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

Yue Yin [1], Miao Liu [1], Haris Gacanin [1], and Fumiyuki Adachi [1]

[1]Affiliation not available

October 30, 2023

Abstract

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

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

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

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

***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 the base station into many content servers [4]–[6]. In a typical communication system, the base station 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 content servers during the content stack phase before it is requested. Consequently the users can be served locally. Some researchers studied the content placement of wireless caching, for example, S. H. Chae, *et al.* [7] control cache-based channel selection diversity with sophisticated considerations of wireless fading channels and interactions among multiple users. J. Kwak, *et al.* [8] 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, C. Liang, *et al.* [9] 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 [10]. NOMA enables different users to occupy the same spectrum, time, space through power multiplexing, further improves spectrum utilization and reduces users' time delay [11]–[14]. 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, Y. Yin *et al.* proposed a dynamic user grouping strategy to decrease the bit error rate [15]. Z. Ding, *et al.* [16] firstly introduced NOMA into WCN, where a fixed power allocation in the caching deliver phase, thus the allocated power might be unable to satisfy the users' QoS demands dynamically. Z. Zhao, *et al.* [17] studied the coverage performance of NOMA-WCN while did not consider about power allocation scheme. To solve this problem, Y. Fu, *et al.* [18] studied the dynamic power allocation driven by deep neural network of WCN based on NOMA. However, this method requires a lot of data training while it is hard to obtain in the realistic systems.

In this paper, we propose a dynamic power allocation strategy for NOMA-WCN based on the QoS orientations of both files and users. A closed form solution of QoS-oriented power allocation is derived to reduce deployment costs for both base stations and content servers. Hence, our proposed QoS-oriented dynamic power allocation method 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 in the two independent time phases, namely content stack phase and content deliver phase as shown in Fig. 1. Firstly, the base station predicts

This work was supported by the Project Funded by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant TC190A3WZ-2, the Jiangsu Specially Appointed Professor under Grant RK002STP16001, the Innovation and Entrepreneurship of Jiangsu High-level Talent under Grant CZ0010617002, the Six Top Talents Program of Jiangsu under Grant XYDXX-010, the 1311 Talent Plan of Nanjing University of Posts and Telecommunications. *(Corresponding author: Guan Gui)*

Y. Yin, M. Liu and G. Gui are with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (E-mails: {1018010407, liumiao, guiguan}@njupt.edu.cn)

H. Gacanin is with Nokia Bell Labs, 2018 Antwerp, Belgium (e-mail: harisg@ieee.org)

F. Adachi is the Research Organization of Electrical Communication, Tohoku University, Sendai 980-8577 Japan (e-mail: adachi@ecei.tohoku.ac.jp)

which files are popular for users in the next period, and then downloads these files from the core network. During the content stack phase, the base station transmits these files to all content servers. At this phase, the base station and the content servers are unable to serve the users, resulting in increasing their time delays. Therefore, the content stack phase should be short. However, if the content stack phase is too short, content servers will not have enough time to download enough files, so we apply NOMA to handle this problem. Then in the content deliver phase, when the users request files from the content servers, if the files exist in their corresponding content servers, the content servers will directly send the files to users. If the content server does not stack the files which are requested by users in advance, the base station will serve users directly. However, base station will increase the load of the backhaul link and cause the network congestion. This case is not considered in this paper.

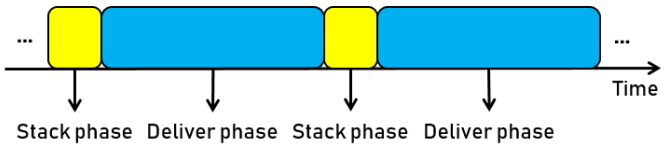

Fig. 1. Time-phase diagram of content stack and content deliver in a WCN. The content stack phase and the content deliver phase are independent of each other. During the content stack phase, the content server cannot serve users. This period is set short to avoid increasing user time delay.

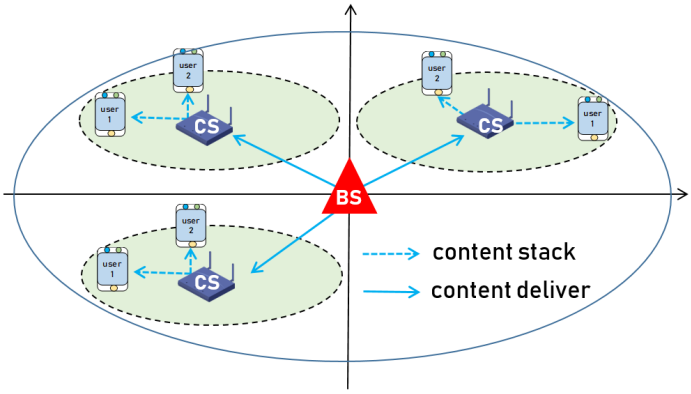

Fig. 2. System model of the WCN, BS refers to the base station, CS refers to the content server. There are multiple content servers within the coverage area of the base station. Each content server serves multiple users.

We consider C content servers distributed within the coverage of a base station, and each content server serves multiple mobile users within its deliver range, as shown in Fig. 2. Moreover, it is assumed that each user is only associated with the nearest content server. In the system model, the position of the base station is set to the origin of the plane coordinates, and the content servers follow a homogeneous Poisson distribution with the base station as the center. The popularity of the files are modeled by the Zipf distribution [19]. Zipf's law has proved that if the words in the dictionary are sorted according to the frequency of their occurrence, the frequency of each word is inversely proportional to its order number,

$$G(r) = \frac{A}{r^{\varepsilon}}, \tag{1}$$

where r represents the frequency order of the words. $G(r)$ is the frequency of occurrence of the r-th word. A and ε are constants, where A is approximately equal to 0.1 and ε is the popularity parameter. Assuming the total number of all the requested files is F_{total}, thus the popularity of file f is computed as:

$$G(f) = \frac{\frac{1}{f^{\varepsilon}}}{\sum_{i=1}^{F_{total}} \frac{1}{i^{\varepsilon}}} \tag{2}$$

The larger the value of ε, the more popular the top-ranked files are.

III. The Proposed QoS-Oriented Dynamic Power Allocation Algorithm

In this section, we present a QoS-oriented dynamic power allocation method in the content stack phase and the content deliver phase.

A. In Content Stack Phase

Base station predicts files which are the most popular ones in the next period, and downloads these files from the core network. During the content stacking phase, the base station send these files to the content servers. All files are allocated with different powers by the NOMA in the base station. 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 content servers. In addition, if the content server is close to the base station, channel conditions is better and it is easier to transmit the files successfully to the content servers. Therefore, the closer the content server is to the base station, the more files can be decoded correctly.

In this paper, we only consider large-scale fading channel, and the path loss channel model from CS_l to base station is:

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

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

$$d_l = \frac{1}{\left(\sqrt{x_l{}^2 + y_l{}^2}\right)^3} \tag{4}$$

The base station sends a superimposed signal of 3 popular files to content servers at one time slot. And it is expected that at least the two most popular files can be correctly decoded by all content servers. The rate for CS_l to decode the most popular file f_1 is:

$$R_{l,1} = \log_2\left(1 + \frac{d_l P_s \alpha_1}{d_l P_s (\alpha_2 + \alpha_3) + n_0}\right) \tag{5}$$

where P_s is the total power sent by the base station, α_1, α_2, and α_3 are the power allocation factors of files f_1, f_2, and f_3, respectively, and n_0 is noise power.

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}. CS_l is able to successfully decode file f_1 only when $R_{l,1} \geq R_1^{\mathrm{QoS}}$,

$$\log_2\left(1+\frac{d_l P_s \alpha_1}{d_l P_s\left(\alpha_2+\alpha_3\right)+n_0}\right) \geq \log_2\left(1+SNR_1^{\mathrm{QoS}}\right) \tag{6}$$

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:

$$\alpha_1 \geq \frac{d_l P_s SNR_1^{\mathrm{QoS}} + n_0 SNR_1^{\mathrm{QoS}}}{\left(SNR_1^{\mathrm{QoS}}+1\right) d_l P_s}. \tag{7}$$

Then, α_1 can be determined by:

$$\alpha_1 = \min\left\{1, \frac{d_l P_s SNR_1^{\mathrm{QoS}} + n_0 SNR_1^{\mathrm{QoS}}}{\left(SNR_1^{\mathrm{QoS}}+1\right) d_l P_s}\right\}. \tag{8}$$

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

$$R_{l,2} = \log_2\left(1+\frac{d_l P_s \alpha_2}{d_l P_s \alpha_3 + n_0}\right). \tag{9}$$

Hence, the power allocation factor α_2 of file f_2 can be further calculated as:

$$\alpha_2 = \min\left\{1-\alpha_1, \frac{d_l P_s SNR_2^{\mathrm{QoS}}\left(1-\alpha_1\right) + n_0 SNR_2^{\mathrm{QoS}}}{\left(SNR_2^{\mathrm{QoS}}+1\right) d_l P_s}\right\}. \tag{10}$$

Finally, file f_3 is decoded by the content server as far as possible. Thus the power allocation factor α_3 can be taken as:

$$\alpha_3 = \max\left\{1-\alpha_1-\alpha_2, 0\right\}. \tag{11}$$

B. In Content Deliver Phase

During the content deliver phase, the users request files from their content servers. 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 who is near the content server is able to use SIC for separating the useful signal. In detail, the superimposed signal received by the near user of CS_l is

$$\begin{aligned} y_{l,near} = {} & h_{l,near}\sqrt{d_{l,near}\beta_{l,near}} f_{l,near} \\ & + h_{l,near}\sqrt{d_{l,near}\beta_{l,far}} f_{l,far} \\ & + \sum_{q\in\Phi C\backslash l} h_{q,lnear}\sqrt{d_{q,lnear}\beta_{q,near}} f_{q,near} \\ & + \sum_{q\in\Phi C\backslash l} h_{q,lnear}\sqrt{d_{q,lnear}\beta_{q,far}} f_{q,far} \\ & + n_{l,near}, \end{aligned} \tag{12}$$

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

$$R_{l,far} = \log_2\left(1+\frac{\beta_{l,far} d_{l,far} P_C}{\beta_{l,near} d_{l,far} P_C + I_{inter} P_C + n_0}\right) \tag{13}$$

where $\beta_{l,far}$ is the power of the far user of CS_l, $d_{l,far}$ is the path loss of the far user of CS_l, and P_C is the transmit power of the content servers. Hence, $I_{inter} = \sum_{q\in\Phi C\backslash l}|h_{q,lfar}|^2 d_{q,lfar}$ which is the intergroup interference of NOMA, where $d_{q,lfar}$ is the path loss of the far user of CS_l to CS_q.

The far user can successfully decode the requested file when the rate is no less than its QoS R_{far}^{QoS}. That is

$$\begin{aligned} \log_2\left(1+\frac{\beta_{l,far} d_{l,far} P_C}{\beta_{l,near} d_{l,far} P_C + I_{inter} P_C + n_0}\right) \\ \geq \log_2\left(1+SNR_{far}^{QoS}\right). \end{aligned} \tag{14}$$

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

$$\begin{aligned} \beta_{l,far} \geq {} & \frac{SINR_{far}^{QoS} d_{l,far} P_C + SINR_{far}^{QoS} n_0}{\left(1+SNR_{far}^{QoS}\right) d_{l,far} P_C} \\ & + \frac{SINR_{far}^{QoS} I_{\mathrm{inter}} P_C}{\left(1+SNR_{far}^{QoS}\right) d_{l,far} P_C}. \end{aligned} \tag{15}$$

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

$$\beta_{l,far} = \min\left\{\begin{array}{l} 1, \frac{SINR_{far}^{QoS} d_{l,far} P_C + SINR_{far}^{QoS} n_{l,far}}{\left(1+SNR_{far}^{QoS}\right) d_{l,far} P_C} \\ + \frac{SINR_{far}^{QoS} I_{\mathrm{inter}} P_C}{\left(1+SNR_{far}^{QoS}\right) d_{l,far} P_C} \end{array}\right\}. \tag{16}$$

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

$$\beta_{l,near} = 1 - \beta_{l,far}. \tag{17}$$

According to the aforementioned presention, 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 α, the total power of base station P_s, the total power of the content servers P_c and the QoS of files and users;
Output: The powers of files and users;
1: Base station generates the popularity of files $G(f)$ according to (2) and downloads the most popular three files;
2: Base station allocates the power of the most popular file according to (8);
3: Then, base station allocates the power of the second most popular file according to (10);
4: The remaining power is given to the third file;
5: Base station superimposes the three files together and sends them to the content servers;
6: The content servers allocate the power to the far users according to (16), 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 the content stack phase and content deliver phase separately. Firstly, we use the content hit probability and user outage probability to measure the performance of the proposed power allocation strategy during the content stack phase and the content delivery phase, respectively. The content hit probability is the probability that the content servers 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

$$Hit\,(l) = \sum_{f=1}^{3} G\,(f)\,(1 - P_{l,f}) \tag{18}$$

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 the content deliver phase, the user outage probability is the probability that the rate the user decodes the requested file is less than the QoS requirement. We set the coverage radius of the base station is 50 meters. The total number of files in the file library is 5, and the popularity parameter ε is 0.5. The total number of content servers is 5. We compare the strategy proposed in this paper with that proposed by Ding *et al.* [16]. During the content stack phase, Ding *et al.*'s power allocation method is to satisfy the most popular file being correctly decoded by all content servers. The power of the second file and third file is 3/4, 1/4 of the remaining power, respectively. In the content deliver phase, the power allocation in paper [16] 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 during the content stack phase. It can be seen from this figure that when the total power is 30 dBm, the outage probability of our proposed power allocation method is 28% and 15% lower than that of the method proposed by Z. Ding *et al.* [16]. In

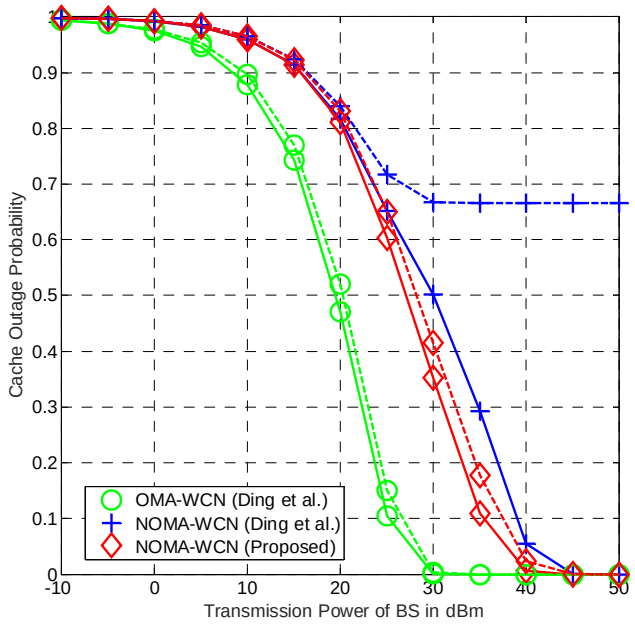

Fig. 3. Cache outage probability vs. transmit power of BS. The dotted line denotes the outage probability when the QoS of all three files is 2 bit per channel use (i.e., $R_1 = R_2 = R_3 = 2$). The solid line denotes the probability of outage when the QoS of all three files is 1.8 bit per channel use (i.e., $R_1 = R_2 = R_3 = 2$).

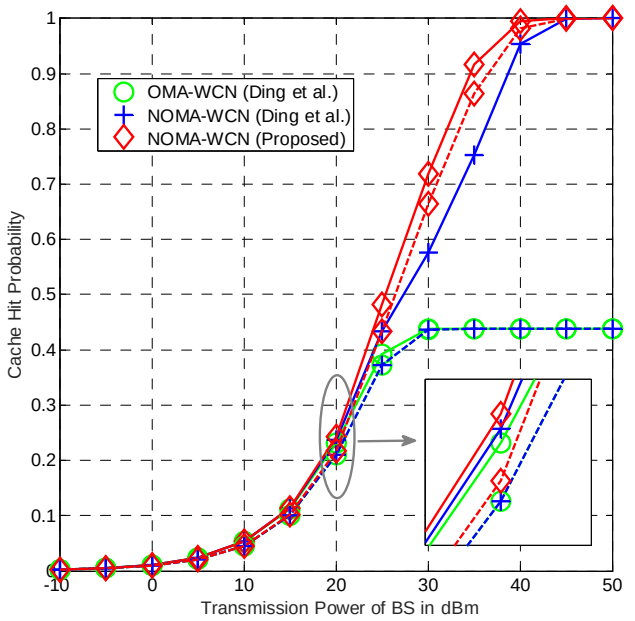

Fig. 4. Cache hit probability vs. transmit power of BS. The dotted line denotes the outage probability when the QoS of all three files is 2 bit per channel use (i.e., $R_1 = R_2 = R_3 = 2$). The solid line denotes the probability of outage when the QoS of all three files is 1.8 bit per channel use (i.e., $R_1 = R_2 = R_3 = 2$).

Fig. 4, the popularity parameter is 0.5 and the total number of file library files is 5, and the cache hit probability is increased by about 20% and 28%, respectively. Fig. 5 shows the outage probability of the near user and the far user during the cache deliver phase. The QoS requirements of the far user and the near user are 1 bit per channel use and 6 bit per 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 [16]. Hence, our proposed QoS-oriented dynamic power allocation method can be considered as a candidate technology to deploy in NOMA-WCN.

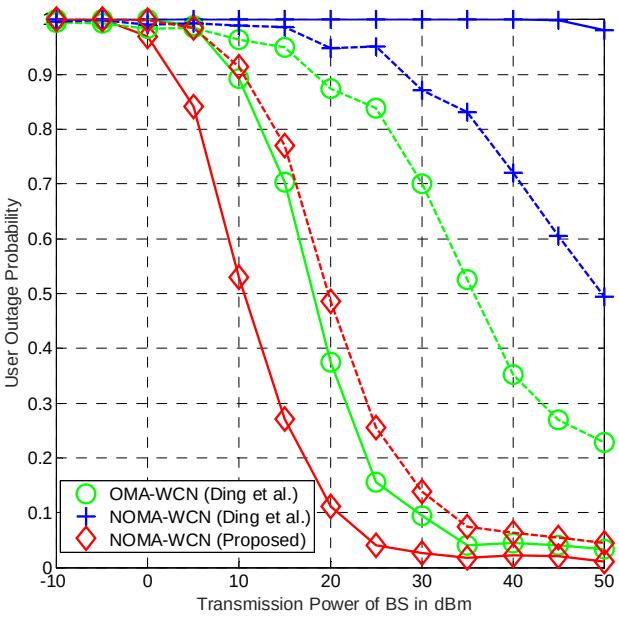

Fig. 5. User outage probability vs. transmit power. The QoS of far user is set as 2 bit/channel use and the QoS of near user is set as 7 bit/channel use. The solid line denotes the probability of outage of the far user. The dashed line denotes the probability of outage of the near user.

V. Conclusion

In this paper, we have proposed a QoS-oriented power allocation strategy in content stack phase and content deliver phase for NOMA-WCN. In the content stack phase, the base station 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 content servers. In the content deliver phase, content servers 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.

Acknowledgement

We thank Prof. Zhiguo Ding from The University of Manchester in assistance with reproducing the power allocation method for NOMA-WCN [16] and also many helpful discussions that greatly improved the manuscript.

References

[1] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network towards 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.

[2] Z. Shi, W. Gao, S. Zhang, J. Liu, N. Kato, "AI-enhanced cooperative spectrum sensing for non-orthogonal multiple access," *IEEE Wireless Commun. Mag.*, in press, doi: 10.1109/MNET.001.1900305

[3] N. Kato, B. Mao, F. Tang, Y. Kawamoto, and J. Liu, "Ten challenges in advancing machine learning technologies towards 6G," *IEEE Wireless Commun. Mag.*, in press, doi: 10.1109/MNET.001.1900476

[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] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, "UAV-relaying-assisted secure transmission with caching," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3140–3153, May 2019.

[6] N. Zhao, F. Cheng, F. Yu, J. Tang, Yu. Chen, G. Gui, and H. Sari, "Caching UAV assisted secure transmission in hyper-dense networks based on interference alignment," *IEEE Trans. Commun.*, pp. 2281–2294, vol. 66, no. 5, May 2018.

[7] 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.

[8] 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.

[9] 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.

[10] 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.*, in press, doi: 10.1109/MWC.001.1900516.

[11] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-oOrthogonal multiple access (NOMA) for cellular future radio access," *Proc. VTC 2013-Spring*, Dresden, Germany, 2-5 June 2013, pp. 1-5.

[12] 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.

[13] Y. Liu, Z. Qin, M. Elkashlan, and Z, Ding, A. Nallanathan, L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Feb. 2017.

[14] M. Liu, J. Yang and G. Gui, "DSF-NOMA: UAV-assisted emergency communication technology in a heterogeneous internet of things," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5508–5519, June 2019.

[15] Y. Yin, Y. Peng, M. Liu, J. Yang, and G. Gui, "Dynamic user grouping-based NOMA over Rayleigh fading channels," *IEEE Access*, vol. 7, no. 1, pp. 110964–110971, 2019.

[16] 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.

[17] 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, July 2018.

[18] 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.

[19] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.