DBS: Delay Based Hierarchical Downlink Scheduling for Real-Time Stream in Cellular Networks

Wenjin Fan(✉), Yu Liu, and Yumei Wang

School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{wjfan,liuy,ymwang}@bupt.edu.cn

Abstract. With the rapid development of cellular networks, demand for real-time stream is increasing dramatically. How to guarantee better quality-of-service (QoS) of real-time stream services under limited resources becomes an increasingly important issue. This paper proposes a delay based hierarchical downlink scheduling (DBS) algorithm for real-time stream in cellular networks. The hierarchical scheduler is divided into two levels. The upper level scheduler offers a prediction of the number of data bits that each real-time stream needs to guarantee the QoS. The lower level scheduler classifies all real-time streams into grade A, B and C according to Head of Line Delay and allocates resources to streams according to their different grades. The simulation results show that our algorithm performs better than other real-time schedulers, such as frame level scheduler (FLS), Modified Largest Weight Delay First (M-LWDF), EXP/PF and EXP-LOG in the aspects of delay and throughput.

Keywords: Downlink schedule · Resource allocation · Stream classify · Cellular network

1 Introduction

The transmission of real-time stream, especially video content, over wireless communications is becoming a major contributor to future Internet application traffic. With the increasing demand for real-time stream, better quality-of-service (QoS) perceived by users will become increasingly important [1].

Such improvement of QoS performance can be achieved with developing up-layer functionality, such as radio resource allocation, and lower-layer functionality, including Orthogonal Frequency-division Multiple Access (OFDMA), Modulation and Coding Schemes (MCSs) and Hybrid Automatic Repeat Request (HARQ), etc. Radio resource allocation is one of most important issue in the above factors. To meet the needs of as many requesters as possible, resources should be allocated with proper scheduling policies. For example, 5G performance goals include to achieve higher data rate, reduced latency, and higher system spectrum efficiency [2], etc. To achieve these targets, in 5G

H. Gao et al. (Eds.): ChinaCom 2019, LNICST 312, pp. 246–257, 2020.
https://doi.org/10.1007/978-3-030-41114-5_19

networks a lot of lower-layer functionality, such as New Radio Frequencies, Massive MIMO, Edge Computing and Small Cell are developed. More importantly, the scheduling algorithm in the high-level functionality should also be improved to achieve better QoS with lower latency and lower packet loss rate [3].

The Media Access Control (MAC) layer carries out the allocation of wireless resources. However, due to the complex and changeable wireless environment and various QoS needs of stream, effective resource allocation is extremely challenging. Some algorithms [3–9] for real-time stream consider different network parameters and achieve real-time stream optimization. More details will be discussed in the next section.

This paper proposes a delay based hierarchical downlink scheduling (DBS) algorithm for real-time stream in cellular networks. The upper level scheduler is the same as that in FLS [5] to offer a predicted number of data that each real-time stream needs to ensure the QoS. Then, the lower level scheduler classifies all real-time streams into grade A, B and C according to $D_{HOL,i}$ (Head of line packet delay), and allocates resources to streams by their different grades. TTI-level delay characteristics are sufficiently applied through the classification to improve Packet Loss Rate (PLR) and throughput.

The remainder of this paper consists of several sections. In Sect. 2, related work is depicted. In Sect. 3, the proposed algorithm is concretely presented. Section 4 describes simulation scenario parameters and discusses the results. Finally, concluding observations are drawn in Sect. 5.

2 Scheduling Policies

In cellular networks, large performance gains can be achieved by properly allocating the number of frequency resources to users at each TTI 5. Due to the high demand of real-time stream for QoS, many QoS-aware scheduling algorithms have been proposed. Related meaning of expressions in these algorithms are shown in Table 1.

Table 1. Notations used in scheduling algorithms.

Expression	Meaning
RB	Resource block
$w_{i,j}$	The priority of the i-th user on the j-th RB
$r_{i,j}$	Attainable data rate
$\overline{R}_i$	The approximated regular precedent data rate
$D_{HOL,i}$	Head of line packet delay
δ_i	Delay threshold
τ_i	Acceptable packet loss rate
∂_j^i	Spectral efficiency for the i-th user over the j-th RB
N	The number of real-time streams
$u_i(k)$	Data to be transmitted during the kth frame

2.1 M-LWDF [6]

The Modified Largest Weighted Delay First (M-LWDF) algorithm not only takes account of the channel quality and fairness, but also provides guarantee for real-time services from the perspective of delay and packet loss rate. The metric is defined as

$$w_{i,j}^{M\text{-}LWDF} = \alpha_i \cdot D_{HOL,i} \cdot \frac{r_{i,j}}{\overline{R}_i} \tag{1}$$

where α_i is further elaborated in (2). α_i weights the metric so that the streams with most urgent requirements are preferred for allocation.

$$\alpha_i = -\frac{\log \delta_i}{\tau_i} \tag{2}$$

In particular, attainable data rate $r_{i,j}$ for each User Equipment (UE) is evaluated by the link adaptation module according to feedbacks on channel quality. In general, the better channel condition is, the larger $r_{i,j}$ UE will have with higher priority. $\overline{R}_i$ is the approximated regular precedent data rate of i-th flow. As shown in Eq. (1), $\overline{R}_i$ is used to restrict users with large traffic and maintain fairness. For example, a high rate user may be constantly allocated with more resources in the past. But with the growth of $\overline{R}_i$, its priority will be reduced. The role of $D_{HOL,i}$ is to increase the priority of packets with longer waiting time in line.

As a whole, M-LWDF guarantees a good balance among spectrum efficiency, fairness and QoS provisioning.

2.2 EXP/PF [7]

It is also important to know that the metric of Proportional Fairness (PF) 4 is computed as the attainable data rate over the average data rate. The Exponential Proportional Fairness (EXP/PF) algorithm adds an exponential term based on the PF algorithm to increase the priority of the real-time stream. As its name states, EXP/PF considers both the characteristics of PF and of an exponential function of the end-to-end delay. The metric is described as

$$w_{i,j}^{EXP/PF} = exp\left(\frac{\alpha_i \cdot D_{HOL,i} - X}{1 + \sqrt{X}}\right) \cdot \frac{r_{i,j}}{\overline{R}_i} \tag{3}$$

where α_i is similarly defined as in (2), and

$$X = \frac{1}{N} \cdot \sum_{i=1}^{N} \alpha_i \cdot D_{HOL,i} \tag{4}$$

2.3 EXP-Rule and LOG-Rule [8]

EXP-rule and LOG-rule are very similar. They can quickly increase the priority of real-time flows approaching the delay threshold. In addition, priority formula considers the overall situation of the network by using the sum of all user delays. As their names

indicate, the main difference between the two algorithms is that one uses the log function and the other uses the exponential function as the main part of the priority formula.

$$w_{i,j}^{EXP\text{-}rule} = b_i \cdot exp\left(\frac{a_i \cdot D_{HOL,i}}{c + \sqrt{1/N \cdot \sum D_{HOL,i}}}\right) \cdot \varepsilon_j^i \tag{5}$$

$$w_{i,j}^{LOG\text{-}rule} = b_i \cdot log(c + a_i \cdot D_{HOL,i}) \cdot \varepsilon_j^i \tag{6}$$

where a_i, b_i and c are defined in (7) and (8) respectively.

$$\begin{cases} b_i = 1/E[\varepsilon_j^i] \\ c = 1.1 \cdot a_i = 5/0.99\tau_i \end{cases} \tag{7}$$

$$\begin{cases} a_i \epsilon \mathrm{E}[5/0.99\tau_i, 10/0.99\tau_i] \\ b_i = 1/\mathrm{E}[\varepsilon_j^i] \\ c = 1 \end{cases} \tag{8}$$

2.4 Delay-Prioritized Scheduler (DPS) [9]

DPS utilizes each user's packet delay information and its instantaneous downlink channel conditions for scheduling. Firstly, DPS orders candidate flows to the remaining time before the deadline expires. Once the user with the highest urgency is selected, the resource allocation step is performed in order to transmit the head of line packet (i.e. the most delayed one). A new iteration is then run on the remaining flows in the list until all RBs are assigned.

2.5 Frame Level Scheduler (FLS) [5]

The FLS algorithm consists of two layers. The upper level of the scheduler first estimates the amount of data that guarantees the QoS requirements of each real-time stream on the LTE radio frame (10 sub-frames). Then the lower level of the scheduler assigns resources to each real-time stream mainly according to the channel quality at each Transmission Time Interval (TTI) (1 ms). The better the channel quality, the higher the priority of resource allocation, and each stream stops resource allocation when the transmission completes the amount of data within the LTE frame. Once there is a surplus, lower level of scheduler allocates resources for the remaining non-real-time streams.

2.6 Other Real-Time Scheduling Algorithms

In [10], a delay-based weighted proportional fairness algorithm (DBWPF) is proposed, which considers weighted average delay of each user in addition to the trade-off between throughput and throughput fairness. The algorithm can improve delay fairness and implementation rate fairness. In [11], a delay–based and QoS–aware scheduling algorithm (DQAS) weights the delay priority of each queue by analyzing the queue buffer of each user stream. This weight is the decision basis for flow scheduling. This algorithm can effectively achieve the balance between experiment and system throughput during heavy loads.

3 Problem Presentation and DBS Algorithm

3.1 Problem Presentation

Although algorithms mentioned in Sect. 2 have good performance for real-time streaming, there is still the possibility of further improvement. The reasons are as follows:

(1) Inadequate use of delay

For real-time streams, the delays and the PLR are closely related. In general, if the packet in the buffer of eNodeB exceeds the delay threshold, then it will be dropped. Similarly, if an overdue packet is already sent, then it will also be dropped by its receiver (i.e. UE) and resource will be wasted. Thus, PLR will increase. Consequently, the real-time flow near the delay threshold has a larger probability of packet dropping, and PLR will increase when the delay increases [12].

In M-LWDF, EXP/PF and other algorithms, metrics become larger with the increasing of delay. But for some streams approaching to the delay threshold, owing to different channel conditions and historical transmission rate, the metrics may be lower. In this way, the priority of these streams cannot be guaranteed, and PLR due to timeout will increase.

Although in FLS algorithm, the upper level scheduling algorithm can calculate the amount of data that the real-time stream needs to transmit with larger scale (10 TTIs), the lower level assigns RBs to each user directly by the channel condition in each TTI. The lack of consideration for delay in the TTI scale could reduce throughput and increase PLR of real-time streams. The TTI-level delay characteristic reflects the relationship between delay of each flow and the time delay threshold, and it is indirectly related to packet loss. For example, there are two flows (flow A and flow B) in the lower level of scheduler. The delay of flow A is larger than flow B and close to the delay threshold. If the channel condition of flow A is slightly poor, flow A will probably lose packets due to timeout, which will affect the overall QoS satisfaction. Promisingly, the attribute of delay can be used at a more detailed scale to improve Packet Loss Rate (PLR) and throughput.

(2) Lack of competition

DPS algorithm assigns RBs depending on each flow's packet delay information and downlink channel conditions. If there are two streams with similar delays, DPS will allocate RBs to the flow with higher delay. But for real-time flow with slightly less delay and much better channel quality, this allocation method is not fair and has lower system efficiency. Therefore, it is more reasonable that the flows with near delay get the RBs through intra grade competition depending on channel quality.

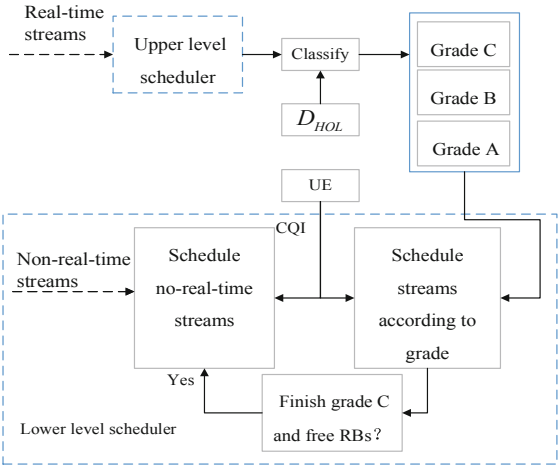


Fig. 1. DBS algorithm

3.2 Upper Level Scheduler

As shown in Fig. 1, the upper level scheduler is only valid for real-time stream. FLS algorithm [5] is exploited as the upper level scheduler. FLS algorithm estimates the amount of data that guarantees the QoS requirements of each real-time stream on the LTE frame. The authors of FLS suppose that packets of each real-time stream waiting for transmission are stored in a queue. Specifically, let $q_i(k)$ indicate the length of i th queue; $d_i(k)$ is the amount of newly filled data to i th queue; $u_i(k)$ is the amount of date corresponding to the flow's QoS requirements during the k th frame. The connection among the above values can be established by the following equation:

$$q_i(k+1) - q_i(k) = d_i(k) - u_i(k) \tag{9}$$

where $0 \leq q_i(k) \leq q_{max}$, $u_i(k) \geq 0$, $d_i(k) \geq 0$ and q_{max} presents the maximum length of queue q_i. To compute $u_i(k)$, FLS designs the following control method:

$$u_i(k) = h_i(k) * q_i(k) \tag{10}$$

where the "$*$" indicates the discrete-time convolution and $h_i(k)$ represents pulse response.

According to the derivation of (9) and (10) in [9], we can get $u_i(k)$ as follows:

$$u_i(k) = q_i(k) + \sum_{n=2}^{M_i} c_i(n)[q_i(k-n+1) - q_i(k-n+2) - q_i(k-n+1)] \tag{11}$$

where M_i indicates the width of the pulse-response, $c_i(n)$ represent a coefficient constrained by (12).

$$\begin{cases} 0 \leq c_i(n) \leq 1, \forall n \\ c_i(n) \geq c_i(n+1), n \geq 1 \text{ with } c_i(n) \epsilon \mathbb{R} \end{cases} \tag{12}$$

On the other hand, delay threshold is

$$\delta_i = (M_i+1)T_f \tag{13}$$

where T_f is the length of sampling interval. Now, we show an example in Fig. 2 to illustrate FLS. Firstly, we set $M_i = 9$ and $c_i(0) = 0$, $c_i(1) = 1$, $c_i(2) = 1/2$, $c_i(3) = 1/4, c_i(4) = 1/8, c_i(5) = 1/16$ and soon on. Then, we set inputs $d_i(k) = 2000$ bits, $d_i(k+1) = 1000$ bits, $d_i(k+2) = 0$ bits. It means that 1000 bits of data submitted to the queue in the frame k and so on. In the light of (11), the enqueued data over kth, (k + 1)th, …, and (k + 9)th frames are 1000, 500, …, 3. Thus, we calculate all data need transported over each frame and obtain $u_i(k)$. Thus, we calculate all data that each frame need to transport and count them up as $u_i(k)$.

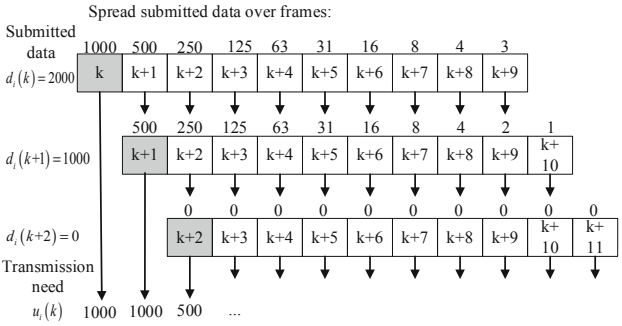


Fig. 2. Example of FLS

3.3 Lower Level Scheduler

We propose DBS algorithm (as shown in Fig. 1) that combines the upper level scheduler of FLS algorithm and a new lower level scheduler. Specifically, when a stream needs scheduling, we should judge whether it is a real-time stream. If yes, the upper estimates the amount of data that guarantees the QoS requirements of each real-time stream on the LTE frame. This also means that once the flow is satisfied with the amount of data required, it will not be allocated resources within the entire LTE frame. Figure 3 shows the flow chart of the DBS algorithm.

Then, all real-time streams are classified into grade A, B and C according to D_{HOL}. The real-time flow approaching the delay threshold has larger probability of packet dropping. We first determine that the interval near the delay threshold is [90%* δ_i, δ_i) and classify all real-time flows in this interval as grade A. These streams will have the highest priority for resources allocation and slightly lower PLR than before.

On the other hand, we should strive to reduce the possibility of streams being classified as grade A due to still larger PRL. The residual streams with small delay have little PLR or being dividing into grade A. Therefore, for real-time flows with medium delay performance and good delay performance, we divide them into grade B and C and with interval [50%* δ_i, 90%* δ_i) and (0, 50%* δ_i) respectively as shown in (14). Thus, each

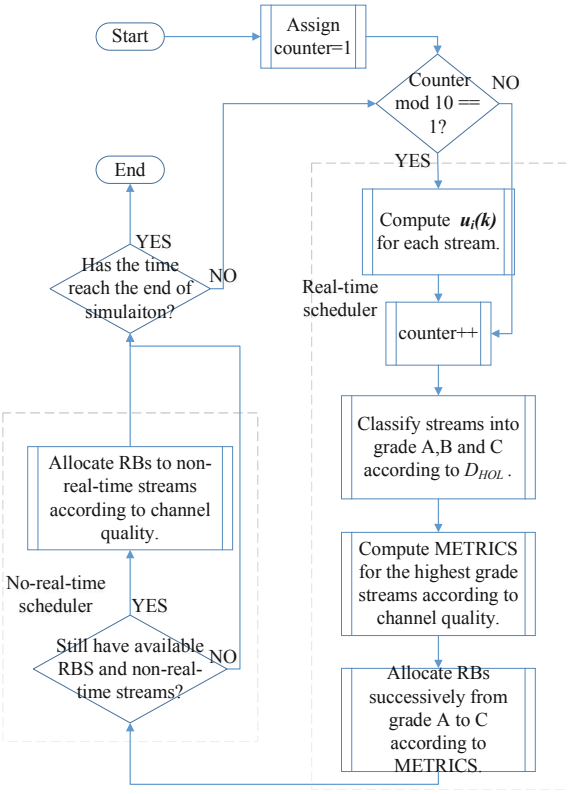


Fig. 3. Flowchart of DBS algorithm

real-time stream will be mapped to a unique grade.

$$grade\left(D_{HOL,i}\right) = \begin{cases} A & 90\% \cdot \delta_i \le D_{HOL,i} < \delta_i \\ B & 50\% \cdot \delta_i \le D_{HOL,i} < 90\% \cdot \delta_i \\ C & 0 < D_{HOL,i} < 50\% \cdot \delta_i \end{cases} \tag{14}$$

This classification method can make full use of delay in the TTI scale on the basis of FLS and guarantee the priority of the flows close to the threshold. Moreover, intra grade competition depending on channel quality can further improve throughput and reduce packet loss rate. The above classification strategy is determined by a large number of simulations, but it could be further improved.

After categorizing, we can allocate resources strictly according to the grades by Maximum Throughput (MT) algorithm [13] as (15).

$$w_{i,j}^{MT} = r_{i,j} \tag{15}$$

MT allocates the RBs only depending on the attainable data rate $r_{i,j}$. The larger $r_{i,j}$, the higher priority $r_{i,j}$ UE has. Attainable data rate $r_{i,j}$ for each UE is computed by the

link adaptation module depending on Channel Quality Indicator (CQI) from UE. The adoption of MT algorithm at lower level algorithm increases throughput and reduces PLR.

Briefly, we firstly assign RBs according to channel quality to real-time streams with grade A. If there is a surplus, RBs will be assigned to the flows with lower grade. In the end, if there is still a surplus, resources will be allocated to non-real-time flows.

4 Performance Evaluation

To evaluate the performance of the proposed algorithm, we conduct simulations with LTE-Sim [14], an open-source system level simulation platform for whole LTE system. Firstly, we introduce simulation scenario; then we compare the simulation results of our DBS to those of M-LWDF, EXP/PF and EXP-LOG.

4.1 Simulation Scenario

We use a single cell scenario with interference as our environment. Users move in a random direction at the speed of 3 km/h. More specific parameters are shown in Table 2.

Table 2. Simulation parameters

Parameter	Value
Simulation duration	40 s
Physical detail	Carrier Frequency: 2 GHz Downlink Bandwidth: 5 MHz Number of RBs: 25
Link adaptation	Modulation Scheme: QPSK, 16QAM, 64QAM Target BLER: 10^{-1}
Cell layout	Radius: 1 km Number of UE: [10, 30]
UE mobility	Mobility model: Random Walk; UE speed: 3 km/h
Traffic model	Real-time flow type: H.264 440 kbps; best effort flow: infinite buffer

4.2 Performance Metrics

We mainly evaluate performance of the DBS algorithm based on PLR and system throughput.

The PLR and system throughput are given as follow:

$$\text{PLR} = \frac{\sum_{i=1}^{K} \sum_{t=1}^{T} pdiscard_i(t)}{\sum_{i=1}^{K} \sum_{t=1}^{T} psize_i(t)} \tag{16}$$

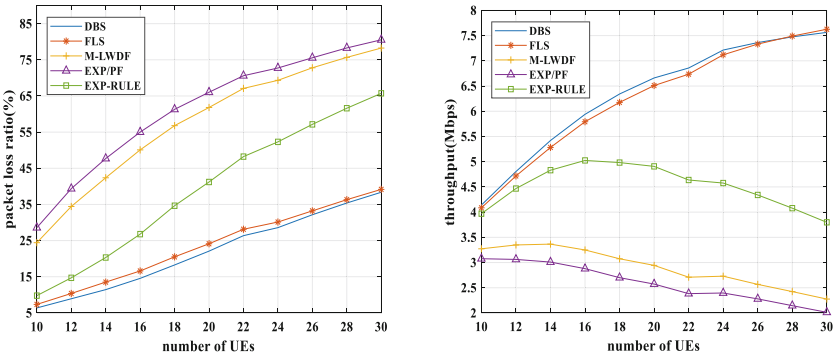
$$\text{system\_throughput} = \frac{1}{T}\sum_{i=1}^{K}\sum_{t=1}^{T}\text{ptransmit}_i(t) \tag{17}$$

where $pdiscard_i(t)$, $psize_i(t)$ and $ptransmit_i(t)$ represent the size of all dropped packets, the size of all packets arriving into the working eNodeB buffer and total size of transmitted packets of user i at time t 9.

4.3 Results and Discussion

We evaluate the performance of DBS, FLS, M-LWDF, EXP/PF and EXP-LOG by changing the number of UEs and the delay threshold of real-time flows.

As shown in Fig. 4(a) and (b), we set the number of users in the system from 10 to 30. We can observe that average PLR and average throughput increase with the number of UEs owing to higher network load. Throughputs of algorithms other than DBS and FLS are different. Because they do not distinguish between real-time flows and non-real-time flows as DBS. Figure 4(a) shows that DBS has lower PLR than FLS. Especially when the number of users is between 12 and 22, the advantage of PLR is more evident. On the whole, the average value of PLR of DBS is 1.54% lower than that of algorithm FLS. The effect is obvious.



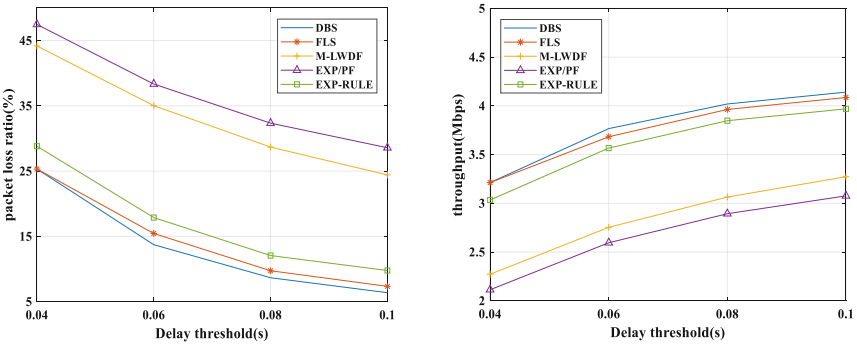
(a) (b)

Fig. 4. (a). Video packet loss ratio (b). Video throughput

Figure 4(b) shows that the throughput of DBS is larger than FLS except for 28 UEs. Because when the number of users is too high, the load is extremely large. A large part of packets cannot be sent in time. That is to say, most packets have a large delay. Therefore, in this scenario the DBS is not as good as that of FLS. In addition, EXP-RULE as an upgrade of EXP/PF has a higher degree of emphasis on delay. Therefore, the performance of EXP-RULE is better than M-LWDF and EXP/PF. The two figures clearly show that DBS can improve throughput and reduce packet loss rate.

In Fig. 5(a) and (b), the curves of PLR and throughput under different delay thresholds are shown when the number of UE is 10. It is noteworthy that due to the limitation of FLS

algorithm, simulation involves only four delay thresholds. Plainly, a larger value of delay threshold implies a lower PLR due to a less number of packets violating the deadline. The trend of throughput is similar to PLR curve. Figure 5(a) clearly indicates that PLR of DBS is slightly lower than that of FLS and far smaller than that of EXP-RULE and other algorithms. But when the threshold is the minimum, the PLR of DBS is nearly the same as that of FLS. On the whole, the advantages of DBS under different delay thresholds are much conspicuous. Figure 5(b) compares the performance of throughput. It shows that DBS can enhance video throughput on the basis of FLS. This is due to the improvement of PLR.



(a) (b)

Fig. 5. (a). Video packet loss ratio for 10 UEs (b). Video throughput for 10 UEs

5 Conclusion

In this paper, we propose a delay based hierarchical downlink scheduling algorithm for real-time stream in cellular networks. Upper level scheduler offers a prediction of the number of data that each real-time stream needs to ensure the QoS. After classifying all real-time streams into grade A, B and C according to their D_{HOL}, the lower level scheduler allocates resource to streams by their different grades. The results of simulation have obviously shown that DBS algorithm improves the performance of FLS algorithm in terms of delay and throughput.

Acknowledgment. This work has been sponsored by National Engineering Laboratory for Next Generation Internet Technologies open fund project, Research on network traffic prediction technology based on spatiotemporal deep learning model, National Science Foundation of China (No. 61201149) and Huawei Research Fund (grant No. YBN2016110032). The authors would also like to thank the reviewers for their constructive comments.

References

1. Wang, Y.C., Hsieh, S.Y.: Service-differentiated downlink flow scheduling to support QoS in long term evolution. Comput. Netw. **94**, 344–359 (2016)

2. Krasilov, E., Krasilov, A., Malyshev, A.: Radio resource and traffic management for ultra-reliable low latency communications. In: Proceedings of IEEE WCNC (2018)
3. Shafi, M., et al.: 5G: a tutorial overview of standards trials challenges deployment and practice. IEEE J. Sel. Areas Commun. **35**(6), 1201–1221 (2017)
4. Girici, T., Zhu, C., Agre, J.R., Ephremides, A.: Proportional fair scheduling algorithm in OFDMA-based wireless systems with QoS constraints. J. Commun. Netw. **12**(1), 30–42 (2010)
5. Piro, G., et al.: Two-level downlink scheduling for real-time multimedia services in LTE networks. IEEE Trans. Multimedia **13**(5), 1052–1065 (2011)
6. Ramli, H., Basukala, R., Sandrasegaran, K., Patachaianand, R.: Performance of well-known packet scheduling algorithms in the downlink 3GPP LTE system. In: Proceedings of MICC, pp. 815–820 (2009)
7. Basukala, R., Ramli, H.M., Sandrasegaran, K.: Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system. In: Proceedings of AH – ICI (2009)
8. Sadiq, B., Baek, S.J., de Veciana, G.: Delay-optimal opportunistic scheduling and approximations: the Log rule. IEEE Trans. Netw. **19**(2), 406–418 (2011)
9. Sandrasegaran, K., Ramli, H.A.M., Basukala, R.: Delay-prioritized scheduling (DPS) for real time-traffic in 3GPP LTE system. In: Proceedings of IEEE WCNC, pp. 18–21 (2010)
10. Liu, S., Zhang, C., Zhou, Y., Zhang, Y.: Delay-based weighted proportional fair algorithm for LTE downlink packet scheduling. Wireless Pers. Commun. **82**(3), 1955–1965 (2015)
11. Madi, N.K., Hanapi, Z.M., Othman, M., Subramaniam, S.K.: Delay-based and QoS-aware packet scheduling for RT and NRT multimedia services in LTE downlink systems. EURASIP J. Wirel. Commun. Netw. **2018**, 180 (2018)
12. Hendaoui, S., Zangar, N., Tabbane, S.: Downlink scheduling for real time application over LTE-A network: delay aware scheduling. In: Proceedings of COMNET, pp. 4–7 (2015)
13. Kela, P., et al.: Dynamic packet scheduling performance in UTRA long term evolution downlink. In: Proceedings of ISWPC (2008)
14. Piro, G., Grieco, L., Boggia, G., Capozzi, F., Camarda, P.: Simulating LTE cellular systems: an open-source framework. IEEE Trans. Veh. Technol. **60**(2), 498–513 (2011)