Optimal Buffer Resource Allocation in Wireless Caching Networks

Tingting Liu*†, Zheng Chang‡, Jun Li†, Feng Shu†, Tapani Ristaniemi‡, and Zhu Han§
*School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China
†School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
‡ Faculty of Information Technology, University of Jyvaskyla, FI 40014 Jyvaskyla, Finland
§Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA
E-mail: liutt@njit.edu.cn; zheng.chang@jyu.fi; jun.li@njust.edu.cn;
shufeng@njust.edu.cn; tapani.ristaniemi@jyu.fi; zhan2@uh.edu

***Abstract*—Wireless caching systems have been exhaustively investigated in recent years. Due to limited buffer capacity, and unbalanced arrival and service rates, the backlogs may exist in the caching node and even cause buffer overflow. In this paper, we first investigate the relationship among backlogs, buffer capacity, data arrival rate and service rate, utilizing the martingale theory which is flexible in handling any arrival and service processes. Then given a target buffer overflow probability, the minimal required buffer portion is determined. If the devoted buffer capacity can fulfill all serving users' minimal buffer requirements, an optimization problem is constructed with the objective to minimize the overall buffer overflow probability. The optimization solution is obtained by a modified water-filling scheme. Finally, the numerical results are illustrated to demonstrate the superiority of the proposed scheme.**

I. Introduction

The surge of a huge amount of duplicated data in cellular networks poses significant challenges to the current network architecture [1]. Instead of upgrading the network architecture, the wireless caching technology has been proposed as an economical and efficient solution in addressing the confronted challenges [2, 3]. The wireless caching technology has been demonstrated with the advantages of releasing traffic pressures on backhaul channels, enhancing delivery performance and improving the overall network energy efficiency. However, due to limited caching storage, some files or part of a file cannot be always stored in the caching nodes. Upon request, the cached files/parts can be delivered instantly to the requesting users, while the uncached ones will be delivered from remote servers, which may constitute a performance bottleneck. Buffer, which serves as a short-term memory to temporally store files inside a caching node, can enable buffer-aided relay that has been proposed to improve delivery performance of the uncached files [4, 5]. Recently, in order to enhance the performance of file delivery, researchers begin studying the joint design of buffer and cache operations in wireless caching networks [6–8], in which, some caching nodes are assumed to be equipped with sufficient buffer capacity to achieve a close-to-optimum delivery performance.

However, in practice, caching nodes may only devote a limited buffer capacity to help file transmissions. Meanwhile, due to the unbalanced data arrival and service rates, backlogs may exist inside a caching node. Too many backlogs may lead to queueing delay, buffer overflow, or even transmission failure. How to optimally utilize the limited buffer capacity to minimize the overall buffer overflow probability has arisen as a critical problem in realizing wireless caching. To our best knowledge, this problem has not been well addressed in the existing works. In order to solve this problem, we first need to figure out what is the relationship among backlogs, buffer capacity, data arrival rate and service rate.

Effective bandwidth theory can provide a useful framework to analyze the backlog/delay characteristics of broad classes of arrivals [9, 10]. It poses a very attractive property, i.e., additive property, which is useful in controlling the summation of flows' effective bandwidth to satisfy certain quality of service (QoS) requirements [11, 12]. However, the effective bandwidth theory may lead to loose estimations for non-Poisson, such as bursty arrival processes. Also, it is not always realistic to assume a user's arrival data follow a Poisson process in the concerned wireless caching networks.

Martingale theory is another valuable option in analyzing backlog/delay characteristics [13]. It is flexible and suitable for any arrival and service processes. Moreover, it can provide very tight estimations in a bursty traffic scenario. The authors in [14] first study the scheduling strategy for multimedia heterogeneous high-speed train networks, aiming to minimize the end-to-end delay based on the martingale theory. The simulation results are provided to show the tightness of the derived bounds compared with the real data trace. Then, they apply the martingale theory in multi-hop vehicular ad hoc networks [15]. Besides analyzing system delay or backlog bounds, the martingale theory-derived results can be used to perform system optimization. The authors in [16] propose to maximize network energy efficiency in machine type communication networks by taking the martingale theory-derived delay bounds as a constraint. The authors in [17] aim to minimize the network delay violation probability in a computation offloading scenario, where the objective function is derived from the martingale theory.

In this work, we aim to investigate the buffer resource allocation problem in the caching networks, by utilizing the martingale theory. The main contributions of this paper are listed as follows,

1) The buffer resource allocation problem is investigated in a wireless caching network. The relationships among backlogs, buffer capacity, data arrival and service rates

978-1-5386-6528-2/19/$31.00 ©2019 IEEE

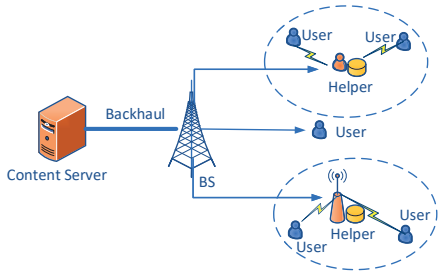

Fig. 1. A simple example on wireless caching system.

are first investigated, and then the buffer overflow probability is provided.

2) Next, given a required buffer overflow probability, each user's minimal buffer portion can be determined. There are three possible cases that are related to the summation of the users' minimal buffer portions and the overall portion. The last case is detailed discussed.
3) Furthermore, the third case, i.e., the summation of the users' minimal buffer portion is smaller than the overall portion, is studied. An optimization problem aiming to minimize the overall buffer overflow probability is constructed. The optimal solution is obtained by a modified water-filling scheme.
4) Numerical results are illustrated to demonstrate effectiveness of the proposed scheme. The proposed scheme is verified to have a smaller overflow probability than the equal allocation scheme.

The remainders of this paper are organized as follows: the system model and the martingale theory-derived buffer overflow probability are presented in Section II. The optimal buffer allocation scheme is discussed in Section III. Numerical results are illustrated in Section IV, and finally conclusions are drawn in Section V.

II. System Model

In the considered system, network users $\mathcal{U}_i,\ i = 1, \ldots, N$ have some file transmission requirements. Instead of transmitting files directly from the base station (BS) to the requesting users, a nearby caching node, as shown in Fig. 1, is employed to enhance the delivery performance.

A. Data Arrival Model

When the BS receives the file transmission requests, it will send the requested files to the corresponding caching nodes. In the following, we use the term 'helper' to denote the caching node for simplicity. On the helper's side, we assume that a user $\mathcal{U}_i$'s bursty arrival data amount $a_i(k)$ at time k follows a Markov-modulated on off (MMOO) process which has two static status $\Pi_i^a \triangleq [\pi_i^0, \pi_i^1]$, On state π_i^0, there is no data arriving, i.e., $a_i(k) = 0$, while on state π_i^1, $a_i(k) = R_i$ bits/s, and $R_i > 0$. The corresponding state transition matrix is defined as $\mathrm{Tr}_i^a \triangleq \begin{bmatrix} 1-p_i & p_i \\ q_i & 1-q_i \end{bmatrix}$, where p_i represents the transition probability from state π_i^0 to state π_i^1, while q_i represents the transition probability from π_i^1 to π_i^0. Accordingly, the steady state distribution Π_i^a of $a_i(k)$ is calculated as $[q_i/(p_i+q_i), p_i/(p_i+q_i)]$. Therefore, the cumulative arrival data over time interval $[m, n]$ is represented as

$$A_i(m,n) = \sum_{k=m}^{n} a_i(k), \tag{1}$$

where $A_i(m,n)$ can be regarded as a bivariate arrival process. If $m = 0$, we use $A_i(0,n) \triangleq A_i(n)$ for brevity.

B. Service Model

A helper is assumed to devote a limited buffer capacity C_B to relay files. The helper will firstly buffer each user's data, and then transmit to the corresponding users. The transmission rate at time k from the helper to the corresponding user can be represented as,

$$s_i(k) = \frac{B}{N}\log_2\left(1 + \frac{P_{\mathrm{tr}} d_i^{-l}}{\frac{B}{N}\sigma_0^2}\right), \tag{2}$$

where B is the bandwidth, N is the number of serving users, P_{tr} is the transmission power, d_i is the distance between a user $\mathcal{U}_i$ and the helper, l is the pathloss exponent, and σ_0^2 is the noise density. If the small scale channel fading is taken into consideration, a truncated channel inversion scheme can be employed to maintain a constant receiving power which leads to a similar expression shown in (2). This will not affect the following discussions. Thereby, the small scale channel fading is not considered here for simplicity. The accumulative transmission amount over time interval $[m, n]$ can be written as

$$S_i(m,n) = \sum_{k=m}^{n} s_i(k). \tag{3}$$

When $m = 0$, $S_i(0,n)$ can be rewritten as $S_i(n)$. It can be seen from (2) that given a user's location d_i, the service rate $s_i(k)$ is a constant irrelevant of time k. In the following, we use s_i for brevity.

In order to make the problem non-trivial, we assume

$$\mathbb{E}[a_i(k)] < s_i < \sup a_i(k), \forall i. \tag{4}$$

It means that a user $\mathcal{U}_i$'s service rate s_i is larger than the expectation of $a_i(k)$, but it is smaller than the peak rate of $a_i(k)$. In this case, on the one hand, data may experience certain queueing delay before it is transmitted, and on the other hand, there will be some backlogs inside the helper's buffer.

C. Buffer Overflow Probability

Since a helper only devotes a limited buffer capacity, the backlogs inside a helper become a non-negligible issue. In order to serve all the users at the same time, the helper will allocate each user with a dedicated buffer capacity. In other words, the files intended for the same user will line in the same queue, and each user's backlogs inside the helper can be represented as follows,

$$Q_i = \sup_{n \geq 0}\{A_i(n) - S_i(n)\}. \tag{5}$$

We use $\alpha_i C_B$ to denote the capacity allocated to a user $\mathcal{U}_i$, where $\alpha_i, 0 \leq \alpha_i \leq 1$ represents the portion allocated to a user $\mathcal{U}_i$, and we have $\sum_{i=1}^{N} \alpha_i \leq 1$. The allocated portion α_i is associated with each user's buffer overflow probability which is defined as follows.

Definition 1. Buffer Overflow Probability *is the probability that a user $\mathcal{U}_i$'s backlog Q_i is larger than the allocated buffer capacity $\alpha_i C_B$, i.e., $Pr\left(Q_i \geq (\alpha_i C_B)\right)$.*

Since the backlogs show stochastic characteristics, it is challenging to precisely determine its amount. In the following, we will resort to the martingale theory that can provide stochastic analysis on the buffer overflow probability.

Firstly, we need to construct two martingales [13], i.e., arrival martingales $M_A(n)$,

$$M_A(n) \triangleq h_a\left(a\left(n\right)\right) e^{\theta(A(n)-nK_a)}, n \geq 0, \tag{6}$$

and service martingales $M_S(n)$,

$$M_S(n) \triangleq h_s\left(s\left(n\right)\right) e^{\theta(nK_s-S(n))}, n \geq 0. \tag{7}$$

In this paper, each user's arrival data to the helper is constructed as arrival martingales. The transmission data from the helper to the corresponding user is constructed as service martingales. In this way, the arrival and service martingales are described by parameters $h_a\left(a\left(n\right)\right)$, K_a, $h_s\left(s\left(n\right)\right)$, and K_s, respectively.

In this paper, first in first out (FIFO) scheduling policy is assumed. In the following, we will provide the buffer overflow probability in **Theorem 1**. In order to make notations clear, we add index i to indicate the specific martingale parameters associated to the user $\mathcal{U}_i$.

Theorem 1. *The buffer overflow probability is calculated as*

$$Pr\left(Q_i \geq (\alpha_i C_B)\right) \leq \frac{\mathbb{E}\left[h_i^a\left(a\left(0\right)\right)\right]}{H_i} e^{-\theta_i^* \alpha_i C_B}, \tag{8}$$

where $H_i = \min\{h_i^a\left(a\right) h_i^s\left(s\right) | a_i - s_i > 0\}$.

Proof: Due to page limitation, proof is omitted here. Please refer to [13, 14] for the rigorous deduction. ■

Given a target buffer overflow probability ε, the minimal required buffer portion is calculated as

$$\alpha_i^{\min} = \frac{1}{\theta_i^* C_B} \ln \frac{\Delta_i}{\varepsilon}, \tag{9}$$

where $\Delta_i \triangleq \frac{\mathbb{E}[h_i^a(a(0))]}{H_i}$. Considering the relationship between a user's minimal portion $\alpha_i^{\min}$ and the overall portion 1, there are three possible cases:

- If $\alpha_i^{\min} > 1$, the overall buffer capacity cannot fulfill this user's requirements. This user should be suggested to connect to the BS directly.
- If $\alpha_i^{\min} \leq 1$ and $\sum_{i=1}^{N} \alpha_i^{\min} > 1$, that is to say the devoted buffer resource cannot fulfill all the users' buffer requirements, the helper may choose to serve some users according to certain criteria. Due to page limitation, we do not discuss this case in this paper.
- If the summation of the users' minimal buffer requirements can be fulfilled by the available buffer resources, i.e., $\sum_{i=1}^{N} \alpha_i^{\min} \leq 1$, an optimization problem can be further constructed to minimize the overall buffer overflow probability. We will discuss this case in the next section.

III. Optimal Buffer Resource Allocation Scheme

When $\sum_{i=1}^{N} \alpha_i^{\min} \leq 1$, in order to minimize the overall buffer overflow probability, we formulate a buffer portion allocation problem as follows,

$$\begin{aligned} \min_{\alpha_i} \ & \sum_{i=1}^{N} \Delta_i e^{-\theta_i^* \alpha_i C_B}, \\ \text{s.t.,} \ & \sum_{i=1}^{N} \alpha_i \leq 1, \ \alpha_i \geq \alpha_i^{\min}. \end{aligned} \tag{10}$$

Since $\frac{\partial \Delta_i e^{-\theta_i^* \alpha_i C_B}}{\partial \alpha_i} < 0$ and $\frac{\partial^2 \Delta_i e^{-\theta_i^* \alpha_i C_B}}{\partial {\alpha_i}^2} > 0$, thus (10) is a convex optimization problem. Using the Lagrangian function, we have

$$L = \sum_{i=1}^{N} \left(\Delta_i e^{-\theta_i^* \alpha_i C_B} + \lambda \alpha_i + \gamma_i (\alpha_i - \alpha_i^{\min}) \right), \tag{11}$$

where parameter $\{\gamma_1, ..., \gamma_N\}$, and λ represent the Lagrangian multipliers. The necessary and sufficient Karush-Kuhn-Tucher (KKT) conditions are listed as follows,

$$\begin{cases} \frac{\partial L}{\partial \alpha_i} = -\Delta_i \theta_i^* C_B e^{-\theta_i^* \alpha_i C_B} + \lambda + \gamma_i = 0, \\ \gamma_i \leq 0, \\ \gamma_i (\alpha_i - \alpha_i^{\min}) = 0. \end{cases} \tag{12}$$

The third condition in (12), i.e., $\gamma_i(\alpha_i - \alpha_i^{\min}) = 0$, leads to $\gamma_i \neq 0$ and $\alpha_i = \alpha_i^{\min}$, or $\gamma_i = 0$ and

$$\alpha_i = \frac{1}{\theta_i^* C_B} \left(\ln \frac{1}{\lambda^*} - \ln \frac{1}{\Delta_i \theta_i^* C_B} \right), \tag{13}$$

where λ^* is a tunable variable to satisfy the constraints $\sum_{i=1}^{N^*} \alpha_i \leq 1$ and $\alpha_i \geq \alpha_i^{\min}$. Thus the solution of (10) can be written as

$$\alpha_i^* = \left(\frac{1}{\Gamma_i} \left(\ln \frac{1}{\lambda^*} - \ln \frac{1}{\Delta_i \Gamma_i} \right) - \alpha_i^{\min} \right)^+ + \alpha_i^{\min}, \tag{14}$$

where $\Gamma_i \triangleq \theta_i^* C_B$, and $(x)^+ \triangleq \max(0, x)$.

Observing (14), it can be noted that the solution of the proposed problem is a modified water-filling scheme. Fig. 2 depicts the illustration of the solution. Firstly, each user is allocated with an amount of $\alpha_i^{\min}$ which is the minimal portion that guarantees its buffer overflow probability. Since we have $\sum_{i=1}^{N} \alpha_i^{\min} \leq 1$, the remaining portions, i.e., $1 - \sum_{i=1}^{N} \alpha_i^{\min}$, can be utilized to further minimize the overall buffer overflow probability. The filled water is adjustable by the water level $\ln(\frac{1}{\lambda^*})$ as shown in Fig. 2. The bowl bottom of each user is determined by $\ln(\frac{1}{\Delta_i \Gamma_i})$ which is associated with the state of each user, and can be regarded as a user indicator. When $\ln(\frac{1}{\lambda^*})$ is larger than $\ln(\frac{1}{\Delta_i \Gamma_i})$, and $\frac{1}{\Gamma_i}\left(\ln \frac{1}{\lambda^*} - \ln \frac{1}{\Delta_i \Gamma_i}\right)$ is larger than $\alpha_i^{\min}$, such as the users in Fig. 2 with indices 2, 4

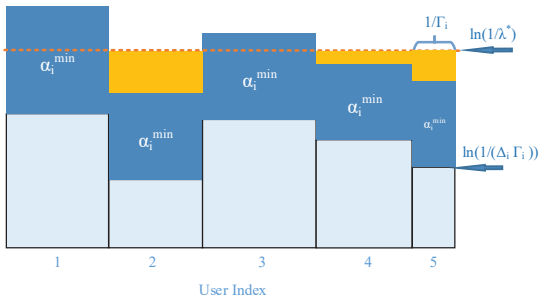

Fig. 2. Illustration of the optimal solution.

and 5, this kind of users will be allocated with more portions except for the minimal portion $\alpha_i^{\min}$. When $\ln(\frac{1}{\lambda^*})$ is larger than $\ln(\frac{1}{\Delta_i\Gamma_i})$, but $\frac{1}{\Gamma_i}\left(\ln\frac{1}{\lambda^*}-\ln\frac{1}{\Delta_i\Gamma_i}\right)$ is smaller than $\alpha_i^{\min}$, such as the users in Fig. 2 with indices 1 and 3. This kind of users will be only allocated with the minimal portion $\alpha_i^{\min}$. At the same time, it can be seen that the filled water is determined by both the hight value $\left(\ln\frac{1}{\lambda^*}-\ln\frac{1}{\Delta_i\Gamma_i}\right)$ and the width value $\frac{1}{\Gamma_i}$.

IV. Numerical Results

In this section, we conduct numerical simulations to demonstrate effectiveness of the proposed buffer allocation scheme. The simulation parameters are listed in Table I. We also provide the specific simulation settings in each simulation scenario.

Firstly, in Fig. 3, we present the relationship between the minimal required buffer portion $\alpha_i^{\min}$ and the user's arrival rate, as well as $\alpha_i^{\min}$ and the devoted buffer capacity C_B. Given $p_i = 0.4$, the user's arrival rate is assumed to vary from 4 Mega bits/s (Mbits/s) to 5 Mbits/s, and the user's distance from the helper is 20 m. Intuitively, as the user's arrival rate growing, given the same service rate, the backlog will be increasing. Thus, the minimal required buffer portion should be increased in order to guarantee the same buffer overflow probability. The above intuition can be verified from Fig. 3 that given $C_B = 120$ Mbits, as a user's arrival rate varies from 4 Mbits/s to 5 Mbits/s, the minimal buffer portion $\alpha_i^{\min}$ increases dramatically from 0.0615 to 0.2141. Then, we vary the helper's devoted buffer capacity from 120 Mbits to 600 Mbits. When $C_B = 600$ Mbits, i.e., 5 times increase of the devoted buffer capacity compared to $C_B = 120$ Mbits, a user's minimal buffer portion increases from 0.0123 to 0.0428, which is a 5 times drop compared to the value of $\alpha_i^{\min}$ when $C_B = 120$ Mbits. It can be concluded that the minimal buffer portion $\alpha_i^{\min}$ is inversely proportional to the devoted buffer capacity C_B.

Fig. 4 and Fig. 5 demonstrate the performance of the proposed buffer allocation scheme. Given 6 users with arrival rate $R_i = [4, 4.2, 4.4, 4.6, 4.8, 5]$ Mbits/s, $C_B = 90$ Mbits and 120 Mbits. The minimal buffer portion that can fulfill the buffer overflow probability requirement is shown in Fig. 4. It can be seen that as the user's index increases, i.e., the corresponding arrival rate increases, the minimal required portion grows. This observation is consistent with the trend

TABLE I
Simulation Parameters

Parameter	Value
Transition probability from 0 to 1	$p_i = 0.4$
Transition probability from 1 to 0	$q_i = 0.5$
User number	$N = 6$
Arrival rate of each user	$R_i \in [4, 5]$ Mbits/s
Transmitter bandwidth	$B = 0.5$ MHz
Transmission power	$P_{\text{tr}} = 20$ dBm
Noise density	$\sigma_o^2 = -174$ dBm
User's antenna gain	$G = 3$ dBi
Distance between helper and user	$d_i \in [20,\ 90]$ m
Pathloss exponent	$l = 4$
Buffer overflow probability	$\varepsilon = 0.05$
Devoted buffer capacity	$C_B \in [90, 600]$ Mbits

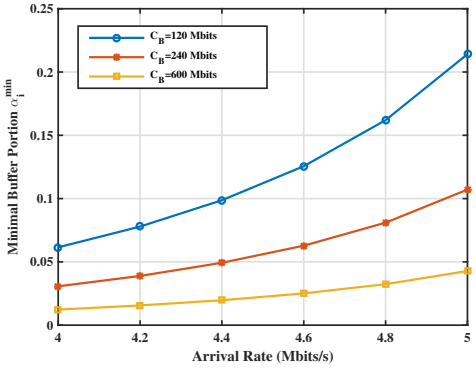

Fig. 3. The relationship between the minimal required buffer portion $\alpha_i^{\min}$ and a user's arrival rate, as well as $\alpha_i^{\min}$ and devoted buffer capacity C_B.

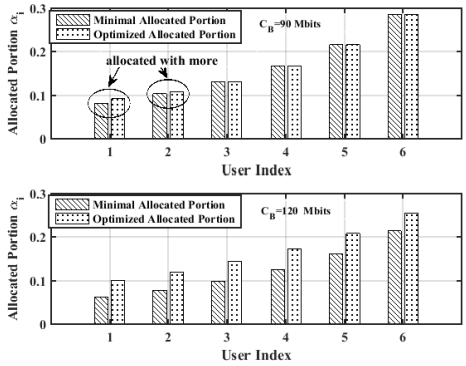

Fig. 4. The comparisons between the minimal allocated portion and the optimized allocated portion.

shown in Fig. 3. The optimized allocated portions which are obtained by the optimal solution shown in (14) are also shown in Fig. 4. The users $\mathcal{U}_1$ and $\mathcal{U}_2$ are firstly allocated with the remaining portion when $C_B = 90$ Mbits. However, since the buffer overflow probability is a decreasing and convex function over buffer portion α_i, it will not gain more if more portions are allocated to the users with small index. Therefore, in order to minimize the overall buffer overflow probability, when $C_B = 120$ Mbits, as seen in Fig. 4, the remaining portions are allocated to all the serving users.

Fig. 5 plots variation of the buffer overflow probability between the original value and the value after using the proposed scheme. It can be seen in Fig. 5 that when $C_B = 90$ Mbits, the buffer overflow probabilities of the users $\mathcal{U}_1$ and $\mathcal{U}_2$ are firstly reduced. When $C_B = 120$ Mbits, as the remaining portions

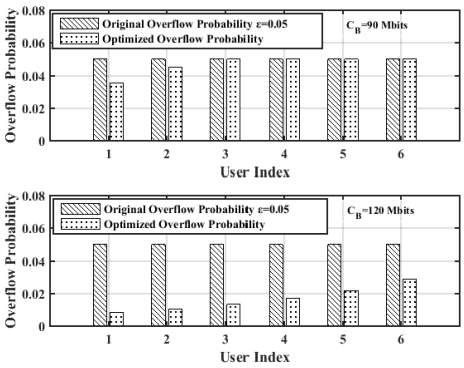

Fig. 5. The buffer overflow probability comparisons between the original value and the optimized value.

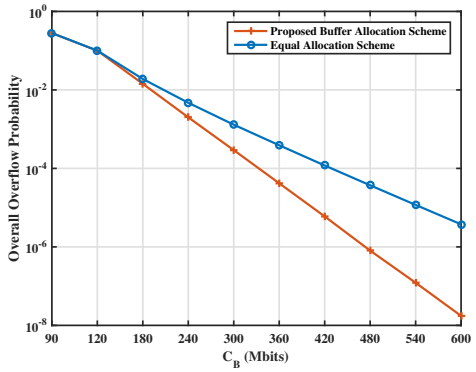

Fig. 6. Overall buffer overflow probability comparisons between the proposed scheme and the equal allocation scheme.

are allocated to all the users, the buffer overflow probabilities are significantly reduced by employing the proposed allocation scheme compared to the original probability $\varepsilon = 0.05$.

Fig. 6 compares the overall buffer overflow probability between the proposed scheme and the equal allocation scheme. The equal allocation scheme simply divides the remaining portions equally. We vary the devoted buffer capacity C_B from 90 Mbits to 600 Mbits. It can be seen from Fig. 6 that, as the devoted buffer capacity C_B increases, the overall buffer overflow probability decreases. The phenomena is consistent with the intuition that a large buffer can achieve a small buffer overflow probability. As shown in Fig. 6, when C_B is small, such as $C_B = 90$ Mbits and 120 Mbits, the two allocation schemes have comparable performance. When C_B is large, such as $C_B \geq 180$ Mbits, the proposed scheme achieves a much smaller probability compared to the equal allocation scheme does. Overall, the proposed allocation scheme has a much better performance compared to the benchmark.

V. Conclusion

This paper proposes an optimal buffer allocation scheme in wireless caching networks. The buffer overflow probability is derived based on the martingale theory. Given the required buffer overflow probability, the minimal buffer portion amount can be obtained. Considering the relationship between the summation of the users' minimal portion requirements and the overall buffer resources, there are three possible cases. This paper focuses on the third case, i.e., the summation of all the minimal portions is less than 1, the remaining portions can be further utilized to minimize the overall buffer overflow probability. The optimized allocation portion can be obtained by a modified water-filling scheme. Numerical results are provided to demonstrate effectiveness of the proposed scheme. Moreover, the proposed scheme has a much smaller overall buffer overflow probability compared to the equal allocation scheme.

Acknowledgment

This work is supported in part by the national natural science foundation of China under grant No. 61702258, 61771244, 61872184, in part by natural science foundation of Hebei Province of China, in part by US MURI, US NSF CNS-1717454, CNS-1731424, CNS-1702850, CNS-1646607.

References

[1] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

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

[3] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: proactive content caching and push with energy-harvesting-based small cells," *IEEE Communications Magazine*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[4] N. Zlatanov, R. Schober, and P. Popovski, "Buffer-aided relaying with adaptive link selection," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 8, pp. 1530–1542, Aug. 2013.

[5] N. Zlatanov, A. Ikhlef, T. Islam, and R. Schober, "Buffer-aided cooperative communications: opportunities and challenges," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 146–153, Apr. 2014.

[6] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, V. W. S. Wong, and J. Wang, "Cross-layer optimization of fast video delivery in cache- and buffer-enabled relaying networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11 366–11 382, Dec. 2017.

[7] G. Zheng, H. A. Suraweera, and I. Krikidis, "Optimization of hybrid cache placement for collaborative relaying," *IEEE Communications Letters*, vol. 21, no. 2, pp. 442–445, Feb. 2017.

[8] Z. Yang, C. Pan, Y. Pan, Y. Wu, W. Xu, M. Shikh-Bahaei, and M. Chen, "Cache placement in two-tier hetnets with limited storage capacity: Cache or buffer?" *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5415–5429, Nov. 2018.

[9] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[10] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[11] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[12] ——, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.

[13] F. Poloczek and F. Ciucu, "Service-martingales: Theory and applications to the delay analysis of random access protocols," in *IEEE Conference on Computer Communications (INFOCOM), Hong Kong, China*, Apr. 2015, pp. 945–953.

[14] Y. Hu, H. Li, Z. Chang, and Z. Han, "Scheduling strategy for multimedia heterogeneous high-speed train networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3265–3279, Apr. 2017.

[15] ——, "End-to-end backlog and delay bound analysis for multi-hop vehicular ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6808–6821, Oct. 2017.

[16] L. Zhao, X. Chi, and Y. Zhu, "Martingales-based energy-efficient d-aloha algorithms for mtc networks with delay-insensitive/urllc terminals co-existence," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1285–1298, Apr. 2018.

[17] T. Liu, J. Li, F. Shu, and Z. Han, "Quality-of-Service driven resource allocation based on martingale theory," in *IEEE Global Communications Conference: Communication QoS, Reliability and Modeling (Globecom CQRM)*, Abu Dhabi, United Arab Emirates, Dec. 2018.