# Dynamic Cache Placement, Node Association, and Power Allocation in Fog Aided Networks

Ruoguang Li, Li Wang, Yanmin Gong, Mei Song, Miao Pan, and Zhu Han

***Abstract*—Proactively caching content at the edge is a promising way of alleviating the traffic load in the core networks, especially for latency sensitive video services. The content popularity herein is an effective measure for making efficient cache placement decisions. Most state-of-the-art works design the caching strategies under the assumption that video popularity is time-invariant. In practice, however, the popularity of different videos is constantly evolving over time, which will influence the quality of service and providers' operating efficiency. Therefore, in this paper, by considering a specified model of evolutionary video popularity, we investigate dynamic content placement, node association, and power allocation strategy in fog aided networks to maximize the time-average utility, meanwhile guaranteeing the queue in each edge node to remain stable. Utilizing the framework of the Lyapunov optimization, the formulated problem is transformed into an instantaneous mixed integer nonlinear programming problem at each time slot, which is then solved by our proposed iterative algorithm which jointly utilizes the greedy method and Generalized Benders decomposition. The simulation results demonstrate the effectiveness of our proposed algorithm and the tradeoff between the utility and average queue length.**

***Index Terms*—fog caching, dynamic video popularity, Lyapunov optimization, mix integer nonlinear programming.**

## I. Introduction

Due to the rapid development of mobile video service, the accompanied increase of data traffic causes a nonnegligible burden for wireless networks. Apart from the demand of high capacity for the video service, low latency also poses stringent challenges, especially in high-definition real time video streaming. Proactively distributing popular contents to femto-cell access points at the off-peak hours, caching is an effective way of avoiding duplicate content transmission from central servers [1]. Thus, caching at the edge node (EN) in the heterogeneous wireless networks, namely, fog caching [2], becomes a potentially practical solution for relieving backhaul pressures and providing more reliable video services to the mobile users.

Content popularity, which is referred to the demands for a given content from users, plays an important role in caching systems [3]. To date, designing content caching strategies mostly in an offline manner by ideally assuming that the content popularity is known and follows a certain distribution over time (e.g. Zipf distribution). However, such a time-invariant assumption is impractical since content popularity is dynamic in the real world. First, new contents are steadily flowed in the requested set, whilst part of the old contents tend to vanish [4]. Second, the popularity of different content varies with different patterns, especially in the video streaming [5]. Therefore, when taking content popularity evolution into consideration, the new problem arises for tackling two classic challenges in the wireless caching networks: 1) *content placement* problem and 2) *node association* and *resource allocation* issues [6]. In the context of time-variant content popularity, few works focused on the cache placement, node association, and resource allocation. Concretely, [7] proposed a hybrid content cache placement without the *prior* knowledge of content popularity, for guaranteeing the network stability. In [8], the authors studied service caching and user-base station association for dense cellular networks. However, in these works, the service request process, which should highly depend on the evolution of content popularity, was simply modeled by a homogeneous Possion process, ignoring the intrinsically non-stationary effects of users' requests.

In this paper, we introduce a newly proposed traffic model, namely, Shot Noise Model (SNM) [9], to accurately capture the dynamic of non-stationary video popularity. Unlike the existing works in which the popularity profile is considered as a rectangular function, we instead use a more practical model termed Evomodel proposed in [10] to quantify such a parameter. Thus, the patterns on how content popularity evolves can be predicted. Based on this, we correspondingly propose a joint dynamic caching placement, node association, and power allocation scheme for a cache-enabled wireless fog-aided network, to maximize the average utility over time, while considering network stability, storage capacity, and long term energy power consumption simultaneously. In order to fulfill this scheme, we employ the Lyapunov optimization technique to solve the formulated problem only with current state of the system at each time slot. To this end, we propose an iterative algorithm which combines the greedy method and Generalized Benders decomposition to solve such a mixed integer nonlinear (MINLP) problem.

The remainder of the paper is organized as follows: Section II presents the system model. Section III presents the optimization problem formulation. For solving the problem, the Lyapunov optimization is proposed in Section IV, and Section V presents our proposed iterative algorithm. The simulation results are

R. Li, L. Wang (corresponding author), and M. Song are with School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China. L. Wang is also with the Key Laboratory of the Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China. Y. Gong is with School of Electrical and Computer Engineering, University of Texas at San Antonio, TX, USA. P. Miao and Z. Han are with Department of Electrical and Computer Engineering, University of Houston, TX, USA.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

illustrated in Section VI, and Section VII draws the conclusions.

## II. System Model

Our system considers a wireless heterogeneous cache-enabled network, which consists of a central controller, and $N$ edge nodes (ENs) which serve a set of $M$ mobile users over a shared wireless channel. Through reliable backhaul links, ENs are connected with the controller, and then linked with the cloud and core network. Each EN is endowed with caching capability, and each of them caches a subset of popular files. Besides, since the popularity of video content dynamically evolves, we assume that the network operation is time-slotted with index $t \in \{0, 1, 2...\}$, and the duration of time slot can be measured at the centralized controller. The communication channels between the ENs and their connected users are also time-varying, but since the user locations are relatively static at the same time slot so that channels change slowly. Thus the channel coefficients remain constant within a time slot and to be drawn in an independent identically distributed (i.i.d.) manner from a continuous distribution. Besides, all the noise components are modeled as additive white Gaussian noise (AWGN) with zero mean and variance $N_0$. We also assume that a multi-user scheduling is employed between each user and the associated EN, such as time division multiple access (TDMA) based scheduling.

### A. Dynamic Model of Video File Popularity

It is practical and reasonable that video popularity evolves over time. SNM is an effective tool to replace traditional traffic models, capturing the non-stationary evolution of video popularity. Basically, for a given video file $f$ at a cache, the arrival process of requests in SNM is formulated by a non-homogeneous Poisson process with the following features: 1) The average number of requests $V_f$; 2) Content arrival time $t_f$; 3) Normalized popularity profile $\Lambda_f(t)$, i.e., $\Lambda_f(t) > 0$ and $\int_0^\infty \Lambda_f(t)dt = 1$. Consequently, the request arrival rate of the content $f$ at the time slot $t$ can be expressed as

$$\mu_f(t) = V_f \Lambda_f(t - t_f). \tag{1}$$

Generally, the particular shape of $\Lambda_f(t)$ is modeled by a rectangular shape. Such a simplification needs improvement if we consider a more practical scenario where the popularity profile changes over time as well. Fortunately, the authors in [10] demonstrated how the video popularity changes based on real data. They proposed a model named *Evomodel* that captures the way how the video is recommended to the user. The spread of information is illustrated in two ways: Direct recommendation and word-of-mouth recommendation, respectively. In specific, at the time slot $t$, $x_f(t)$ represents the number of users who know and prefer to watch the video $f$, and $y_f(t)$ is the number of users who do not know video $f$. The process of how the number of users who know and want to watch the video $f$ can be quantified through the following ordinary differential equation (ODE)

$$\lambda_f(t) = \dot{x}(t) = (\alpha_f y_f(t) + \beta_f y_f(t) x_f(t)) q_f, \tag{2}$$

where $\alpha_f$ and $\beta_f$ are the rate of direct recommendation and word-of-mouth recommendation to the video file $f$, respectively, and $q_f$ is the video attraction. The first term in the parenthesis represents the number of users who get the knowledge of video $f$ through direct recommendation with constant rate $\alpha_f$, and the second term represents that the users who get the knowledge of video $f$ through word-of-mouth with rate $\beta_f x_f(t)$. Because $\lambda_f(t)$ implies the increase rate of user number, so here we normalize $\lambda_f(t)$ and use it to represent the popularity profile of video $f$, i.e., $\Lambda_f(t) = \frac{\lambda_f(t)}{\lambda_{f \in \mathcal{F}} \lambda_f(t)}$. Finally, with SNM, the estimated popularity of $f$ can be expressed as [11]

$$\begin{aligned} p_f(t) &= \mathbb{E}(\mu_f | V_f, \tau_f) \\ &= \frac{\int_0^{\mu_f(t)} \eta_f (\eta_f t_f)^{N_f} (e^{-\eta_f t_f}/V_f) f(\eta_f) d\eta_f}{\int_0^{\mu_f(t)} (\eta_f t_f)^{N_f} (e^{-\eta_f t_f}/V_f) f(\eta_f) d\eta_f}, \end{aligned} \tag{3}$$

where $f(\eta_f)$ is the power-law density.

### B. Caching Model and Transmission Model

We assume the presence of a library of $F$ video files, which represent the content that may be requested by the mobile users. The set of the video files is denoted by $\mathcal{F} = \{1, 2, ..., F\}$, where the size of content $f$ is denoted as $s_f$ (in Mbits). In addition, caching control decisions are made by the centralized controller in the slot-by-slot basis. we introduce a binary variable $\pi_{i,f}(t)$ which represents the content placement decision at the time slot $t$. Specifically, $\pi_{i,f}(t)$ is equal to 1 if the video file $f$ is cached in $i$-th EN, and 0 otherwise. In addition, we denote $\rho_{j,f}(t) \in [0, 1]$ as the user $j$'s preference to the video $f$. Therefore, the probability of user $j$ requesting for video $f$ is written as $w_{j,f}(t) = p_f(t)\rho_{j,f}(t)$.

On the other hand, for the transmission phase, let us denote another indicator $\theta_{i,j}(t)$ which is equal to 1 if EN $i$ is associated with user $j$ at time slot $t$, and 0 otherwise. We assume that the wireless channel power gain between EN $i$ and user $j$ at the time slot $t$ is $|h_{i,j}(t)|^2$, and the transmit power is $P_{i,j}(t)$. Accordingly, the instantaneous rate from EN $i$ to user $j$, which is given by

$$r_{i,j}(t) = \pi_{i,f}(t)\theta_{i,j}(t) \log_2 \left(1 + \frac{|h_{i,j}(t)|^2 P_{i,j}(t)}{N_0}\right). \tag{4}$$

### C. Queue Model

The arrived but not yet served requests will be queued in the buffers at each EN. Thus, for EN $i$, the arrival of content delivering request from user $j$ is

$$O_{i,j}(t) = \pi_{i,f}(t)\theta_{i,j}(t) w_{j,f}(t) s_f(t). \tag{5}$$

We assume that $O_{i,t}(t) \leq O_{i,j}^{max}$ to ensure that the EN can process the video request task in an acceptable delay. To capture the dynamics of content placement and requests, we introduce queue between each EN $i$ and user $j$, the backlog in the queue is denoted as $Q_{i,j}(t)$, evolving in each time slot $t$ as follows:

$$Q_{i,j}(t+1) = [Q_{i,j}(t) - r_{i,j}(t)T_p]^+ + O_{i,j}(t), \tag{6}$$

where $[x]^+ = \max\{x, 0\}$. We assume that $T_p \leq T_p^{max}$ at each time slot $t$ to guarantee the minimum service delay. The queue length of current unserved request buffers will in turn influence the controller's decision about EN assignment in the next slot.

For any average requested content data rate inside the capacity region, all queues must remain mean stable, i.e.,

$$\lim_{\tau\to\infty}\sup\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbb{E}\{\sum_{i=1}^{N}\sum_{j=1}^{M}Q_{i,j}\left(t\right)\}<\infty. \tag{7}$$

## III. Utility Maximization Problem Formulation

At the beginning of each time slot $t$, the centralized controller make the joint caching placement, node association, and power allocation decisions based on the the information of requested data queue state information (QSI) and channel state information (CSI). Firstly, We define the system utility which represents the benefits minus cost of the system at the time slot $t$ as

$$U(t)= \tag{8}$$
$$\sum_{f=1}^{F}\sum_{i=1}^{N}\sum_{j=1}^{M}w_{j,f}(t)\pi_{i,f}\left(t\right)\theta_{i,j}(t)\left(\vartheta\sigma_j\log_{10}\left(s_f\left(t\right)\right)-\phi P_{i,j}\left(t\right)\right).$$

Specifically, $\sigma_j \log_{10}(s_f(t))$ maps the ideal video quality to the users' experience, wherein $\sigma_j$ is a predefined parameter regarding the quality of the video file (from standard-definition (SD) to high-definition (HD)) requested by user $j$. In other words, the larger $\sigma_j$ is, the higher quality of the video file with more details. $\vartheta$ is the unit benefit obtained from the content delivery, and $\phi$ is the unit cost during the transmission.

It is more reasonable to investigate the long-term performance in such a time-variant system. Therefore, the average utility of the considered all ENs is equal to

$$\overline{U(t)}=\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=0}^{\tau-1}U(t). \tag{9}$$

Correspondingly, the time-average utility maximization problem can be formulated as

$$\max_{\pi_{i,f}(t),\theta_{i,j}(t),P_{i,j}(t)}\ \overline{U(t)} \tag{10}$$
$$\begin{aligned}
s.t.\quad & \text{C1:}\pi_{i,f}(t)\in\{0,1\},\forall i,f,\quad \text{C2:}\theta_{i,j}(t)\in\{0,1\},\forall i,j,\\
& \text{C3:}\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=0}^{\tau-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\theta_{i,j}\left(t\right)\pi_{i,f}\left(t\right)P_{i,j}\left(t\right)\\
& \quad\leq P_i^{max},\forall l,f,\\
& \text{C4:}r_{i,j}(t)\geq r_{req},\forall i,j,\\
& \text{C5:}\sum_{f=1}^{F}\sum_{l=1}^{L}\pi_{i,f}(t)C_{l,f}(t)\leq C_i,\forall i,\\
& \text{C6:}\sum_{i=1}^{N}\theta_{i,j}(t)=1,\forall j,\quad \text{C7:}\sum_{i=1}^{N}\pi_{i,f}(t)=1,\forall f,\\
& \text{C8:}\lim_{\tau\to\infty}\sup\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbb{E}\{\sum_{i=1}^{N}\sum_{j=1}^{M}Q_{i,j}\left(t\right)\}<\infty,
\end{aligned}$$

where C1 is the caching decision constraint; C2 is the association decision constraint; C3 is a bound on the limit of the maximum average transmit power; C4 represents the minimum QoS requirement; C5 is the finite caching capacity constraint for EN $i$; C6 represents that each user is associated with only one EN at each time slot $t$; C7 guarantees that each video can only be deployed in one EN to reduce the duplicity; C8 is the network stability constraint that guarantees a finite queue length for each data queue in the EN.

## IV. Solution Development Based on Lyapunov Optimization Framework

In this section, we first give the following lemma to show how to tackle the average constraint C3. Then, based on the Lyapunov optimization framework, we build a connection between the utility maximization and queue stability.

**Lemma.** Construct a virtual queue, the queue dynamics of which are

$$H_{i,j}(t+1)=[H_{i,j}(t)-P_i^{max}]^+ +\theta_{i,j}\left(t\right)\pi_{i,f}\left(t\right)P_{i,j}\left(t\right). \tag{11}$$

Suppose $\mathbb{E}\{H_{i,j}(0)\}\leq\infty$. Then, the virtual queue $H_{i,j}(t)$ is mean-rate stable, the inequality holds.

We define $Q_i(t) = \sum_{j=1}^{j} Q_{i,j}(t)$ and $H_i(t) = \sum_{j=1}^{j} H_{i,j}(t)$. Let $\Theta\left(t\right)=\left[Q_i\left(t\right),H_i\left(t\right)\right]$ be a concatenated vector. Then the Lyapunov function is defined as a scalar metric of queue congestion

$$L(\Theta\left(t\right))=\frac{1}{2}\left(\sum_{i=1}^{N}Q_i^2\left(t\right)+\sum_{i=1}^{N}H_i^2\left(t\right)\right). \tag{12}$$

Since the arrival process is non-homogeneous Poisson Process which is non-i.i.d, the conventional one-slot Lyapunov optimization cannot guarantee the stable state. Thus, we use the delayed Lyapunov drift proposed in [12]. Without loss of generality, we assume that all queues are empty when $t$ = 0 such that $L(\Theta(0)) = 0$. Define the $T$-slot Lyapunov drift $\triangle(\Theta(t))$ as follows:

$$\triangle_T(\Theta(t))=\mathbb{E}\{L(\Theta(t+T))-L(\Theta(t))|\Theta(t)\}, \tag{13}$$

where $T$ used in the above can be considered as the time required for the system to reach near stable state. Minimizing the Lyapunov drift every time slot $t$ would stabilize the system, pushing queue backlog towards a lower congestion state. But we also want to maximize the objective function defined in (10) to maximize the utility. In order to build a connection between such two problems, we define *drift-minus-reward* function as

$$\triangle_T(\Theta(t))-V\mathbb{E}\{U(t)|\Theta(t)\}, \tag{14}$$

where $V$ is a non-negative control parameter which is used to adjust the tradeoff between $(\Theta(t))$ and $\mathbb{E}\{U\left(t\right)|\Theta(t)\}$. More specifically, $V$ is a weight that reflects how important the overall average utility maximization is in the optimization, i.e., larger $V$ represents the overall average utility is closer to its optimum at the expense of linearly increasing queue lengths, and vice versa.

Under the Lyapunov optimization technique, we obtain the joint cache placement, user association, and power allocation decisions by minimizing the upper bound of (14) at each time slot, as shown in (15), where $B$ is a positive constant which

**Algorithm 1** Dynamic Cache Placement, Node Association and Power Allocation Scheme (DCNP)

**Input:** $Q_{i,j}(0) \leftarrow 0$, $H_{i,j}(0) \leftarrow 0$, $h_{i,j}(t)$, $\Lambda_f(t)$ , $s_f$, $P_i^{max}$, $V$, $C_i$, $r_{req}$, $O_{i,j}^{max}$, $T_p^{max}$.
**Output:** $P_{i,j}(t)$, $\pi_{i,f}(t)$, $\theta_{i,j}(t)$.
1: **for** $t = 0$ to $\tau + T - 1$ **do**
2: Predict the service demands based on the evolution of content popularity.
3: Observe the current queue length $Q_{i,j}(t)$, virtual queue length $H_{i,j}(t)$, and channel coefficient $h_{i,j}(t)$.
4: **repeat**
5: Solve (16) by proposed iterative algorithm.
6: **until** Convergence=**true**
7: Update the queue length according to (6) and (11).
8: **end for**

is defined as $B \triangleq \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}[(O_{i,j}^{max})^2 + (r_{i,j}T_p^{max})^2] + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}(P_i^{max})^2$ [12].

Thus, the corresponding optimization problem can be transformed as

$$\max_{\pi_{i.f}(t),\theta_{i,j}(t),P_{i,j}(t)} G(\pi_{i,f}(t),\theta_{i,j}(t),P_{i,j}(t)) \quad (16)$$
$$s.t. \quad \text{C1, C2, C4, C5, C6, and C7,}$$

where $G(\pi_{i,f}(t),\theta_{i,j}(t),P_{i,j}(t))$ is shown in (17). The process is summarized in Algorithm 1.

## V. Iterative Algorithm Design for Near Optimal solution

Due to the existence of integer variables $\pi_{i,j}(t)$, $\theta_{i,j}(t)$, and continuous variable $P_{i,j}(t)$ in (16), the joint cache placement, node association, and power allocation problem is formulated as a MINLP problem which is usually NP-hard. Therefore, we design an iterative algorithm for solving such a problem to obtain the near optimal solution.

We solve (16) by decomposing it into a two-stage problem. Specifically, we first solve the cache placement problem with the given node association and power allocation, which is a knapsack problem. Then, based on the current cache placement, the joint node association and power allocation problem can be solved by the Generalized Benders decomposition method. It is intuitive to separate the problem into two parts since they occur in two consecutive phases in the caching system. The sub-optimal solution can be obtained after several iterations.

### A. Cache Placement

For the given node association and power allocation decision $(\tilde{\theta}_{i,j}(t), \tilde{P}_{i,j}(t))$ in which EN $i$ is in the set $\Omega_j(t) = \{i|\theta_{i,j}(t) = 1, P_{i,j}(t) \neq 0, i \in \{1,2...,N\}\}, \forall j \in \{1,2,...,M\}$, the cache placement problem can be equivalently transformed into the following knapsack problem

$$\max_{\pi_{i.f}(t)} G(\pi_{i,f}(t),\tilde{\theta}_{i,j}(t),\tilde{P}_{i,j}(t)) \quad (18)$$
$$s.t. \quad \pi_{i,f}(t) \in \{0,1\}, \forall f \in \mathcal{F}, i \in \Omega_j(t),$$
C4, C5, and C7.

In order to reduce the computational complexity, we use greedy algorithm to solve it. The above problem can be interpreted as follows: Given the benefit of caching file $f$ at EN $i$, which files should EN $i$ cache? We first define utility function for single cache placement, node association, and power allocation as

$$G_{i,j}^{f}(t) = (A(t) - B(t) - C(t))\pi_{i,f}(t), \quad (19)$$

where$A(t)$=$Vw_{i,f}(t)\theta_{i,j}(t)\left(\vartheta\sigma_f \log_{10}\left(s_f\left(t\right)\right) - \varphi P_{i,j}\left(t\right)\right)$, $B(t)$=$Q_{i,j}(t)w_{j,f}(t)\theta_{i,j}(t)s_f(t)$, $C(t)$=$H_{i,j}(t)\theta_{i,j}(t)s_f(t)$.

The feasible solution of (18) is simply caching the files that have the largest benefits, i.e.,

$$\pi_{i,f}(t) = \begin{cases} 0, & \text{if } G_{i,j}^{f}(t) < 0, \\ 1, & \text{if } G_{i,j}^{f}(t) > 0, i = \arg\min_{\delta\in\Omega_j(t)} G_{i,\sigma}^{f}(t), \\ 0, & \text{if } G_{i,j}^{f}(t) > 0, i \neq \arg\min_{\delta\in\Omega_j(t)} G_{i,\sigma}^{f}(t). \end{cases} \quad (20)$$

### B. User Association and Power Allocation

We update node association and power allocation based on the current cache placement $\tilde{\pi}_{i,j}(t)$ afterwards, which can be expressed as

$$\max_{\theta_{i,j}(t),P_{i,j}(t)} G(\tilde{\pi}_{i,f}(t),\theta_{i,j}(t),P_{i,j}(t)), \quad s.t. \text{ C2, C4, and C6.} \quad (21)$$

Generalized Benders decomposition (GBD) is an efficient tool for solving (21). The principle of GBD is to decompose problem into two subCproblems: a primal problem which aims to solve the problem only with pure continuous variable, and a master problem which aims to solve the pure integer programming problem. Specifically, solving the primal problem provides the information about the lower bound, and solving the master problem gives the information about the upper bound. When the upper bound meets the lower bound, the iterative process converges. The details of GBD are presented as follows.

*1) Solve the Primal Problem:* The primal problem represents the optimization problem that fixes the integer variables and solve a particular 0-1 combination denoted by $\theta_{i,j}^{(v)}(t)$, where $\nu$ is the current iteration counter. Therefore, the primal problem is expressed as

$$\max_{P_{i,j}(t)} G(\tilde{\pi}_{i,f}(t),\theta_{i,j}^{(\nu)}(t),P_{i,j}(t)), \ s.t. \quad \text{C4.} \quad (22)$$

$$\triangle_T(\Theta(t)) - V\mathbb{E}\{U\left(t\right)|\Theta(t)\} \leq BT^2 - V\mathbb{E}\{\sum_{t=\tau}^{\tau+T-1}\sum_{f=1}^{F}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{j,f}(t)\pi_{i,f}\left(t\right)\theta_{i,j}(t)\left(\vartheta\sigma_j \log_{10}\left(s_f\left(t\right)\right) - \varphi P_{i,j}\left(t\right)\right)|\Theta(t)\} \quad (15)$$
$$+ \mathbb{E}\{\sum_{t=\tau}^{\tau+T-1}\sum_{f=1}^{F}\sum_{i=1}^{N}\sum_{j=1}^{M} Q_{i,j}(t)[O_{i,j}(t) - r_{i,j}(t)T_p]|\Theta(t)\} + \mathbb{E}\{\sum_{t=\tau}^{\tau+T-1}\sum_{f=1}^{F}\sum_{i=1}^{N}\sum_{j=1}^{M} H_{i,j}(t)[\theta_{i,j}\left(t\right)\pi_{i,f}\left(t\right)P_{i,j}(t) - P_i^{max}]|\Theta(t)\}.$$

We can solve (22) by using the Lagrangian dual method. In addition, since the optimal solution to the primal problem (if exists) is also a feasible solution to the original problem, the optimal value provides a lower bound to the original problem. In general, however, not all choices of binary variables lead to a feasible the primal problem. Therefore, for a given choice of $\theta_{i,j}^{(v)}(t)$, there are two cases for primal problem: feasible problem and infeasible problem.

**Primal is feasible:** If the primal problem at the $\nu$-th iteration is feasible, its solution provides information on the transmit power of source nodes and relay. Then we can have the optimality cut, adding it to the master problem.

$$\alpha \leq \mathcal{L}(P_{i,j}(t), \lambda), \tag{23}$$

where $\mathcal{L}(P_{i,j}(t), \lambda)$ is the Lagrangian function of (22), and $\lambda$ is the Lagrangian multiplier.

**Primal is infeasible:** If the primal problem is infeasible, we can formulate an $l_1$ minimization problem to find a feasible point, which is given by

$$\min \quad s \quad s.t. \quad r_{i,j}(t) - r_{req} \leq s, \forall i, j, \tag{24}$$

and then we have the feasibility cut

$$0 \leq \bar{\mathcal{L}}(\bar{P}_{i,j}(t), \bar{\lambda}) = \bar{\lambda}(r_{i,j}(t) - r_{req} - s), \tag{25}$$

where $\bar{P}_{i,j}(t)$ is the solution of the $l_1$ norm optimization. And $\bar{\lambda}$ is the Lagrangian multiplier in this problem.

*2) Solving the master problem:* We can see the subproblem is reduced to a problem only with respect to the continuous variables. The problem can be further formulated as

$$\begin{aligned} \min_{\theta_{i,j}(t),\alpha} \quad & \alpha \\ s.t. \quad & \alpha \leq \mathcal{L}(\theta_{i,j}(t), P_{i,j}^{(p)}(t), \lambda^{(p)}), \forall (p \in \{1, 2, ..., S_1\}, \\ & 0 \leq \bar{\mathcal{L}}(\theta_{i,j}(t), \bar{P}_{i,j}^{(q)}(t), \bar{\lambda}^{(q)}), \forall q \in \{1, 2, ..., S_2\}, \\ & \text{C2, and C6,} \end{aligned} \tag{26}$$

where $S_1 + S_2 = \nu$. Such a pure integer programming problem can be solved by exploiting many methods, such as Branch-and-bound, cutting plane, etc.

**Theorem.** Let $\Psi(t)$ denote the set of all feasible solutions for $(\pi_{i,f}(t), \theta_{i,j}(t), P_{i,j}(t))$ and assume that for any decision $\eta(t) \in \Psi(t)$, the expectation of utility $U(\eta(t))$ is bounded as $U^{min} \leq \mathbb{E}\{U(\eta(t))\} \leq U^{max}$, where $U^{min}$ and $U^{max}$ are finite constants with respect to $\eta(t)$. If $O_{i,j}(\eta(t))$ is non.i.i.d., and assume $\tau = rT$ where $r = 1, 2, ..., r-1$, we have

$$\lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} U\left(\eta(t)\right) \geq \lim_{R\to\infty} \frac{1}{R} \sum_{t=0}^{R-1} U_r^* - \frac{DT}{V}, \tag{27}$$

$$\lim_{\tau\to\infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} \sum_{i=1}^{N} \sum_{j=1}^{M} Q_{i,j}\left(\eta(t)\right) \leq \frac{DT}{\epsilon} + \frac{V\left(U^{max} - U^{min}\right)}{\epsilon} + \frac{T-1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} \max\left[O_{i,j}^{max}, r_{i,j} T_p^{max}\right], \tag{28}$$

where $U_r^*$ is the optimal value in the $T$-slot for given $r$, and $\epsilon$ is a positive constant.

According to Theorem, we thus see a $[\mathcal{O}(V), \mathcal{O}(1/V)]$ tradeoff between the utility and backlog in this context.

## VI. Simulation Results

In this section, we evaluate the performance by carrying out the simulations. Default parameters are summarized in Table I. The parameters regarding the SNM ans Evomodel are selected from [9] and [10], respectively. For convenience, we assume that each video has the same size $s_f$. The types of videos are moive, news, music video, and episodes, respectively. The preference of user $j$ to video $f$ $\rho_{j,f}(t)$ is uniformly distributed from 0 to 1. The unit benefit $\vartheta$ and unit cost $\phi$ both are set to be 0.6.

TABLE I: Simulation parameters

| Parameter | Value |
|---|---|
| The number of EN $N$ | 5 |
| The number of users $M$ | 20 |
| The number of time slot $\tau$ | 500 |
| The number of near stable time slot $T$ | 20 |
| Total transmit power $P_i^{max}$ | 200 mW |
| Storage capacity of $i$-th EN $C_i$ | [300,4800] MB |
| Transmit delay $T_p$ | 2s |
| $N_0$ | -100 dBm/Hz |
| $O_{i,j}^{max}$ | 100 MB |
| $R^{max}$ | 100 MB/s/Hz |

In Fig. 1, the average queue length over time slot $t$ is presented with different values of the control parameter $V$. $C_i$ is set to be 2400 MB, and $s_f$=100 MB. It is observed that the value of the average queue length increases with $t$ and finally oscillates around a certain value, maintaining a "near" stable state, which is validated that the queue never exceeds the bound in (28). Besides, in the same time slot, a larger value of $V$ leads to a larger average queue length. It is due to the fact in (14) that when $V$ is larger, we pay more attention on the utility maximization. So each EN receives more requests from the users, resulting in boosted data traffic in the buffer.

Additionally, we compare the performance of the proposed algorithm with exhaustive search and caching without node association (NA) and power allocation (PA). Specifically, caching

$$\begin{aligned} G(\pi_{i,f}(t), \theta_{i,j}(t), P_{i,j}(t)) =& V \sum_{f=1}^{F} \sum_{i=1}^{N} \sum_{j=1}^{M} w_{j,f}(t) \pi_{i,f}\left(t\right) \theta_{i,j}(t) \left(\vartheta \sigma_f \log_{10}\left(s_f\left(t\right)\right) - \varphi P_{i,j}\left(t\right)\right) \\ & - \sum_{f=1}^{F} \sum_{i=1}^{N} \sum_{j=1}^{M} Q_{i,j}(t)[O_{i,j}(t) - r_{i,j}(t) T_p] - \sum_{f=1}^{F} \sum_{i=1}^{N} \sum_{j=1}^{M} \theta_{i,j}\left(t\right) \pi_{i,f}\left(t\right) H_{i,j}(t) P_{i,j}(t). \end{aligned} \tag{17}$$

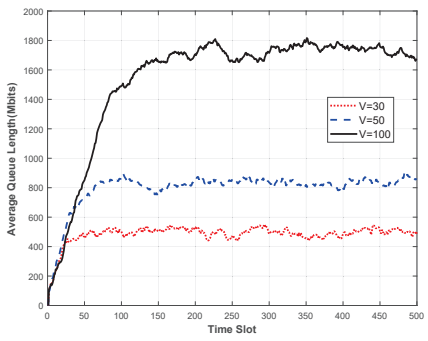


Fig. 1: Average queue length vs. V.

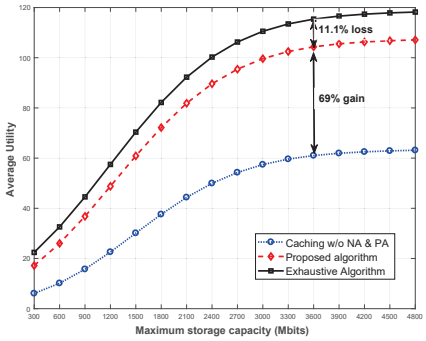


Fig. 2: Performance comparison.

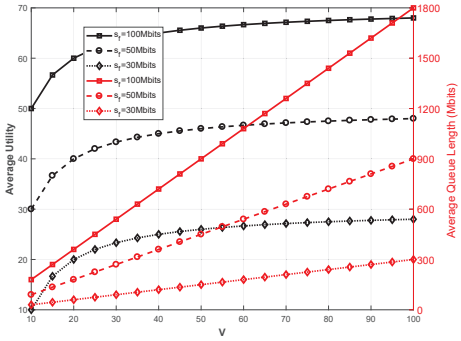


Fig. 3: Tradeoff of utility and Queue.

without (w/o) NA and PA means that we only consider caching the videos based on their popularity and queue stability, instead of optimizing the EN-user association and power allocation. As shown in Fig. 2, with $V$=100 and $s_f$=100 MB, the average utility increases as the EN's maximum storage capacity $C_i$ increases, because the EN can process more video requests from users. But the utility growth is slowing down when $C_i$ is sufficiently large. This happens because as more video requests come into the EN, the transmit power increases as well to make sure that the queue could maintain stable. It is also shown that the proposed algorithm outperforms the caching w/o NA and PA, achieving a 69% performance gain. However, compared to exhaustive search, there is only a 11% performance loss. In general, exhaustive search is much more time-consuming, so the performance loss is trivial especially in a delay-sensitive system.

Fig. 3, which has y-axes with both sides, illustrates the relationship between the average utility and average queue length with $C_i$=2400 MB. Apparently, we can see that the larger the value of video size $s_f$ is, the higher utility the system can obtain, and the larger queue length is as well. This fact derives from the mapping function of video quality in (8) that the video with larger size can provide higher quality of service, meanwhile it needs more time to be processed and transmitted, leading to higher level of backlog. Besides, we can see that the average utility grows at the speed of $\mathcal{O}(\frac{1}{V})$, but the average queue length grows with linear speed as $\mathcal{O}(V)$. This observation validates the tradeoff which is presented in Theorem.

## VII. Conclusion

In this paper, we have proposed a joint cache placement, node association, and power allocation scheme in the context of fog-aided wireless caching network with dynamic video popularity. Based on the Shot Noise Model and Evomodel, the dynamics of the video popularity was built to illustrate how the popularity changes over time. Accordingly, we aimed at maximizing the time-average network utility by considering the stability of the queue length of edge nodes. Utilizing the Lyapunov optimization, the formulated mix integer nonlinear programming problem was solved by our proposed algorithm in each time slot. Simulation results demonstrated the effectiveness of our proposed scheme and verified the analysis of a $[\mathcal{O}(\frac{1}{V}), \mathcal{O}(V)]$ utility-backlog tradeoff in the system.

## Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 61571056, 61871416, in part by the Beijing Science and Technology Nova Program under Grant xx2018083, in part by the Fundamental Research Funds for the Central Universities under Grant 2018XKJC03, and in part by the Beijing Municipal Natural Science Foundation under Grant L172010. The work of M. Pan was supported in part by the U.S. National Science Foundation under grants US CNS-1350230 (CAREER), CNS-1646607, CNS-1702850, and CNS-1801925.

## References

[1] L. Wang, H. Wu, Z. Han, P. Zhang, and H. V. Poor, "Multi-hop cooperative caching in social IoT using matching theory," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2127-2145, Apr. 2018.

[2] T. Liu, J Li, B. Kim, C. Lin, S. Shiraishi, J. Xie, and Z. Han, "Distributed file allocation using matching game in mobile fog-caching service network," in *proc. IEEE Conf. Comp. Commun. Workshops (INFOCOM WKSHPS)*, Honolulu, HI, Apr. 2018, pp. 499-504.

[3] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "The role of caching in future communication systems and networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1111-1125, Jun. 2018.

[4] M. Garetto, E. Leonardi, and S. Traverso, "Efficient analysis of caching strategies under dynamic content popularity," in *Proc. IEEE International Conf. Comput. Commun. (INFOCOM)*, Hongkong, China, Apr./May, 2015, pp. 2263-2271.

[5] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. S. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, early access.

[6] M. Choi, J. Kim, and J. Moon, "Wireless video caching and dynamic streaming under differentiated quality requirements," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1245-1257, Jun. 2018.

[7] J. Kwak, Y. Kim, L. B. Le, and S. Chong, "Hybrid content caching in 5G wireless networks: cloud versus edge caching," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3030-3045, May 2018.

[8] J. Xu, L. Chen, and P. Zhou, "Joint service caching and task offloading for mobile edge computing in dense networks," in *proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Honolulu, HI, Apr. 2018, pp. 207-215.

[9] K. Qi, S. Han, and C. Yang, "Dynamic popularity driven caching optimization at base station," in *proc. IEEE Annu. Int. Symp. Pers., Indoor, and Mob. Radio Commun. (PIMRC)*, Montreal, QC, Oct. 2017, pp. 1-5.

[10] J. Wu, Y. Zhou, D. M. Chiu, and Z. Zhu, "Modeling dynamics of online video popularity," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1882-1895, Sept. 2016.

[11] J. Ji, K. Zhu, R. Wang, B. Chen, and C. Dai, "Energy efficient caching in backhaul-aware cellular networks with dynamic content Popularity," *Wireless Commun. and Mob. Compt.*, vol. 2018, Article ID 7532049, 12 pages, 2018.

[12] M. Neely, *Stochastic network pptimization with application to communication and queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.