QoS-Oriented Dynamic Power Allocation in NOMA-Based Wireless Caching Networks

Yue Yin, Member, IEEE, Miao Liu, Member, IEEE, Guan Gui, Senior Member, IEEE, Haris Gacanin, Senior Member, IEEE, Hikmet Sari, Life Fellow, IEEE, and Fumiyuki Adachi, Life Fellow, IEEE

Abstract—Non-orthogonal multiple access (NOMA) based wireless caching network (WCN) is considered as a promising technology for next-generation wireless communications since it can significantly improve the spectral efficiency. In this letter, we propose a quality of service (QoS)-oriented dynamic power allocation strategy for NOMA-WCN. In content placement phase, base station (BS) sends multiple files to helpers by allocating different powers according to the different QoS targets of files, for ensuring that all helpers can successfully decode the two most popular files. In content delivery phase, helpers serve two users at the same time by allocating the minimum power to far user according to the QoS requirement, and then all the remaining power is allocated to near user. Hence, our proposed method is able to increase the hit probability and drop the outage probability compared with conventional methods. Simulation results confirm that the proposed power allocation method can significantly improve the caching hit probability and reduce the user outage probability.

Index Terms—Non-orthogonal multiple access, wireless caching networks, dynamic power allocation, quality of service.

I. INTRODUCTION

SPECTRAL efficiency is one of the critical design challenges in the sixth generation (6G) wireless communication systems [1]–[3]. To improve the spectral efficiency, wireless caching techniques have been proposed to divide traffic load at BS into many helpers [4], [5]. In a typical communication system, BS accesses the core network to download the file and then transmits it to users. On the other hand, the main idea of wireless caching is to download the popular content by the helpers during content placement phase before it is requested. Consequently the users can be served locally. Some researchers studied the content placement of wireless caching, for example, Chae and Choi [6] control cache-based channel selection diversity with sophisticated considerations of wireless fading channels and interactions among multiple users. Kwak et al. [7] proposed a hybrid content caching scheme design that does not require the knowledge of content popularity. They optimize the content caching locations by employing the Lyapunov optimization approach. To decrease the latency and improve the utilization of caches, Liang et al. [8] designed a mechanism to jointly provide proactive caching, bandwidth provisioning, and adaptive video streaming.

Non-orthogonal multiple access (NOMA) has been recognized as one of the most promising wireless technologies for improving the spectrum efficiency in 6G mobile communications [9]. NOMA [10]–[13] enables different users to occupy the same spectrum, time, space through power multiplexing, further improves spectrum utilization and reduces users’ time delay. At the receiver, the signals of multiple users are separated by the successive interference cancellation (SIC). User grouping and power allocation are the two most important parts of the NOMA system. For user grouping, Yin et al. proposed a dynamic user grouping strategy to decrease the bit error rate [14]. Ding et al. [15] firstly introduced NOMA into WCN, where a fixed power allocation in content delivery phase, thus the allocated power might be unable to satisfy the users’ QoS demands dynamically. Zhao et al. [16] studied the coverage performance of NOMA-WCN while did not consider about power allocation scheme. To solve this problem, Fu et al. [17] studied the dynamic power allocation driven by deep neural network of NOMA-WCN according to the deadline of user transmission and the constraints of total power. For this method, they need to generate training data samples, validation data samples and test data samples. The total number of data samples is $10^5$, and this method of power allocation has high complexity. In [18], the authors considered channel selection, user requests and cache contents in given time and proposed an optimized power allocation strategy in a cache-aided cellular network with Rayleigh fading channels. Recently, they proposed a divide-and-conquer-based method and a deep reinforcement learning-based method in cache-aided NOMA systems [19]. The simulation results are given to prove the superiority of their methods. Further more, Gurugopinath et al. [20] introduced cache-aided NOMA for...
vehicular networks considering the case of split the content placement. They formulated a joint power allocation optimization problem, where power allocation across vehicles and cached split files is investigated.

In this letter, we propose a dynamic power allocation strategy for NOMA-WCN based on the QoS of both files and users within the constraints on BS’s power. Compared with deep learning-based methods, the complexity of our proposed QoS-oriented dynamic power allocation method is very low and a closed-form solution of QoS-oriented power allocation is derived to reduce deployment costs for both BS and helpers. Hence it can be utilized in a realistic NOMA-WCN. Computer simulations are conducted to evaluate the proposed method with respect to hit probability and the outage probability illustrating the effectiveness of the proposed method.

II. PROBLEM FORMULATION

The transmission of WCN is divided to two independent time phases, namely content placement phase and content delivery phase as shown in Fig. 1. Firstly, BS predicts which files are popular for users in the next period, and then downloads these files from the core network. During content placement phase, BS transmits these files to all helpers. At this phase, BS and the helpers are unable to serve the users, resulting in increasing their time delays. Therefore, content placement phase should be short. However, if content placement phase is too short, helpers will not have enough time to download enough files, so we apply NOMA to handle this problem. Then in content delivery phase, when the users request files from the helpers, if the files exist in their corresponding helpers, the helpers will directly send the files to users. If the helper does not store the files which are requested by users in advance, BS will serve users directly. However, BS will increase the load of the backhaul link and cause the network congestion. This case is not considered in this letter.

We consider multiple helpers distributed within the coverage of BS, and each helper serves multiple mobile users within its delivery range, as shown in Fig. 2. Moreover, it is assumed that each user is only associated with the nearest helper. The popularity of the files are modeled by the Zipf distribution. Assuming the total number of all the requested files is $F_{total}$, thus the popularity of file $f$ is computed as:

$$G(f) = \frac{1}{\sum_{i=1}^{F_{total}} \frac{1}{f^i}}$$

where $\varepsilon$ is the popularity factor of the file. In reality, the popularity of files in the coverage of each helper are different, because users in different area have different preferences [21]. However, this letter doesn’t consider this situation, because file popularity is not the focus of this letter.

III. THE PROPOSED QoS-ORIENTED DYNAMIC POWER ALLOCATION ALGORITHM

In this section, we present a QoS-oriented dynamic power allocation method in content placement phase and content delivery phase.

A. In Content Placement Phase

BS predicts files which are the most popular ones in the next period, and downloads these files from the core network. During content placement phase, BS send these files to the helpers. All files are allocated with different powers by the NOMA in BS. Each file $f$ has QoS requirement $R_f^{QoS}$, and it can be decoded correctly on the receiver only when the QoS requirement is satisfied. If the file gets more power, it is more likely to be decoded correctly at the helpers. In addition, if the helper is close to BS, channel conditions is better and it is easier to transmit the files successfully to the helpers. Therefore, the closer the helper is to BS, the more files can be decoded correctly.

In content placement phase, we only consider large-scale fading channel, because when arranging helpers, we can choose the location of helpers at a position with line-of-sight connection with BS. While in content delivery phase, we have to consider small-scale fading, because the line-of-sight connection cannot be ensured between the user and the helper all the time [15]. The path loss channel model from $CS_l$ to BS is:

$$d_l = \frac{1}{\left(\sqrt{|x_l - x_0|^2 + |y_l - y_0|^2}\right)^3},$$

where $CS_l$’s position is $(x_l, y_l)$, BS’s position is $(x_0, y_0)$, and the path loss factor is 3. In order to reduce the computational complexity, we assume that BS is located at the origin of the two-dimensional plane, hence the path loss model of $CS_l$ can be simplified as

$$d_l = \frac{1}{\sqrt{x_l^2 + y_l^2}}$$

BS sends a superimposed signal of 3 popular files to helpers at one time slot. And it is expected that at least the two most popular files can be correctly decoded by all helpers. The rate
for CS1 to decode the most popular file $f_1$ is:
\[
R_{l,1} = \log_2 \left( 1 + \frac{d_1 P_s \alpha_1}{d_1 P_s (2 + \alpha_3) + n_0} \right),
\]
where $P_s$ is the total power sent by BS, $\alpha_1$, $\alpha_2$, and $\alpha_3$ are the power allocation factors of files $f_1$, $f_2$, and $f_3$, respectively, and $n_0$ is noise power. The bandwidth is normalized to 1.

After passing through the fading channel, the file with higher power is more likely to be decoded correctly, so power is allocated to the most popular file $f_1$ first. The QoS of file $f_1$ is $R_1^{QoS}$. CS1 is able to successfully decode file $f_1$ only when $R_{l,1} \geq R_1^{QoS}$.

\[
\log_2 \left( 1 + \frac{d_1 P_s \alpha_1}{d_1 P_s (2 + \alpha_3) + n_0} \right) \geq \log_2 \left( 1 + SNR_1^{QoS} \right),
\]
where $SNR_1^{QoS}$ is SNR requirements for file $f_1$. The power allocation factors for all files satisfy $\sum_{j=1}^{3} \alpha_j = 1$. As a result, the power allocation factor for file 1 must satisfy:

\[
\alpha_1 \geq \frac{d_1 P_s SNR_1^{QoS} + n_0 SNR_1^{QoS}}{SNR_1^{QoS} + 1} \frac{d_1 P_s}{d_1 P_s (2 + \alpha_3) + n_0}.
\]

Then, $\alpha_1$ can be determined by:

\[
\alpha_1 = \min \left\{ 1, \frac{d_1 P_s SNR_1^{QoS} + n_0 SNR_1^{QoS}}{SNR_1^{QoS} + 1} \frac{d_1 P_s}{d_1 P_s (2 + \alpha_3) + n_0} \right\}.
\]

Similarly, after perfect SIC, the interference power from file $f_1$ can be canceled, and then the data rate for CS1 to decode the second most popular file $f_2$ is:

\[
R_{l,2} = \log_2 \left( 1 + \frac{d_1 P_s \alpha_2}{d_1 P_s (2 + \alpha_3) + n_0} \right).
\]

Hence, the power allocation factor $\alpha_2$ of file $f_2$ can be further calculated as:

\[
\alpha_2 = \min \left\{ 1 - \alpha_1, \frac{d_1 P_s SNR_2^{QoS} (1 - \alpha_1) + n_0 SNR_2^{QoS}}{SNR_2^{QoS} + 1} \frac{d_1 P_s}{d_1 P_s (2 + \alpha_3) + n_0} \right\}.
\]

Finally, file $f_3$ is decoded by the helper as far as possible. Thus the power allocation factor $\alpha_3$ can be taken as:

\[
\alpha_3 = \max \{1 - \alpha_1 - \alpha_2, 0\}.
\]

### B. In Content Delivery Phase

During content delivery phase, the users request files from their helpers. By applying NOMA, the files requested by two users can be superimposed and transmitted on the same spectrum simultaneously. After receiving the signal, the user near the helper is able to use SIC for separating the useful signal. In detail, the superimposed signal received by the near user of CS1 is

\[
y_{l,\text{near}} = h_{l,\text{near}} \sqrt{d_{l,\text{near}} \beta_{l,\text{near}} f_{l,\text{near}}} + h_{l,\text{near}} \sqrt{d_{l,\text{near}} \beta_{l,\text{far}} f_{l,\text{far}}} + \sum_{q \in \Phi \setminus C \setminus I} h_{q,\text{near}} \sqrt{d_{q,\text{near}} \beta_{q,\text{near}} f_{q,\text{near}}} + n_{l,\text{near}},
\]

where $h_{l,\text{near}}$ is the channel coefficient of CS1 to its near user, $\beta_{l,\text{near}}$ is the power factor of the near user of CS1 and $\beta_{l,\text{far}} + \beta_{l,\text{near}} = 1$, $f_{l,\text{near}}$ is the file of the near user of CS1, $f_{l,\text{far}}$ is the file of the far user of CS1, $q \in \Phi \setminus C \setminus I$ means the elements in set of helpers C except $l$. The first item on the right side is the signal of the near user of CS1, the second is the interference from the far user to the near user, the third and the fourth is the interference from other helpers to the near user of CS1, and the last is noise.

When meeting the users’ QoS, the users can correctly decode the requested files. The strategy target of allocating delivery power is to meet the QoS requirement of the far user first, and the remaining power is fully distributed to the near user. According to Shannon’s capacity formula, the rate of the far user decoding the file is:

\[
R_{l,\text{far}} = \log_2 \left( 1 + \frac{\beta_{l,\text{far}} d_{l,\text{far}} P_C}{\beta_{l,\text{near}} d_{l,\text{near}} P_C + I_{\text{inter}} P_C + n_0} \right),
\]

where $\beta_{l,\text{far}}$ is the power of the far user of CS1, $d_{l,\text{far}}$ is the path loss of the far user of CS1, and $P_C$ is the transmit power of the helpers. Hence, $I_{\text{inter}} = \sum_{q \in \Phi \setminus C \setminus I} |h_{q,\text{far}}|^2 d_{q,\text{far}}$ which is the intergroup interference of NOMA, where $d_{q,\text{far}}$ is the path loss of the far user of CS1 to CS$q$. The far user can successfully decode the requested file when the rate is no less than its QoS $R_{l,\text{far}}^{QoS}$. That is

\[
\log_2 \left( 1 + \frac{\beta_{l,\text{far}} d_{l,\text{far}} P_C}{\beta_{l,\text{near}} d_{l,\text{near}} P_C + I_{\text{inter}} P_C + n_0} \right) \geq \log_2 \left( 1 + SNR_{l,\text{far}}^{QoS} \right).
\]

Therefore, the power allocation factor $\beta_{l,\text{far}}$ for the far user should satisfy

\[
\beta_{l,\text{far}} \geq \frac{\text{SINR}_{l,\text{far}} d_{l,\text{far}} P_C + \text{SINR}_{l,\text{far}} n_0}{(1 + \text{SNR}_{l,\text{far}}^{QoS}) d_{l,\text{far}} P_C} + \frac{\text{SINR}_{l,\text{far}} I_{\text{inter}} P_C}{(1 + \text{SNR}_{l,\text{far}}^{QoS}) d_{l,\text{far}} P_C}.
\]

Since $\beta_{l,\text{far}} + \beta_{l,\text{near}} = 1$, we obtain

\[
\beta_{l,\text{far}} = \min \left\{ 1, \frac{\text{SINR}_{l,\text{far}} d_{l,\text{far}} P_C + \text{SINR}_{l,\text{far}} n_0}{(1 + \text{SNR}_{l,\text{far}}^{QoS}) d_{l,\text{far}} P_C} + \frac{\text{SINR}_{l,\text{far}} I_{\text{inter}} P_C}{(1 + \text{SNR}_{l,\text{far}}^{QoS}) d_{l,\text{far}} P_C} \right\}.
\]

In addition, the power allocation factor of the near user is

\[
\beta_{l,\text{near}} = 1 - \beta_{l,\text{far}}.
\]

According to the aforementioned presentation, the proposed QoS-oriented dynamic power allocation algorithm for NOMA-WCN is summarized in Algorithm 1.
Algorithm 1 The Proposed QoS-Oriented Dynamic Power Allocation Algorithm for NOMA-WCN

**Input:** the total number of files $F_{total}$, the popularity parameter $\alpha$, the total power of BS $P_s$, the total power of the helpers $P_c$ and the QoS of files and users;

**Output:** The powers of files $\alpha_1, \alpha_2$ and the power of users $\beta_{l,near}, \beta_{l,far}$;

1: BS generates the popularity of files $G(f)$ according to (1) and downloads the most popular three files;
2: BS allocates the power of the most popular file according to (7);
3: Then, BS allocates the power of the second most popular file according to (9);
4: The remaining power is given to the third file;
5: BS superimposes the three files together and sends them to the helpers;
6: The helpers allocate the power to the far users according to (15), and the remaining power is fully distributed to the near users;

**IV. SIMULATION RESULTS**

In this section, we evaluate the proposed power allocation method in NOMA-WCN during content placement phase and content delivery phase separately. Firstly, we use content hit probability and user outage probability to measure the performance of the proposed power allocation strategy during content placement phase and content delivery phase, respectively. The position of BS is set to the origin of the plane coordinates, and the locations of helpers are modeled as a homogeneous Poisson point process. The content hit probability is the probability that the helpers have the files, which the users request. Since the content hit probability is related to the file popularity and the file outage probability, we can express $CS_l$’s hit probability as

$$\text{Hit}(l) = \sum_{f=1}^{3} G(f)(1 - P_{l,f})$$  \hspace{1cm} (17)

where $G(f)$ represents the popularity of file $f$, $P_{l,f}$ is the outage probability for $CS_l$ to decode file $f$. When the rate of $CS_l$ to decode the file $f$ is less than the file $f$’s QoS requirement, the transmission of file $f$ is regarded as the outage.

In content delivery phase, the user outage probability is the probability that the rate of the user when decoding the requested file is less than the QoS. We set the coverage radius of BS is 50 meters. The total number of helpers is 5. We compare the strategy proposed in this letter with that proposed by Ding et al. [15]. During content placement phase, Ding et al.’s power allocation method is to satisfy the most popular file being correctly decoded by all helpers. The power of the second file and third file is $3/4, 1/4$ of the remaining power, respectively. In content delivery phase, the power allocation of Ding et al. is fixed. The power allocation factor of the near user is $1/4$, and the power allocation factor of the far user is $3/4$.

Fig. 3 shows the outage probability of the cache content during content placement phase. It can be seen from this figure that when the total power is $30$ dBm, the outage probability of the proposed dynamic power allocation method is $28\%$ and $15\%$ lower than that of the method proposed by Ding et al. [15].

Fig. 4 shows the cache hit probability vs. transmit power of BS. The QoS of all three files is $1.8$ bit/channel use (i.e., $R_{QoS}^1 = R_{QoS}^2 = R_{QoS}^3 = 1.8$). Fig. 5 shows the outage probability of near user and far user during content delivery phase. The QoS of the far user and the near user are $1$ bit/channel use and $6$ bit/channel use, respectively. When the total power is $15$ dBm, the outage probability of the far user is reduced by $5\%$ and the
outage probability of the near user is reduced by about 10%. From the simulation results, we can observe that the proposed method can effectively improve the cache hit probability of the WCN and reduce the probability of user outage compared with the state-of-the-art method [15]. Hence, our proposed QoS-oriented dynamic power allocation method can be considered as a candidate technology to deploy in NOMA-WCN.

V. Conclusion

In this letter, we have proposed a QoS-oriented power allocation strategy in content placement phase and content delivery phase for NOMA-WCN. In content placement phase, BS applies NOMA to assign different powers to multiple files for superimposition and transmission. The strategy of the power allocation in this phase is to ensure that the two most popular files can be correctly decoded by the all helpers. In content delivery phase, helpers use NOMA technology to serve two users on the same spectrum at the same time. The goal of power allocation in this phase is to ensure that files of far users can be correctly decoded, reducing user delay and improving spectrum efficiency. Simulation results confirmed that the proposed power allocation strategy improves the cache hit probability and reduces the user outage probability compared with the OMA scheme and a fixed power allocation based NOMA scheme. In future work, we will consider researching the user grouping technology of NOMA in WCN to further increase the cache hit probability and reduce the user outage probability.

ACKNOWLEDGMENT

The authors thank Prof. Zhiguo Ding from the University of Manchester in assistance with reproducing the power allocation method for NOMA-WCN [15] and also many helpful discussions that greatly improved this letter.

REFERENCES

[1] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, “Future intelligent and secure vehicular network toward 6G: Machine-learning approaches,” Proc. IEEE, vol. 108, no. 2, pp. 292–307, Feb. 2020.
[2] Z. Shi, W. Gao, S. Zhang, J. Liu, and N. Kato, “AI-enhanced cooperative spectrum sensing for non-orthogonal multiple access,” IEEE Wireless Commun. Mag., vol. 27, no. 3, pp. 173–179, Apr. 2020.
[3] N. Kato, B. Mao, F. Tang, Y. Kawamoto, and J. Liu, “Ten challenges in advancing machine learning technologies toward 6G,” IEEE Wireless Commun. Mag., vol. 27, no. 3, pp. 96–103, Jun. 2020.
[4] C. Yang, B. Xia, W. Xie, K. Huang, Y. Yao, and Y. Zhao, “Interference cancelation at receivers in cache-enabled wireless networks,” IEEE Trans. Veh. Technol., vol. 67, no. 1, pp. 842–846, Jan. 2018.
[5] 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.
[6] S. H. Chae and W. Choi, “Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching,” IEEE Trans. Wireless Commun., vol. 15, no. 10, pp. 6626–6637, Oct. 2016.
[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] C. Liang, Y. He, F. R. Yu, and N. Zhao, “Enhancing QoE-aware wireless edge caching with software-defined wireless networks,” IEEE Trans. Wireless Commun., vol. 16, no. 10, pp. 6912–6925, Oct. 2017.
[9] G. Gui, M. Liu, F. Tang, N. Kato, and F. Adachi, “6G: Opening new horizons for integration of comfort, security and intelligence,” IEEE Wireless Commun. Mag., early access, Mar. 3, 2020, doi: 10.1109/MWC.001.1900516.
[10] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, “Non-orthogonal multiple access (NOMA) for cellular future radio access,” in Proc. Veh. Technol. Conf. (VTC Spring), Dresden, Germany, Jun. 2013, pp. 1–5.
[11] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, “On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users,” IEEE Signal Process. Lett., vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
[12] Y. Liu, Z. Qin, M. Elkashlan, and Z. Ding, A. Nallanathan, and L. Hanzo, “Nonorthogonal multiple access for 5G and beyond,” Proc. IEEE, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.
[13] M. Liu, J. Yang, and G. Gui, “DSP-NOMA: UAV-assisted emergency communication technology in a heterogeneous Internet of Things,” IEEE Internet Things J., vol. 6, no. 3, pp. 5508–5519, Jun. 2019.
[14] Y. Yin, Y. Peng, M. Liu, J. Yang, and G. Gui, “Dynamic user grouping-based NOMA over Rayleigh fading channels,” IEEE Access, vol. 7, pp. 110964–110971, 2019.
[15] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, “NOMA-assisted wireless caching: Strategies and performance analysis,” IEEE Trans. Commun., vol. 66, no. 10, pp. 4854–4876, Oct. 2018.
[16] Z. Zhao, M. Xu, W. Xie, Z. Ding, and G. K. Karagiannidis, “Coverage performance of NOMA in wireless caching networks,” IEEE Commun. Lett., vol. 22, no. 7, pp. 1458–1461, Jul. 2018.
[17] Y. Fu, W. Wen, Z. Zhao, T. Q. S. Quek, S. Jin, and F. Zheng, "Dynamic power control for NOMA transmissions in wireless caching networks," IEEE Wireless Commun. Lett., vol. 8, no. 5, pp. 1485–1488, Oct. 2019.
[18] K. N. Doan, W. Shin, M. Vaeezi, H. V. Poor, and T. Q. S. Quek, “Optimal power allocation in cache-aided non-orthogonal multiple access systems,” in Proc. Int. Conf. Commun. Workshops (ICC Workshops), Kansas City, MO, USA, May 2018, pp. 1–6.
[19] K. N. Doan, M. Vaeezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, “Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches,” IEEE Trans. Commun., vol. 68, no. 1, pp. 630–644, Jan. 2020.
[20] S. Gurugopinath, P. C. Sofotasios, Y. Al-Hammadi, and S. Muhaidat, “Cache-aided non-orthogonal multiple access for 5G-enabled vehicular networks,” IEEE Trans. Veh. Technol., vol. 68, no. 9, pp. 8359–8371, Sep. 2019.
[21] D. Liu and C. Yang, “Caching at base stations with heterogeneous user demands and spatial locality,” IEEE Trans. Commun., vol. 67, no. 2, pp. 1554–1569, Feb. 2019.