tag{10}$$

where the left- and right-hand sides are denoted by $g_l(\eta)$ and $g_r(\eta)$, respectively.

Proposition 2. Under the small-cache scenario, i.e., $\frac{N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{S}}} \approx 0$, the solution of equation (10) can be approximated as

$$\eta_0 \approx \begin{cases} h^{-1}(KN_{\mathrm{c}}), & \text{if } 1 < \beta < 2, \\ 1 - \left(k_\mu \frac{\beta-2}{\beta-1}\right)^{\beta} \frac{N_{\mathrm{f}}^{\mathrm{I}}}{N_{\mathrm{c}}}, & \text{if } \beta > 2, \end{cases} \tag{11}$$

where $K = \left[\frac{(A_\beta)^\beta \left(N_{\mathrm{f}}^{\mathrm{S}}\right)^{(2-\beta)\beta}}{k_\mu^\beta N_{\mathrm{f}}^{\mathrm{I}}}\right]^{\frac{1}{(\beta-1)^2}} > 0$ for $1 < \beta < 2$, $A_\beta = \beta\left(\frac{\beta-1}{\beta}\right)^{\beta-1}\Gamma(2-\beta)$, $h(x) = \frac{x^{\frac{(2-\beta)\beta}{(\beta-1)^2}}}{(1-x)^{\frac{1}{(\beta-1)^2}}}$, and $h^{-1}(\cdot)$ is the inverse function of $h(\cdot)$. Apparently, $h^{-1}(KN_{\mathrm{c}})$ monotonically increases with N_{c}. The proof is omitted due to lack of space (refer to [11] for details).

In the following, we solve equation (10) in closed form for three special cases.

1) Case 1: $\beta \approx 1$, i.e., the popularity distribution is quite skewed with $\delta^{\mathrm{I}} = \delta \approx 1$. In this case, $g_l(\eta)$ can be approximated by $\frac{N_{\mathrm{f}}^{\mathrm{S}}}{k_\mu \eta N_{\mathrm{c}}}$, then the solution to equation (10) can be obtained from (11) as $\eta_0 \approx \frac{N_{\mathrm{f}}^{\mathrm{S}}}{N_{\mathrm{f}}^{\mathrm{S}} + k_\mu N_{\mathrm{f}}^{\mathrm{I}}}$.

TABLE I
SIMULATION PARAMETERS

Evaluation duration, T_0	30 days
Content lifespan, T	5 days
Average content arrival rate, λ	10^3 contents/day
Average request volume of each SNM content during lifespan, $\mathbb{E}[V_m]$	3 [4]
Content catalogue of IRM, $N_{\mathrm{f}}^{\mathrm{I}}$	5×10^3 contents
Cache size of BS, N_{c}	100 contents
Average request arrival rate ratio of each IRM content to each SNM content, k_μ	0.5
Popularity skewness, $\delta = \beta^{-1}$	$2/3$

Compared with (2), we find that $\eta_0 \approx w^{\mathrm{S}}$, which is irrelevant to N_{c}. This can be explained as follows. When content popularity is quite skewed, almost all requests are generated by very few popular contents. Therefore, it is sufficient to only proactively cache a small number of popular contents and then reactively cache the contents with short lifespan according to the fraction of requests for SNM contents w_{S}.

2) Case 2: $\beta = 3$, i.e., the popularity distribution has moderate skewness. In this case, the expression of η_0 can be obtained from (11) as $\eta_0 \approx \max\left\{1 - \frac{k_\mu^3 N_{\mathrm{f}}^{\mathrm{I}}}{8N_{\mathrm{c}}}, 0\right\}$.

It is shown that for less skewed popularity distribution, the fraction of cache resource allocated to IRM contents, i.e., $1 - \eta_0$, is proportional to $N_{\mathrm{f}}^{\mathrm{I}}$, but inverse proportional to N_{c}.

3) Case 3: β is very large, i.e., the popularity distribution tends to be uniform. Then, we can derive η_0 from (11) based on the approximation $\left(\frac{\beta-2}{\beta-1}\right)^{\beta} \approx e^{-1}$ as

$$\eta_0 \approx \begin{cases} 0, & \text{if } k_\mu > 1, \\ \max\left\{1 - e^{-1}\frac{N_{\mathrm{f}}^{\mathrm{I}}}{N_{\mathrm{c}}}, 0\right\}, & \text{if } k_\mu = 1, \\ 1, & \text{if } k_\mu < 1. \end{cases} \tag{12}$$

It follows that the allocation of cache resource simply depends on k_μ: the BS only caches the category of contents with higher average request arrival rate. When the IRM and SNM contents have the same average request arrival rate, i.e., $k_\mu = 1$, and the cache size is small so that $\frac{N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{I}}} \le e^{-1} = 0.37$ holds (which is often the case in real-world cellular networks), then only IRM contents are cached.

IV. Numerical and Simulation Results

In this section, we evaluate the accuracy of introduced approximations in Sec. III, and then compare the performance of the proposed caching resource allocation policy with two benchmark policies, finally, we observe the impact of equal-lifespan modeling. The simulation parameters are listed in Table I, which are used throughout the simulations if not otherwise specified. All results are averaged over 100 random realizations of content request processes.

A. Accuracy of the Approximations

We first evaluate the accuracy of the approximations adopted to obtain the numerical solution η^*. The results are shown in Fig. 1, where legend "Num" denotes the numerical solution and legend "Simu" denotes the solution obtained by exhaustive searching of η with step size 0.02. It is shown that the numerical solution is very close to the solution by exhaustive searching, meaning that the employed approximations, including Che's approximation and the approximations to derive $\bar{p}_{\mathrm{h}}^{\mathrm{S}}(\eta)$ and $\bar{p}_{\mathrm{h}}^{\mathrm{I}}(\eta)$, have minor impact on the optimization of cache resource allocation.

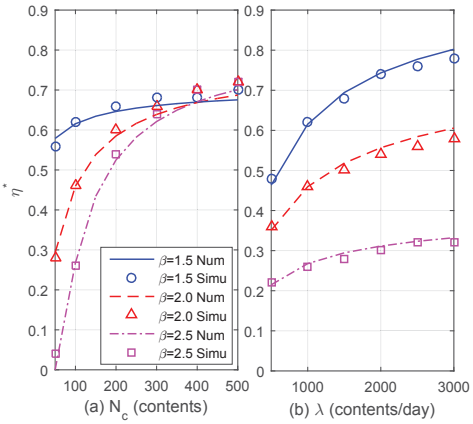

Fig. 1. Optimal cache allocation fraction versus cache size of BS and average content arrival rate

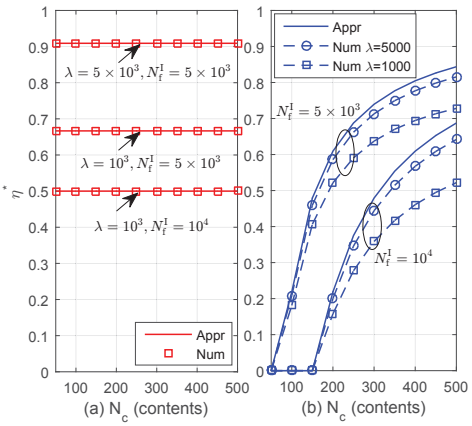

Fig. 2. Accuracy of approximations in special cases.

We next evaluate the accuracy of the approximations used to obtain the results under the special cases. The results are depicted in Fig. 2(a) and (b) for the cases with $\beta \approx 1$ and $\beta = 3$, respectively, where legend "Appr" denotes approximation results, and different λ and $N_{\mathrm{f}}^{\mathrm{I}}$ are considered. It is shown that the approximations used in the case with $\beta \approx 1$ is very accurate, but the accuracy for the case with $\beta = 3$ relies on $\frac{N_{\mathrm{c}}}{N_{\mathrm{f}}^{\mathrm{S}}}$ (note that $N_{\mathrm{f}}^{\mathrm{S}} = \lambda T$). Nevertheless, one can find that the numerical and approximation results exhibit the same trends, making the insights gained from the approximation result in *Case 2* still valid.

B. Performance Comparison

We compare the proposed caching resource allocation policy (with legend "Propose") with two benchmark caching policies, LRU and popular caching policy (with legends "LRU" and

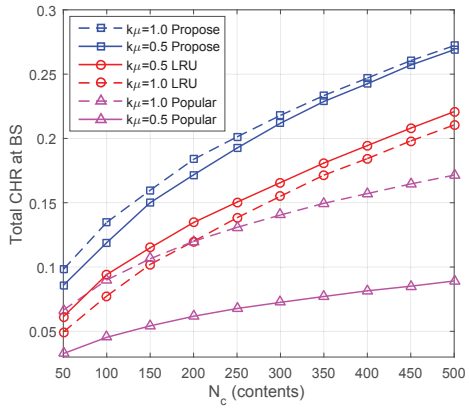

Fig. 3. Performance comparison of three caching policies.

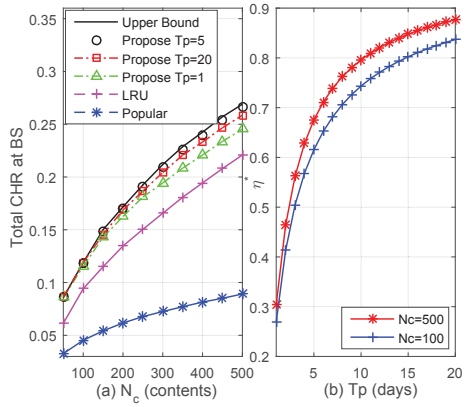

Fig. 4. Impact of random lifespan of SNM contents.

"Popular", respectively). The total CHR achieved by the three policies are depicted in Fig. 3. We can observe that the proposed policy obviously outperforms the two benchmark policies for both small and large average request arrival rate ratios k_μ. In particular, the gain of the proposed policy over the popular caching policy is remarkable for small k_μ. This is because popular caching policy implicitly assumes static popularity, so that the cached SNM contents are seldom requested beyond their lifespan, leading to the waste of cache resource.

C. Impact of Equal-lifespan Modeling

The proposed cache resource allocation policy is based on the simplified model that all SNM contents have equal lifespan T [10]. In Fig. 4, we simulate the performance of the proposed policy under random lifespan, where the lifespans of SNM contents follow uniform distribution between [1, 9] days with a mean value of 5 days. To apply the proposed policy in this scenario, we replace the fixed lifespan T with a predicted mean value of the random lifespan, denoted by T_p. We consider different values of T_p to reflect different prediction accuracy. If $T_\mathrm{p} = 5$ days, then the prediction is accurate, otherwise it is inaccurate.

In Fig. 4(a), the curve with legend "Upper Bound" denotes the simulation results where all SNM contents have the same lifespan of 5 days. We can find that considering random lifespan but with accurate prediction of mean lifespan (i.e., $T_\mathrm{p} = 5$ days) only leads to slight performance loss. When inaccurate prediction is considered (e.g., $T_\mathrm{p} = 1$ day or 20 days), the proposed policy exhibits a larger performance loss as expected, which however still outperforms the two benchmark policies. Interestingly, the performance with $T_\mathrm{p} = 20$ days is better than that with $T_\mathrm{p} = 1$ day, which implies that an aggressive prediction of the mean lifespan is more desirable than a conservative prediction. This can be explained by Fig. 4(b). It is shown that the increasing speed of η^* decreases with T_p, which means that a conservative prediction of mean lifespan tends to lead to a larger bias of the optimal caching resource allocation.

V. Conclusions

In this paper, we studied the wireless edge caching problem under a hybrid traffic model consisting of two categories of contents with long and short lifespans, which are respectively characterized by IRM and SNM models. Aimed at maximizing the total cache hit ratio of the BS, a cache resource allocation problem was formulated and solved, and the behavior of the optimal solution was analyzed in several special cases. Numerical and simulation results demonstrated the performance gain of the proposed hybrid caching policy over non-hybrid caching policies for the scenarios with either equal or random lifespan for SNM contents.

References

[1] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[2] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[3] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[4] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Unravelling the impact of temporal and geographical locality in content caching systems," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1839–1854, Oct. 2015.

[5] F. Olmos, B. Kauffmann, A. Simonian, and Y. Carlinet, "Catalog dynamics: Impact of content publishing and perishing on the performance of a LRU cache," in *Proc. IEEE ITC*, 2014.

[6] J. G. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," in *Proc. IEEE/WIC/ACM WI-IAT*, 2010.

[7] Y. Zhou, L. Chen, C. Yang, and D. M. Chiu, "Video popularity dynamics and its implication for replication," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1273–1285, Aug. 2015.

[8] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of Internet short video sharing: A YouTube-based measurement study," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug. 2013.

[9] B. Lee, C. Pei, F. Li, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, 1999.

[10] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc. IEEE INFOCOM*, 2016.

[11] K. Qi, S. Han, and C. Yang, "Dynamic popularity driven caching optimization at base station," *Journal paper to be submitted.*

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambrige University Press, 2004.

[13] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.