IET Communications

Research Article

Journals
The Institution of
Engineering and Technology

Multicast scheduling for delay-energy trade-off under bursty request arrivals in cellular networks

ISSN 1751-8628
Received on 30th December 2018
Revised 20th March 2019
Accepted on 2nd April 2019
E-First on 15th May 2019
doi: 10.1049/iet-com.2018.6286
www.ietdl.org

Yifan Zhou[1] ✉, *Zhifeng Zhao*[1], *Chen Qi*[1], *Rongpeng Li*[1], *Yves Louet*[2], *Jacques Palicot*[2], *Honggang Zhang*[1]

[1]*College of Information Science and Electronic Engineering, Zhejiang University, Zheda Road 38, Hangzhou, People's Republic of China*

[2]*Université Européenne de Bretagne & CentraleSupélec & IETR, Avenue de la Boulaie, Cesson-Sévigné Cedex, France*

✉ *E-mail: zhouyftt@zju.edu.cn*

Abstract: In this study, the authors consider the utilisation of multicast technology in cellular networks given different arrival patterns for the content requests of mobile users. Traditionally, the performance evaluation of multicast in the literature usually depends on the adoption of temporal Poisson processes for content requests, which is not accurate any more according to many real data measurements. Therefore, to make use of the bursty nature of content requests, they propose a hybrid unicast/multicast strategy where the base station (BS) can perform the unicast or multicast procedure according to its serving status. By modelling the complete process into a circular Markov chain, they derive the average latency of content requests and the average power consumption of BSs under different arrival patterns and serving configurations in theoretical and/or simulative ways. Moreover, the multicast threshold introduced in their strategy can be dynamically adjusted to achieve a joint optimisation between average latency and power consumption when confronted with varied demands. Numerous results show that the proposed strategy can not only reduce the average latency of content requests but also decrease the average power consumption of BSs, especially under the bursty request arrival patterns.

1 Introduction

According to Shannon's theory, the technical gains brought by the physical layer gradually become saturated, which cannot match the rapid increase of traffic demands in the current mobile internet era [1]. In this study, we devote to make use of the temporal characteristics of user requests in cellular networks, based on which many pertinent technologies have been extensively studied in recent years.

Among those technologies, multicast is considered to improve the throughout and energy efficiency of cellular networks [2]. With multicast, the base station (BS) or server can send the same content to many users within its coverage area, achieving the purpose of one transmission and multiple receptions [3]. Such a strategy not only increases the capacity of entire access networks but also effectively reduced the energy consumption of the transmitter [4]. However, it works at the cost of increasing user delay, because each request needs to wait for the multicast until the number of requests achieves a pre-defined threshold. In practice, multicast can be combined with other promising technologies to further exploit the characteristics of cellular networks for potential capacity improvement.

For example, it can be combined with device to device (D2D) short-distance transmissions [5]. Among user equipments (UE), which request the same content, some of them have better channel qualities and thus can serve as relays. After these relay users obtain the requested content from the BS multicast, they spread the content to all other requesting users through D2D transmissions.

Similarly, network coding can be combined with multicast technology to meet UE diverse rate requirements [6]. Simulations showed that such a hybrid strategy can achieve a throughput improvement of 30–45%. Works in [7, 8] consider the optimal allocation of dynamic multicast in a content-centric caching-enabled wireless network, aiming to minimise user delay and power consumption. Unlike caching contents on the BSs, the authors of [9, 10] consider caching directly on UE and combine it with the macro-BS broadcast. The results show that such a mixed strategy can achieve significant performance gains.

In summary, multicast-related technologies use the clustering nature of content requests in time, space and content domains to improve the overall performance, such as spectrum efficiency and energy consumption. However, the user request patterns adopted by most related works cannot accurately reflect the reality. For example, the arrival pattern of content requests is usually assumed to be a uniform Poisson process for tractability [4], which is inconsistent with the bursty nature claim we made in previous works [11].

To utilise the bursty nature of content requests, in this study, we propose a hybrid strategy combining unicast and multicast techniques to improve the overall performance. Specifically, we try to identify the variation of user's content access latency and BS's average power consumption with respect to different arrival and serving patterns. Furthermore, in order to reach a trade-off between the delay and power consumption, we jointly optimised them in different scenarios to find the optimal multicast threshold in our proposed hybrid strategy.

In detail, we abstract the unicast/multicast process in cellular networks into a queuing model [12], where users' requests for a specific content enter a multicast queue of this content. When the number of requests in the queue is less than a prescribed threshold, the BS performs a conventional unicast, i.e. only one user is served at a time. As new users continue to join, the BS may start a one-to-many multicast process if the number of requests reaches the threshold. Based on this abstraction, the average latency of the mobile user and the average power consumption of the BS as two performance metrics will be theoretically and/or simulatively derived hereafter for different request arrival patterns.

The rest of this paper is organised as follows. Section 2 will deal with the Poisson arrival of content requests in our strategy, focusing on the derivation of average latency. On the other hand, the simulative analysis is conducted for the bursty case in Section 3 including requests' latency and power consumption of BS. To jointly optimise these two performance metrics, we proposed a dynamic multicast threshold strategy in Section 4. After that, the conclusion is given in Section 5.

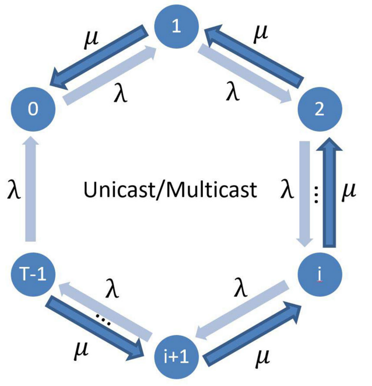

Fig. 1 *Circular Markov chain characterisation of the hybrid strategy*

2 Average latency performance under Poisson arrivals

First, we characterise the arrival and departure process of content requests in the BS's unicast/multicast paradigm into a circular Markov chain as shown in Fig. 1.

This diagram depicts the transition process of a multicast queue (with threshold T), which can be thought as a circular Markov chain, as requests for the content keep on arriving and the BS keeps on performing unicast or multicast. In detail, the multicast queue with T possible states starts from state 0, and the state of the chain represents the number of content requests waiting in the queue. When the first request arrives, the Markov chain goes to state 1. In general, when the multicast queue is in state i $(0 < i < T)$, it has two different ways to transfer, one of which to state $i+1$ due to the arrival of a new request (when $i = T-1$, go directly to state 0), the other to state $i-1$ if the BS finishes one unicast before new request arrives. Actually, the memoryless property of the Poisson process makes the transition of each state independent. The probability of each occurrence in different states depends on the arrival rate λ of content requests and the unicast service rate μ of BS, indicating the number of content requests that arrive or served within 1 ms.

Furthermore, we calculate the average latency, which is defined as the time difference between the content request's arrival and the completion of this request served by the BS. Specifically, this problem can be divided into two separate parts. First, when a content request arrives at a BS, it will be categorised into the service queue of this specific content. Therefore, requests arriving at a different time will encounter the service queue in different states whose probability is also the stable time proportion within the Markov chain (assuming $P_i, 0 \le i < T$). Second, for a random request, assuming that the service queue is in state i when it joins the multicast queue, thus this user is listed in order i of the unicast service. Since this request may be served by unicast or by multicast (triggered when queue length reaches T), we assume that the average waiting time is D_i. Thus, for a random content request, its average latency can be written as

$$D = \sum_{i=0}^{T-2} P_i D_{i+1}, \tag{1}$$

where P_i is the limiting probability of the Markov chain in state i and D_i is the average waiting time of the user who arrives as the ith unicast member. The reason why the index equals $i+1$ here is that the Markov chain status will be shifted from i to $i+1$ after the new request joins. Next, we need to calculate P_i and D_i separately.

2.1 Limiting probabilities of different states

In order to simplify the theoretical derivation of the average latency, we assume that the arrivals of content requests obey the Poisson process. First, we derive the stable distribution of the Markov chain in Fig. 1 based on the content request's arrival rate λ, the BS unicast service rate μ, and the multicast threshold T (as Markov chain with finite states should have limiting probabilities). According to the rate principle, we can obtain the following equilibrium equations:

$$\begin{aligned} P_1(\lambda + \mu) &= P_0\lambda + P_2\mu, \\ P_i(\lambda + \mu) &= P_{i-1}\lambda + P_{i+1}\mu, \\ P_{T-1}(\lambda + \mu) &= P_{T-2}\lambda, \\ P_0\lambda &= P_1\mu + P_{T-1}\lambda. \end{aligned} \tag{2}$$

For example, the left side of the first equation represents the departure rate of state 1 and the right two parts are the arriving rates from states 0 and 2 to state 1, respectively. The same equation also applies to state i, when $0 < i < T-1$. In particular, if the queue is in state $T-1$, then the departure rate is $P_{T-1}(\lambda + \mu)$, and the arrival rate is $P_{T-2}\lambda$. Similarly, for state 0, the departure rate is $P_0\lambda$ and the arrival rate is from state 1 for unicast and from state $T-1$ for multicast.

Combine the above equilibrium equations with the definition that limiting probabilities sum up to 1, we can get the following non-trivial solution when $\lambda \neq \mu$:

$$P_i = c_1\left(\frac{\lambda}{\mu}\right)^i + c_2, \tag{3}$$

where the values of c_1 and c_2 are calculated as follows:

$$c_1 = \frac{1}{\frac{1-(\frac{\lambda}{\mu})^{\mathrm{T}}}{1-\frac{\lambda}{\mu}} - T(\frac{\lambda}{\mu})^{\mathrm{T}}}, c_2 = -c_1\left(\frac{\lambda}{\mu}\right)^{\mathrm{T}}. \tag{4}$$

For $\lambda = \mu$, we can get the following trivial solution:

$$P_i = \frac{2(T-i+1)}{T(T+1)}. \tag{5}$$

To verify the correctness of the derivation, we will next calculate the corresponding theoretical and simulation values with different parameter settings (λ, μ, T).

From Fig. 2 we can see that for $\lambda < \mu$ $(\lambda = 5,\ \mu = 10)$, the limiting probabilities P_i of the Markov chain decrease in i. Since the unicast service rate of the BS is obviously greater than the arrival rate of the content request, most of which are served in time through unicasts, thus the service queue rarely enters the multicast mode. For $\lambda > \mu$ $(\lambda = 15,\ \mu = 10)$, the curves also follow the decreasing pattern, while the rate of decrement is increasing, which is opposite with the former case. In addition, when $\lambda = \mu$, the limit probabilities show a linearly decreasing trend.

2.2 Average waiting time for random request

Since BS's serving procedure is assumed here to be a mixed process of unicast and multicast, the average latency for a user in the service queue is not only related to the queue length, which affects the multicast time, but also related to its unicast order, which determines the unicast time. Thus, we can use a tuple (m, n) to fully describe any user in the service queue, where m represents the unicast order and n means that the queue currently requires another n requests to trigger the multicast, indicating that there are $T - n$ users waiting in line. Accordingly, we hope to derive the recursion pattern and then the general formula of the user's average wait time $W(m, n)$ with parameters (m,n,T) according to the process in Fig. 1.

In detail, each transition in the Markov chain may refer to queuing length plus one due to the arrival of a new request or queuing length minus one due to the unicast of the foremost request. Combined with the (m, n) description, user's unicast order remains (m unchanged) when a new request arrives, while the number of new requests necessary for the queue to trigger multicast become $n-1$, then the tuple becomes $(m, n-1)$. When a unicast is accomplished, the unicast order of the user becomes $m-1$ $(m > 1$, otherwise it will be served immediately), and the

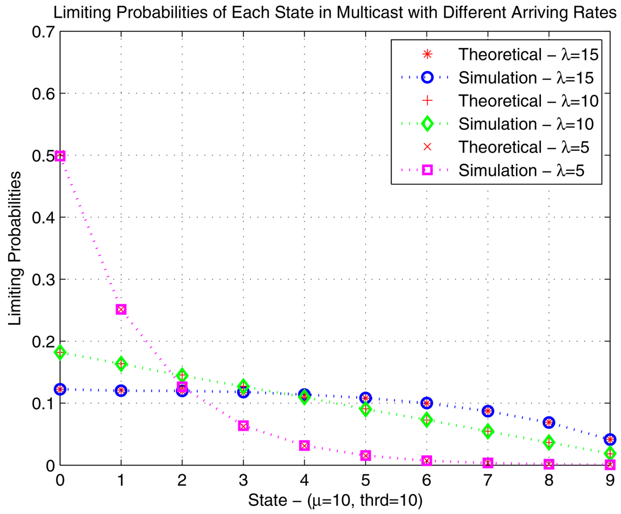

Fig. 2 *Limiting probabilities of the unicast/multicast Markov chain with threshold $T = 10$ under different Poisson arrival rates*

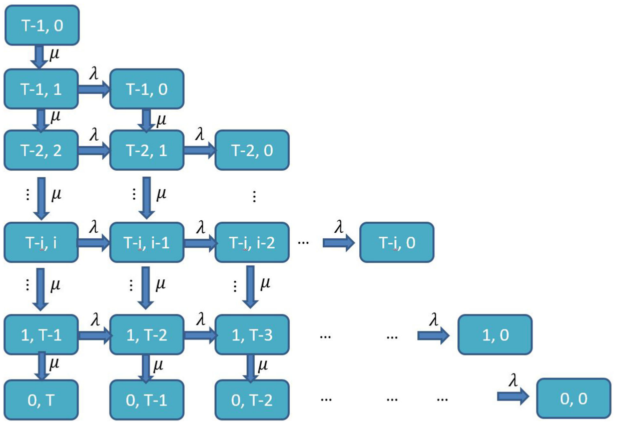

Fig. 3 *Two-dimensional state transition diagram illustrating unicast/ multicast service queueing*

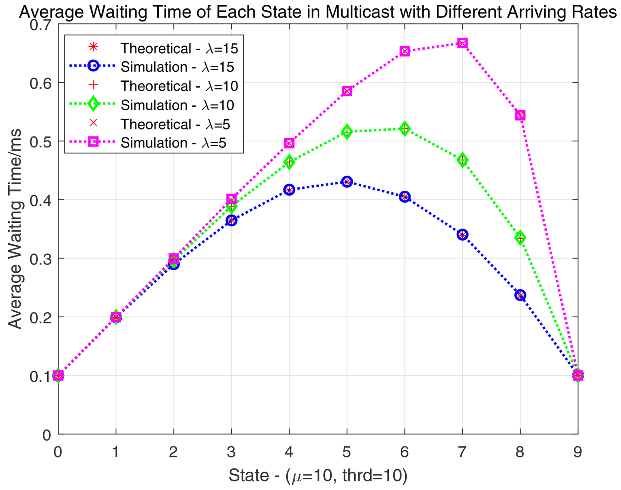

Fig. 4 *Average waiting time for each state in the Markov chain with $T = 10$ under different parameter setups*

number of requests necessary for multicast will be $n + 1$, thus the tuple becomes $(m - 1, n + 1)$. According to the above description, we can obtain a two-dimensional state transition diagram as shown in Fig. 3, from which the following recursion pattern for $W(m, n)$ is derived:

$$W(m,n) = \frac{1}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} W(m, n - 1) + \frac{\mu}{\lambda + \mu} W(m - 1, n + 1)\,. \tag{6}$$

where m and n meet the constrains $m + n \le T$, $m > 0$, and $n > 0$. Besides, the boundary values for $W(m, n)$ satisfy

$$W(m, 0) = \frac{1}{\mu}, \quad W(0, n) = 0. \tag{7}$$

Based on the above recursion formula and boundary conditions, we can derive the general formula of $W(m, n)$

$$W(m,n) = \frac{\lambda^2 + 2\lambda\mu}{\mu(\lambda + \mu)^2} + \frac{1}{\lambda + \mu} \sum_{i=0}^{m-1} \sum_{j=0}^{n+i-1} C_{i+j}^{i} \left(\frac{\mu}{\lambda + \mu}\right)^i \left(\frac{\lambda}{\lambda + \mu}\right)^j, \tag{8}$$

where C_{i+j}^{i} is the combinatorial number with value $(i + j)!/i!j!$. In order to calculate the average waiting time D_{i+1} for the new request in (1), we need to examine the relationship between D_i and $W(m,n)$. According to the previous definition, D_i represents the average waiting time of users who entered the queue with the unicast order i, while $W(m,n)$ represents the average waiting time for users in the m unicast order with queue length $T - n$. Obviously, according to their respective definitions, we can get the following equation:

$$D_m = W(m, T - m)\,. \tag{9}$$

Combining with (8), we can get the expression of D_m

$$D_m = \frac{\lambda^2 + 2\lambda\mu}{\mu(\lambda + \mu)^2} + \frac{1}{\lambda + \mu} \sum_{i=0}^{m-1} \sum_{j=0}^{T-m+i-1} C_{i+j}^{i} \left(\frac{\mu}{\lambda + \mu}\right)^i \left(\frac{\lambda}{\lambda + \mu}\right)^j \tag{10}$$

After combining (3) of P_i with (10) of D_i according to (1), we can derive the average latency of a random user within the hybrid unicast/multicast strategy under Poisson arrivals. In order to verify the correctness of the above theoretical derivation, we also conduct simulation under different parameter configurations.

First of all, we derive D_i from (10) for several combinations of arrival and service rates, and the theoretical value and corresponding simulation curves are shown in Fig. 4.

In Fig. 4, we can see that for different parameter configurations, D_i first increases and then decreases. Specifically, for $\lambda \le \mu$, the increasing rate of D_i when i is small is significantly less than that when i is greater; for $\lambda > \mu$ ($\lambda = 15$, $\mu = 10$), the increasing rate of D_i of small i is comparable to the decreasing rate of large i. This may be due to the fact that when $\lambda \le \mu$, more content requests are served by unicast, so the maximum value of D_i should be obtained within the range of $i > T/2$, thus D_i shows a right shoulder shape. For $\lambda > \mu$, content requests are relatively less unicast served, so the peak of the curve is left shifted.

3 Hybrid strategy analysis under bursty request arrivals

The previous section explains how average latency varies with relevant parameters under Poisson arrivals. However, in fact, users' content request pattern in cellular networks is far from the traditional Poisson assumption, i.e. the inter-arrival time between requests does not obey the exponential distribution but may obey the log-normal distribution as obtained from real data analysis [11]. Therefore, in this section, we will consider the impact of the burstiness of content requests on the average latency and power consumption.

Different from the Poisson arrivals of content requests, the performance metric cannot be theoretically derived under bursty arrivals, such as the limiting probabilities and average waiting time of different states in service queue [13]. The reason is that the Markov chain cannot be expressed as a stable transfer process. Therefore, in the following performance analysis of non-Poisson arrivals, we mainly compare them through numerical simulations.

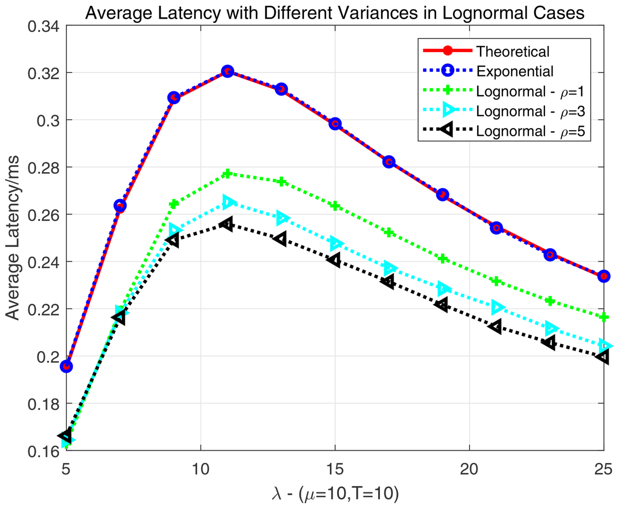

Fig. 5 *Average latency varies with the arriving rates under log-normal distribution with different degrees of aggregation ρ*

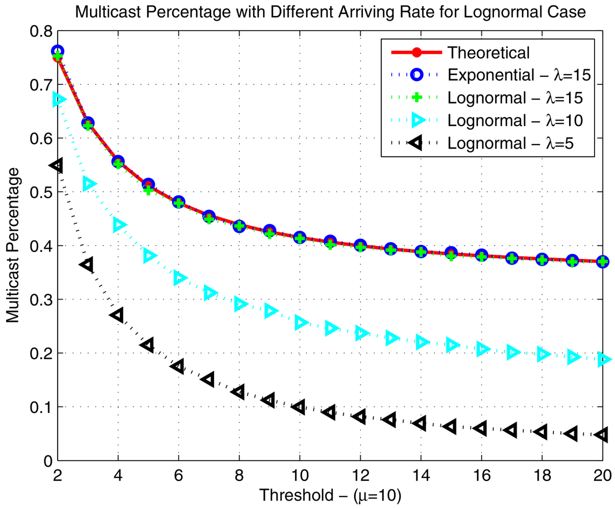

Fig. 6 *Multicast probability varies with the multicast threshold T under log-normal distribution with different average arrival rate λ*

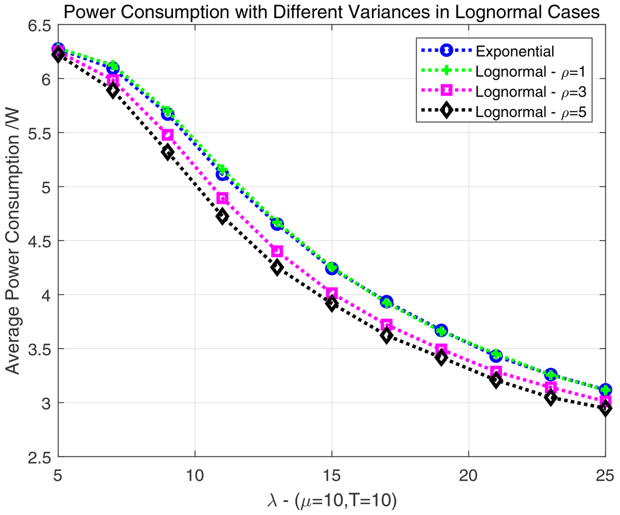

Fig. 7 *Average power consumption varies with the arriving rates under log-normal distribution with different degrees of aggregation ρ*

3.1 Average latency under bursty arrivals

In the exponential case, the mean and standard deviation of the random variable are equal to $1/\lambda$. Therefore, once λ is determined, the mean and variance cannot be adjusted separately. However, in practice, the same number of requests can have a variety of appearances as the arrival pattern changes, which is consistent with the dynamics of mobile content requests. From this point of view, the exponential distribution (Poisson arrival process) does not have the flexibility to describe the temporal pattern of content requests. On the other hand, the log-normal distribution has more flexibility in parameter selection since it can adjust the variance while maintaining the mean value, which is useful to simulate the burst characteristics of request arrivals. Next, we analyse the numerical impact of burstiness on latency performance and energy efficiency for the hybrid strategy by adjusting the ratio of the standard deviation to the mean in log-normal distribution (ρ, which is used to characterise the degree of temporal aggregation of content requests, as defined by (11))

$$\rho = \frac{\sqrt{\mathrm{Var}(t)}}{E(t)}. \tag{11}$$

In Fig. 5, we can see that the average latency of the log-normal distribution with the same degree of aggregation ($\rho = 1$) is smaller than that of the exponential distribution, and both of them show an increase-then-decrease tendency with respect to λ. While among different log-normal distributions, it shows that the greater ρ, the smaller the average latency. A possible explanation is that the increase in ρ indicates greater variance in the arrival time interval for the same number of requests, which results in more burstiness, thus the requests in the congested state is mostly served by multicast, while the requests in the idle state are served through unicast and the superposition of these two cases degrades the overall average latency.

3.2 Average power consumption of BS under bursty arrivals

In order to examine the variation of the average power consumption of the BS with the degree of aggregation ρ under bursty arrivals, the multicast probability of the service queue should be analysed firstly.

In fact, the average power consumption of the BSs and the multicast probability of the hybrid strategy are highly related. If one request is unicast served, the amount of power it consumes (assumed to be random variable W_1) is determined by the channel conditions and the user's distance from the BS. If this request is being served by multicast, then the BS consumes the maximum power ($W_{\max}$) required by T users in this multicast queue, and every single user only consumes $W_{\max}/T$ of the power. Therefore, combined with the multicast probability M_T, the average power consumption of the BS for one single request can be written as

$$W = (1 - M_T)W_1 + M_T\frac{W_{\max}}{T}. \tag{12}$$

In Fig. 6, the log-normal distribution and exponential distribution with the same ρ have no significant difference in the multicast probability. Further in Fig. 7, as the arrival rate and multicast probability increases, the BS can serve most users with less power through more multicast and the average power consumption also decreases as ρ increases.

4 Joint optimisation of average latency and power consumption

Generally, as T increases, no matter the requests are exponentially or log-normally arrived, the average latency always increases while the average power consumption decreases mostly as in Figs. 8 and 9. From this point of view, we can make a trade-off between these two metrics by jointly examining the average latency of request and the average power consumption of the BS, and perform the optimisation of $D + \epsilon W$ on the selection of T. As shown in Fig. 10, we show the joint metric curves with respect to T for the log-normal distributed inter-arrival time with different arrival rates given $\epsilon = 1$. The three curves show a decreasing-then-increasing trend except that the $\lambda = 5$ curve keeps rising. As a result, each curve has the lowest point minimising the target metric.

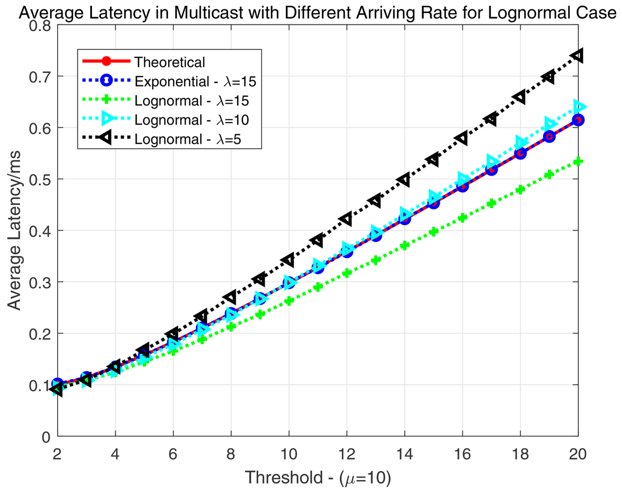

Fig. 8 *Average latency varies with the multicast threshold T under log-normal distribution with different average arrival rate λ*

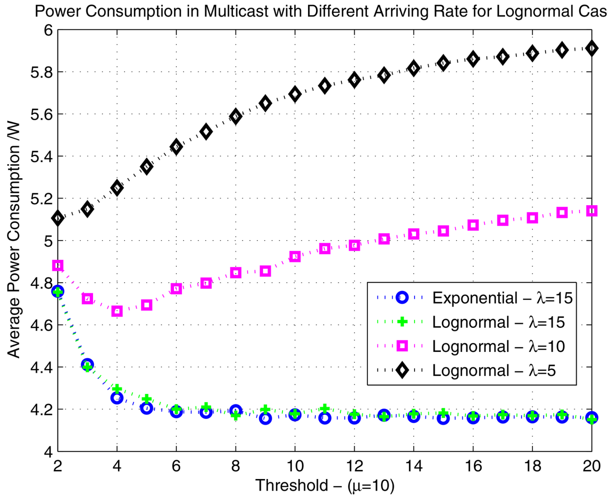

Fig. 9 *Average power consumption varies with the multicast threshold T under log-normal distribution with different average arrival rate λ*

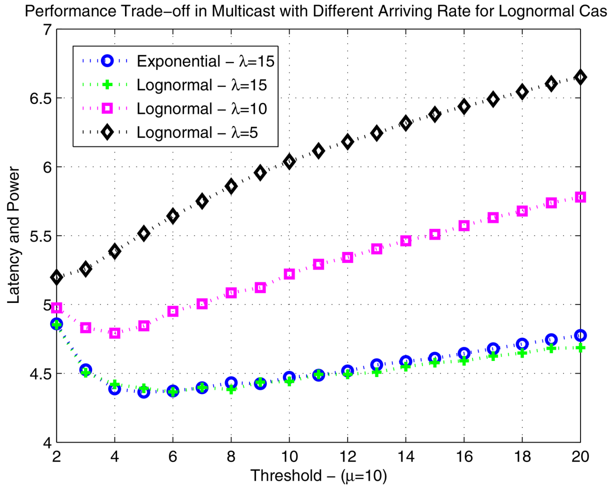

Fig. 10 *Latency and power trade-off with multicast threshold T under different arrival rates for log-normal distributed inter-arrival time*

Specifically, curves continue to decrease as λ increases, and the optimal T value also gradually increases since the average power consumption decreases with T, thereby reducing the joint metric. The reason why different curves show a similarly decreasing-then-increasing trend is the same with the explanation of Fig. 11.

In order to examine the impact of coefficient ϵ on the joint optimisation results, we depict the variation of the optimal threshold T with ϵ in (0,1) for different arrival rates in Fig. 12. It can be seen that regardless of the specific arrival interval, the optimal T value of the joint optimisation shows a stair-like upward trend with increasing ϵ except for the $\lambda = 5$ case. For example, the optimal value in the $\lambda = 15$ case of a log-normal distribution increases gradually from $T = 2$ in $\epsilon = 0$ to $T = 6$ when $\epsilon = 1$. The possible explanation is that as ϵ increases, a larger proportion in the joint optimisation lies on the average power consumption, which generally shows a decreasing-then-increasing trend with respect to T. In order to minimise the joint metric, T needs to be around the lowest point of the average power consumption curve. Therefore, as the value of ϵ increases, the optimal threshold T will keep increasing at first. However, once the value of T reaches the lowest point, it remains constant since both the average latency and power consumption will increase as T increases.

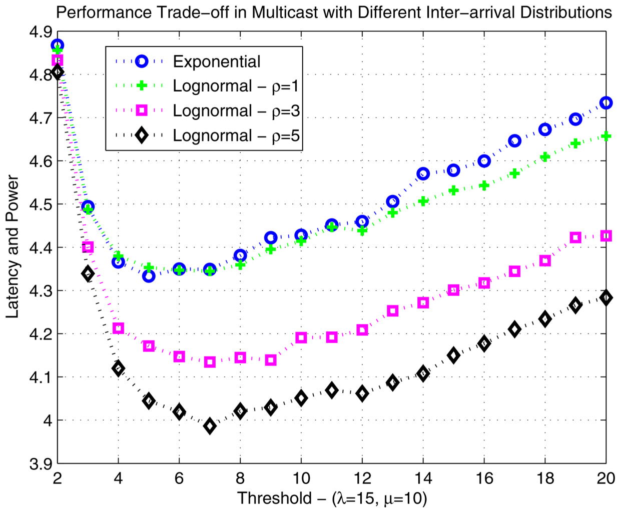

Fig. 11 *Latency and power trade-offs with multicast threshold T under different degrees of aggregation for log-normal distributed inter-arrival time*

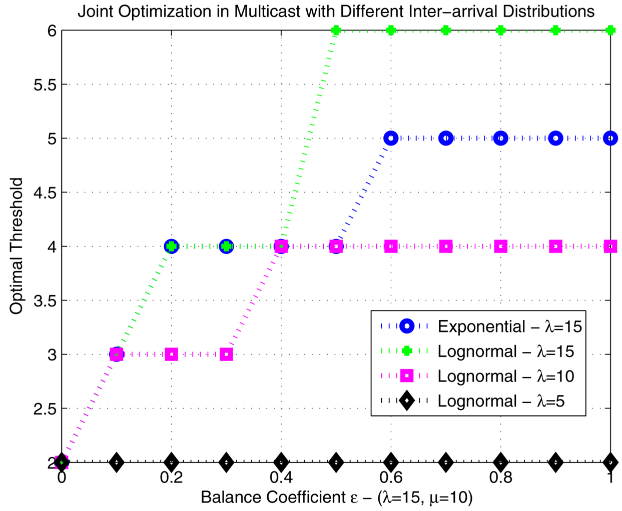

Fig. 12 *Optimal multicast thresholds for joint optimisation of latency and power under different arrival rates for log-normal distributed inter-arrival time*

Furthermore, besides the arrival rate, we analysed the effect of the degree of aggregation ρ on the joint optimisation performance. As shown in Fig. 11, we show the variation curve of the joint metric with a multicast threshold for a different ρ value given $\epsilon = 1$, which is also a decreasing-then-increasing trend. The possible explanation is that when T is small, the multicast effect is significant and the average power consumption decreases rapidly with T where the rate of decline exceeds the growth rate of average latency. When T exceeds a certain value, the decreasing of multicast probability makes the average power consumption no longer significantly drops. On the other hand, the average latency will play a dominant role in the joint metric variation, thus the segment increases with T. Through comparison, it is found that different inter-arrival time corresponds to different optimal multicast thresholds even with the same arrival rate and service rate. For example, in Fig. 11, the optimal value of the exponential case is $T = 5$, while the optimal T values of the other three log-normal curves increase with ρ. In addition, we can obtain different optimal multicast threshold T when choosing different ϵ values to jointly optimise the average latency and power consumption.

5 Conclusion

In this study, we analyse the average latency of content request and the average power consumption of the BS with Poisson or bursty request arrivals in the proposed hybrid unicast/multicast strategy. Firstly, under the traditional Poisson arrivals, we use the Markov chain model to describe the service procedure mathematically and calculate the corresponding limit probabilities, the average waiting time of each state and the overall average latency by both theoretical derivation and simulation verification. Secondly, for the bursty arrival case, since there is no closed-form solution, we use a numerical method to compare the average latency, the multicast probability and the average power consumption for different inter-arrival distributions. Furthermore, in order to comprehensively balance the performance of latency and power consumption, we have jointly optimised them in both scenarios to find the optimal multicast thresholds minimising the weighted sum of these two metrics. According to the results, we found that the adoption of the hybrid strategy can not only solve the problem that the average latency increases infinitely with the arrival rate under congestion but also being more effective in the practical bursty arrival scenario (less latency compared to Poisson arrivals with the same arrival rate). For future direction, theoretical analysis is necessary to unveil the impact of bursty arrival patterns on the overall performance, and some approximation techniques may be helpful since the closed-form derivation is not applicable.

6 Acknowledgments

This work was supported in part by the National Key R&D Program of China (no. 2018YFB0803702), the National Natural Science Foundation of China (nos. 61701439 and 61731002), the Zhejiang Key Research and Development Plan (no. 2018C03056), the National Postdoctoral Program for Innovative Talents of China (no. BX201600133), and the project funded by the China Postdoctoral Science Foundation (no. 2017M610369).

7 References

[1] Cisco visual network index: Global mobile data traffic forecast update, 2015–2020 white paper', 2016. Available at http://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] Araniti, G., Condoluci, M., Scopelliti, P., *et al.*: 'Multicasting over emerging 5G networks: challenges and perspectives', *IEEE Netw.*, 2017, **31**, (2), pp. 80–89

[3] Hartung, F., Horn, U., Huschke, J., *et al.*: 'MBMS IP multicast/broadcast in 3G networks', *Int. J. Digit. Multimedia Broadcast.*, 2009, **1**, pp. 1–25

[4] Huang, C., Zhang, J., Poor, H.V., *et al.*: 'Delay-energy tradeoff in multicast scheduling for green cellular systems', *IEEE J. Sel. Areas Commun.*, 2016, **34**, (5), pp. 1235–1249

[5] Militano, L., Condoluci, M., Araniti, G., *et al.*: 'Single frequency-based device-to-device enhanced video delivery for evolved multimedia broadcast and multicast services', *IEEE Trans. Broadcast.*, 2015, **62**, (2), pp. 263–278

[6] Tassi, A., Chatzigeorgiou, I., Vukobratovic, D.: 'Resource allocation frameworks for network-coded layered multimedia multicast services', *IEEE J. Sel. Areas Commun.*, 2015, **33**, (2), pp. 141–155

[7] Zhou, B., Cui, Y., Tao, M., *et al.*: 'Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks', *IEEE Trans. Commun.*, 2017, **65**, (7), pp. 2956–2970

[8] Cui, Y., Jiang, D.: 'Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks', *IEEE Trans. Wirel. Commun.*, 2017, **16**, (1), pp. 250–264

[9] Su, C.-J., Tassiulas, L.: 'Joint broadcast scheduling and user's cache management for efficient information delivery', *Wirel. Netw.*, 2000, **6**, (4), pp. 279–288

[10] Poularakis, K., Iosifidis, G., Sourlas, V., *et al.*: 'Exploiting caching and multicast for 5G wireless networks', *IEEE Trans. Wirel. Commun.*, 2016, **15**, (4), pp. 2995–3007

[11] Li, R., Zhao, Z., Qi, C., *et al.*: 'Understanding the traffic nature of mobile instantaneous messaging in cellular networks: a revisiting to α-stable models', *IEEE Access*, 2015, **3**, pp. 1416–1422

[12] Ross, S.M.: '*Introduction to probability models*' (Academic Press, San Diego, USA, 2014)

[13] Taleb, T., Ksentini, A., Chen, M., *et al.*: 'Coping with emerging mobile social media applications through dynamic service function chaining', *IEEE Trans. Wirel. Commun.*, 2016, **15**, (4), pp. 2859–2871