Service-differentiated Downlink Flow Scheduling to Support QoS in Long Term Evolution

You-Chiun Wang and Song-Yun Hsieh

Abstract—The growing demand of network services, especially for broadband downlink communication, has triggered the evolution of cellular systems. Recently, 3GPP continuously works out the standard of *long term evolution (LTE)* in response to the oncoming 4G cellular system. LTE adopts the OFDMA technology for downlink communication, and divides the spectral resource into physical resource blocks (PRBs). However, the *LTE flow scheduling problem*, which asks how to allocate PRBs to downlink flows for communication, is not well addressed in the standard but has great impact on transmission efficiency. On the other hand, LTE categorizes flows into *guaranteed bit rate (GBR)* and *non-GBR* classes, where GBR flows usually have higher priorities and shorter delay constraint than non-GBR flows. Although various solutions have been proposed to the LTE flow scheduling problem, many of them generally aim at maximizing transmission efficiency or keeping fair transmission, which may starve non-GBR flows or cannot well support quality of service (QoS) for GBR flows. Therefore, the paper develops a *service-differentiated downlink flow scheduling (S-DFS)* algorithm by taking the aforementioned difference between flows into consideration. Except for increasing transmission efficiency, S-DFS has two major goals: 1) satisfying the QoS requirement of GBR flows, and 2) ensuring the data transmission of non-GBR flows. Specifically, S-DFS first deals out PRBs to flows according to their channel conditions and *QoS class identifier (QCI)* defined in LTE. Then, with the mechanism of *resource reallocation*, S-DFS can assign a dynamic amount of reallocatable PRBs to the flows whose packets are about to be dropped. Experimental results demonstrate that S-DFS can achieve higher LTE transmission efficiency. Furthermore, it not only reduces both dropping ratio and delay of GBR packets, but also improves data throughput of non-GBR flows.

Index Terms—cellular system, downlink flow scheduling, long term evolution (LTE), physical resource block (PRB), quality of service (QoS).

1 Introduction

The telecommunications industry keeps growing and innovating over the past few decades [1]. Nowadays, people expect to freely access Internet anytime, anywhere through mobile devices, and they are consequently hungry for high-speed wireless communication. In response to the above need, Third Generation Partnership Project (3GPP) develops *long term evolution (LTE)* for the current (and possibly future) generation of cellular system, and today, LTE systems have been operated in many countries [2]. On the other hand, it becomes common to use network services with the demand of broadband downlink communication, for example, online gaming, multimedia streaming, and mobile TV [3]. According to the Cisco report in [4], Internet video streaming and downloads have taken a large share of global network bandwidth in 2014, and will grow to more than 80% of all consumer Internet traffics by 2019. These broadband downlink services obviously pose challenges in the design of LTE systems.

LTE employs the technology of *orthogonal frequency division multiple access (OFDMA)* for its downlink communication, and divides the spectral resource into two-dimensional array of *physical resource blocks (PRBs)*. A PRB has the duration of 0.5 ms in the time domain and the length of 180 kHz in the frequency domain. PRBs are 'non-sharable' resources, so each PRB can be given to at most one user. The number of available PRBs depends on the downlink transmission bandwidth. LTE allows the bandwidth to be 1.4, 3, 5, 10, 15, or 20 MHz, which supports 6, 15, 25, 50, 75, or 100 PRBs, respectively. How to allocate PRBs to downlink flows for transmission will significantly affect LTE transmission efficiency [5], and we call it the *LTE flow scheduling problem*. However, 3GPP leaves this problem to the research and industrial communities.

Furthermore, to provide differential QoS for various services, LTE classifies flows into two categories: *guaranteed bit rate (GBR)* and *non-GBR*. GBR flows can support real-time services with strict delay constraints, such as voice over IP (VoIP), live-streaming video, and online games. Non-GBR flows are often used to provide non-real-time services with loose deadlines, for example, TCP-based applications. However, some existing solutions to the LTE flow scheduling problem try to improve the overall transmission efficiency on the cost of flows with bad channel conditions or small priorities, which could thus starve non-GBR flows. On the other hand, other solutions attempt to achieve fair transmissions among flows. However, when the PRBs are not sufficient to support all flows, they may not guarantee QoS for GBR flows.

Therefore, the aforementioned observations motivate us to develop the *service-differentiated downlink flow scheduling (S-DFS)* algorithm to efficiently solve the LTE flow scheduling problem. In addition to improving LTE transmission efficiency, our S-DFS algorithm has two primary objectives: 1) satisfying the QoS requirement of GBR flows and 2) ensuring the data transmission of non-GBR flows. Here, the first objective means that we have to meet the constraints of packet dropping and delay of GBR flows, while the second objective indicates that we should prevent non-GBR flows from starvation. To do so, our S-DFS algorithm considers not only the channel quality, queue status, and head-of-line (HOL) packet delay of each flow, but also its *QoS class identifier (QCI)*, which is a scalar identifier defined by LTE to describe the QoS characteristics of GBR and non-GBR flows. In particular, the S-DFS algorithm allows flows to bid for PRBs by using their channel quality and QCI priority, so it can increase transmission efficiency

The authors are with the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, 80424, Taiwan.
E-mail: ycwang@cse.nsysu.edu.tw; m023040061@student.nsysu.edu.tw

and favor GBR flows. Then, the S-DFS algorithm searches for 'reallocatable' PRBs from the above resource allocation by using two adjustable parameters α and β, which avoids some flows occupying too many PRBs. Finally, these reallocatable PRBs are assigned to the flows with impending packet discard. This resource reallocation mechanism not only reduces potential packet dropping but also allows non-GBR flows to obtain necessary PRBs for transmission.

The contributions of this paper are summarized as follows:

- We develop an efficient S-DFS solution to the LTE flow scheduling problem by differentiating flows based on their QoS characteristics.
- S-DFS adopts a novel mechanism of resource reallocation to 'redistribute' a subset of resource to the flows in urgent need of PRBs.
- By adjusting parameters α and β, S-DFS can flexibly fine tune the amount of resource given to different types of services.
- Experimental results show that S-DFS not only reduces GBR packet dropping and delay, but also improves non-GBR data throughput, as compared with a number of popular LTE flow scheduling methods.

The remainder of this paper is organized as follows: The next section discusses related work. Section 3 presents the network model for the LTE flow scheduling problem, and Section 4 proposes our S-DFS algorithm to solve the problem. We then compare the performance of different LTE flow scheduling schemes in Section 5. Finally, Section 6 concludes this paper.

2 RELATED WORK

Flow scheduling has received a lot of research attention, and various methods are proposed for wireless networks. Both *maximum throughput (MT)* [6] and *proportional fair (PF)* [7] are two typical scheduling methods. MT seeks to maximize transmission efficiency by selecting the UE u_i with the best channel quality:

$$u_i = \arg\max_i (r_i(t)), \tag{1}$$

where $r_i(t)$ is u_i's channel rate at the current time t. PF aims at supporting fair transmission among UEs. Thus, PF compares the current channel rate of each UE with its (past) average data rate, and then selects the one with the largest value:

$$u_i = \arg\max_i \left(\frac{r_i(t)}{r_i^M} \right), \tag{2}$$

where r_i^M is the mean data rate of UE u_i. However, MT and PF do not consider the delay constraint of real-time applications. To address this issue, *modified largest weighted delay first (M-LWDF)* [8] adds a weight factor w_i and the HOL packet delay $d_i(t)$ in the above Eq. (2) as follows:

$$u_i = \arg\max_i \left(w_i d_i(t) \cdot \frac{r_i(t)}{r_i^M} \right), \tag{3}$$

where $w_i = -\log \delta_i / \tau_i$. Here, δ_i denotes the maximum probability that the HOL packet delay exceeds the delay threshold of UE u_i, and τ_i defines the target delay for u_i. To favor real-time traffics over non-real-time ones, *exponential proportional fair (EXP/PF)* [9] further improves PF by taking the mean HOL packet delay $d_i^M(t)$ into consideration:

$$u_i = \arg\max_i \left(\exp\left(\frac{w_i d_i(t) - d_i^M(t)}{1 + \sqrt{d_i^M(t)}} \right) \cdot \frac{r_i(t)}{r_i^M} \right), \tag{4}$$

where $d_i^M(t) = \sum_j w_j d_j(t)/n_R$ and n_R is the number of real-time flows. The work of [10] develops both LOG-RULE and EXP-RULE which consider the channel condition of each UE. Specifically, LOG-RULE selects the UE according to the following equation:

$$u_i = \arg\max_i (b_i \cdot \log(c + a_i d_i(t)) \cdot \Gamma_i), \tag{5}$$

where b_i, c, and a_i are tunable parameters, and Γ_i denotes the spectral efficiency for UE u_i (on a subchannel). EXP-RULE can be viewed as an enhancement of EXP/PF, and its metric is similar to that of LOG-RULE:

$$u_i = \arg\max_i \left(b_i \cdot \exp\left(\frac{a_i d_i(t)}{c + \sqrt{(1/n_R)\sum_j d_j(t)}} \right) \cdot \Gamma_i \right). \tag{6}$$

A number of research efforts also develop their flow scheduling schemes to support real-time services or provide fair transmissions among flows in LTE networks. Luo et al. [11] propose a cross-layer framework to provide smooth video delivery over LTE, which considers the delay constraint, distortion, and historical data rate of a video flow, so as to determine its resource allocation and coding scheme. The work of [12] develops a two-level LTE downlink scheduler for multimedia services. The upper level employs the discrete-time linear control theory [13] to estimate the amount of data that each multimedia flow should transmit within a scheduling period in order to satisfy its delay requirement. Then, the lower level assigns PRBs to flows based on the PF scheme. Iturralde et al. [14] also consider a two-level scheduling strategy to allocate resource in LTE networks. The first level views the flow scheduling problem as a bankruptcy game and employs the Shapley value [15] to provide fair resource distribution among flows. Then, the second level follows EXP-RULE (i.e, Eq. (6)) to assign PRBs. The study in [16] adopts a utility function to calculate the user's degree of satisfaction for each flow, and then follows the similar idea of cooperative game in [17] to let flows compete for PRBs according to their utility values. In [18], Liu et al. combine the PF scheme with the *earliest-deadline-first (EDF)* strategy [19], where EDF always schedules the packet with the closest deadline expiration first. Thus, they seek to take into consideration both the fairness feature of PF and the bounded-delay feature of EDF. The work of [20] classifies flows into *urgent* and *non-urgent* ones, where urgent flows should be granted with the highest priority to quickly send out their packets. Then, non-urgent flows, which contain non-real-time flows and real-time flows whose packet deadlines are yet unexpired, can have the equal opportunity to acquire the spectral resource for transmission. Given the initial scheduling of PRBs by Eq. (1), Lai and Tang [21] develop a *packet prediction mechanism (PPM)* to support QoS for real-time services. PPM uses the virtual queue to estimate the behavior of future incoming packets based on the packets in the current queue. Then, it rearranges the transmission order and discards those packets that cannot meet their delay demand.

Several studies take into consideration the difference between GBR and non-GBR flows. For example, the work of

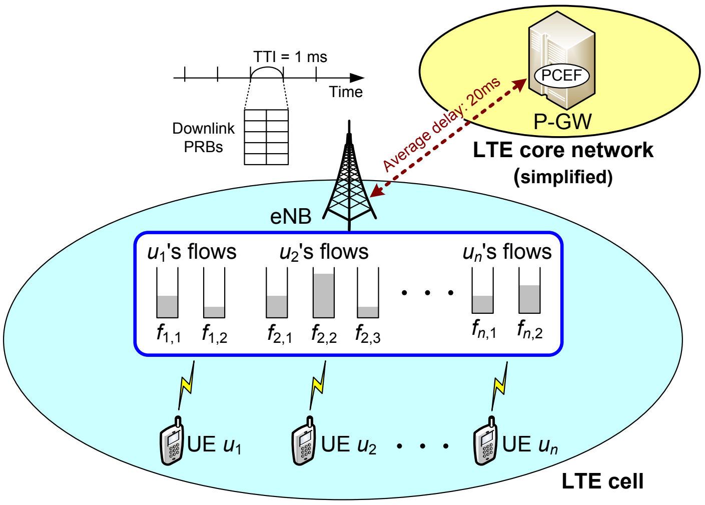


Fig. 1: The LTE network architecture.

[22] develops an LTE MAC scheduler to categorize incoming packets into five QoS classes, which correspond to the services with QCI indices of 1, 2, 7, 8, and 9 (in other words, it considers two types of GBR flows and three types of non-GBR flows). The MAC scheduler performs strict priority scheduling for all GBR flows and then scheduling non-GBR flows. Nevertheless, when the eNB has no sufficient resource to serve all flows, non-GBR flows would be starved, because the MAC scheduler always gives the resource to transmit GBR packets first. On the other hand, Mushtaq et al. [23] propose an opportunistic scheduling approach to calculate the priority of each UE u_i as follows:

$$P_i = \begin{cases} D_i\varepsilon_i \cdot (PF_i \cdot (\frac{r_{GBR}}{r_i^M})^{\phi} + B_i), & u_i \text{ is a GBR UE} \\ \varepsilon_i \cdot (PF_i + B_i), & \text{otherwise} \end{cases} \quad (7)$$

where D_i is the delay factor, ε_i is a discontinuous reception (DRX) indicator, r_i^M is the average throughput, and B_i is the buffer status of u_i. PF_i borrows the idea from the PF scheme (referring to Eq. (2)), r_{GBR} denotes the average throughput of a GBR UE, and ϕ is a tunable exponential factor. Then, PRBs are allocated to UEs according to their priorities. Obviously, Eq. (7) does not take QCI into account.

Comparing with the aforementioned work, this paper seeks to exploit the QCI characteristics in Table 1 along with multiple parameters of flows to allocate PRBs, so that the QoS requirement of GBR flows can be satisfied while non-GBR flows will not be starved. Instead of simply calculating a weight or priority for each UE to assign PRBs (e.g., Eqs. (1)–(7)), our S-DFS algorithm considers a more sophisticated scheduling mechanism. Specifically, it first uses QCI priority, channel quality, and queue status to distribute PRBs among flows. Then, a subset of resource is adaptively reallocated to those flows in urgent need of PRBs to alleviate potential packet dropping. The proposed S-DFS algorithm can be tailored to LTE networks by considering the QCI property, and extensive simulation results in Section 5 will also demonstrate its effectiveness.

3 Network Model

Fig. 1 gives the LTE network architecture, which consists of two major parts: *LTE core network* and *LTE cells*. The LTE core network is responsible for various jobs such as back-end management, billing, and connecting to external networks. On the other hand, *user equipments (UEs)* are served in LTE cells. We thus focus on the flow scheduling problem in a single LTE cell, where the central base station (called *E-UTRAN Node B*, or *eNB*) takes responsibility of distributing downlink PRBs among all UEs. To assist the eNB in selecting the *modulation and coding scheme (MCS)* for data transmission, each UE will report the *channel quality indication (CQI)* value, which reveals its current channel condition, to the eNB for reference. The duration of each flow scheduling process is called a *transmission time interval (TTI)*, whose length is 1 ms. A UE can possess multiple flows, and each flow has a *queue* at the eNB as its packet container (shown in Fig. 1). Packets are stamped with their arrival time in the queue when they are generated, and the eNB transmits the packets in a queue according to the first-in-first-out (FIFO) principle. The HOL packet delay, which is the elapsing time of the first packet in the queue after its arrival, will be recomputed by the eNB in every TTI.

LTE introduces the QCI concept to describe the QoS characteristic of each flow, which includes its *packet delay budget* and *packet error loss rate*, as shown in Table 1. In particular, the packet delay budget is the maximum time that a packet may be delayed between a UE and the *policy and charging enforcement function (PCEF)*, which is a part of the *packet data network gateway (P-GW)* in the LTE core network, as shown in Fig. 1. Since the average delay between the PCEF and an eNB is 20 ms (suggested by the LTE Release 13 standard [24]), we can derive the expected *delay tolerant time* that a flow can wait for a packet delivered from the eNB by subtracting 20 ms from the corresponding packet delay budget. Therefore, we can determine whether to drop a packet by referring to the above delay tolerant time. On the other hand, the packet error loss rate defines the upper bound for the probability that a packet arrives at the eNB but is not received by the UE. Such phenomena may occur when packets are dropped because of serious noise interference or expiration (i.e., they become overdue). Table 1 presents the standardized QCI characteristics defined in LTE, which contains six types of GBR flows and seven types of non-GBR flows[1]. The *priority* term in the QCI table indicates the importance of a flow. Generally speaking, GBR flows have higher QCI priorities than non-GBR flows (except for the IMS signaling with QCI index = 5 and the MCPTT signaling with QCI index = 69), and also smaller packet delay budgets.

Given the CQI value of each UE and the traffic demands of its flows, the LTE flow scheduling problem determines how to allocate downlink PRBs to UEs in every TTI such that their demands are satisfied, under the constraint that each PRB can be assigned to at most one UE. Notice that there exists a feasible solution to this problem only if the eNB has an enough number of PRBs to satisfy the traffic demands of all flows. In case of insufficient PRBs, our objective is to satisfy the QoS requirement of GBR flows, while guarantee non-GBR flows to receive a number of PRBs for transmission (in other words, prevent them from starvation).

4 The Proposed S-DFS Algorithm

In every TTI, the eNB will collect the information of CQI and queue status from each UE, and then execute the S-DFS

[1] There were originally nine types of QCIs defined in LTE Release 8 standard [25]. However, to support more applications such as MCPTT and mission critical data, LTE Release 13 standard adds four new types with QCI indices of 65, 66, 69, and 70 in Table 1.

TABLE 1: Standardized QCI characteristics defined in LTE (Release 13).

QCI index	flow type	QCI priority	packet delay budget	packet error loss rate	example services
1	GBR	2	100 ms	10^{-2}	conversational voice[1]
2	GBR	4	150 ms	10^{-3}	conversational video[2]
3	GBR	3	50 ms	10^{-3}	real-time gaming
4	GBR	5	300 ms	10^{-6}	non-conversational video[3]
65	GBR	0.7	75 ms	10^{-2}	MCPTT[4]
66	GBR	2	100 ms	10^{-2}	non-MCPTT
5	non-GBR	1	100 ms	10^{-6}	IMS[5] signaling
6	non-GBR	6	300 ms	10^{-6}	video[3], TCP-based[6]
7	non-GBR	7	100 ms	10^{-3}	voice, video[2], interactive gaming
8	non-GBR	8	300 ms	10^{-6}	video[3], TCP-based[6]
9	non-GBR	9	300 ms	10^{-6}	video[3], TCP-based[6]
69	non-GBR	0.5	60 ms	10^{-6}	MCPTT signaling
70	non-GBR	5.5	200 ms	10^{-6}	mission critical data

[1] VoIP [2] live streaming [3] buffered streaming
[4] mission critical user plane push to talk voice
[5] IP multimedia subsystem
[6] WWW, e-mail, chat, ftp, p2p file sharing, progressive video, etc.

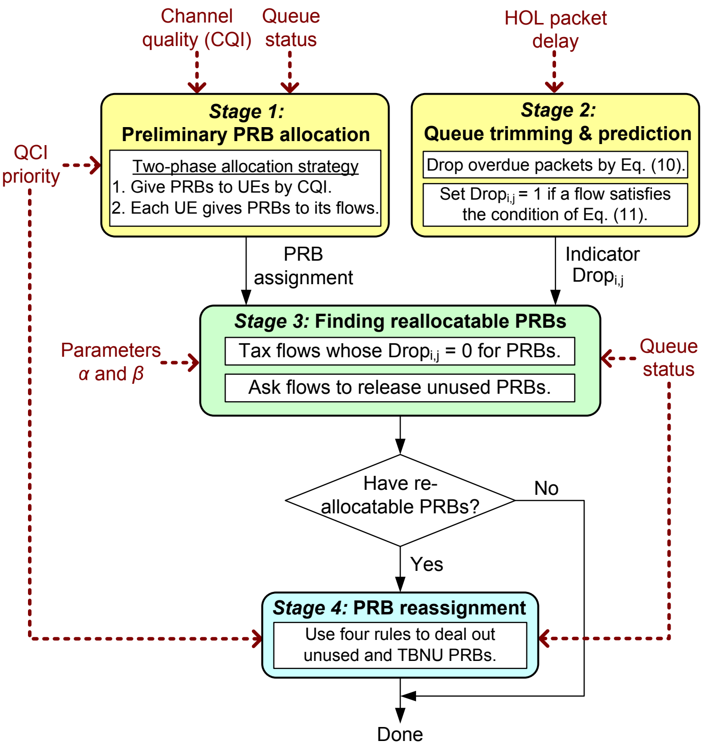

Channel quality (CQI)
Queue status
HOL packet delay
QCI priority
Stage 1: Preliminary PRB allocation
Two-phase allocation strategy
1. Give PRBs to UEs by CQI.
2. Each UE gives PRBs to its flows.
Stage 2: Queue trimming & prediction
Drop overdue packets by Eq. (10).
Set Drop$_{i,j}$ = 1 if a flow satisfies the condition of Eq. (11).
PRB assignment
Indicator Drop$_{i,j}$
Stage 3: Finding reallocatable PRBs
Parameters α and β
Tax flows whose Drop$_{i,j}$ = 0 for PRBs.
Ask flows to release unused PRBs.
Queue status
Have re-allocatable PRBs?
No
Yes
Stage 4: PRB reassignment
Use four rules to deal out unused and TBNU PRBs.
Done

Fig. 2: The flowchart of our proposed S-DFS algorithm, where the dotted lines indicate the parameters required in each stage.

algorithm to deal out the downlink PRBs to UEs' flows. Fig. 2 presents the flowchart and parameters of our S-DFS algorithm, which contains the following four stages:

(1) **Preliminary PRB allocation:** We adopt a two-phase strategy to calculate the preliminary assignment of downlink PRBs. In the first phase, the eNB gives PRBs to UEs based on their channel quality (i.e., CQI values). Then, in the second phase, each UE deals out the acquiring PRBs to its flows according to the QCI priority and queue status.

(2) **Queue trimming and prediction:** By referring to the HOL packet delay and the delay tolerant time of each flow, the eNB trims its corresponding queue by discarding overdue packets. In addition, the eNBs predicts whether this flow will still encounter packet dropping in the next TTI, and marks an indicator Drop$_{i,j}$ accordingly.

(3) **Finding reallocatable PRBs:** From *Stage 1*, some flows may be allocated with multiple PRBs. Therefore, the eNB will ask these flows to 'return' a part of their acquiring resource to the system (which is controlled by two parameters α and β). We call such resource *reallocatable PRBs*, and they can be adaptively assigned to other flows in the next stage to further improve the system performance.

(4) **PRB reassignment:** Based on the result of *Stage 2*, the flows that will suffer from packet dropping soon can have precedence over others to get the reallocatable PRBs. Therefore, these flows can compete for such PRBs by using the queue status and QCI priority. In this way, we can alleviate potential discard of packets in the subsequent TTI.

Below, we present the detailed design in each stage, and then give the rationale of our S-DFS algorithm.

4.1 Stage 1: Preliminary PRB Allocation

In the first stage, we seek to give a 'basic' allocation of PRBs to flows such that the overall transmission efficiency can be increased. To do so, we employ a *two-phase allocation strategy*, as shown in Fig. 2. The eNB first assigns PRBs on the basis of UEs, and then each UE deals out the acquiring PRBs to its flows. Specifically, for each PRB, the eNB allocates it to the UE that has the maximum CQI value to that PRB. In other words, we follow the similar concept in Eq. (1). This method helps each UE well utilize its PRBs by using better MCS to send data and, in consequence, we can improve LTE transmission efficiency.

Each UE then gives the acquiring PRBs to its flows with non-empty queues (i.e., these flows have data for transmission in the current TTI). Broadly speaking, flows with more importance (e.g., real-time or urgent applications) should be allocated with PRBs first. Therefore, each UE sorts its flows according to the QCI priority defined in Table 1 (from higher to lower), which indicates the importance of a flow. Let n_i be the number of PRBs acquired by UE u_i, and $F_i = \{f_{i,1}, f_{i,2}, \cdots, f_{i,k}\}$ be the set of u_i's flows, where $k \leq 13$ and $f_{i,j}$ has a higher QCI priority than $f_{i,j+1}$, $j = 1..k-1$. Then, we consider two possible cases as follows:

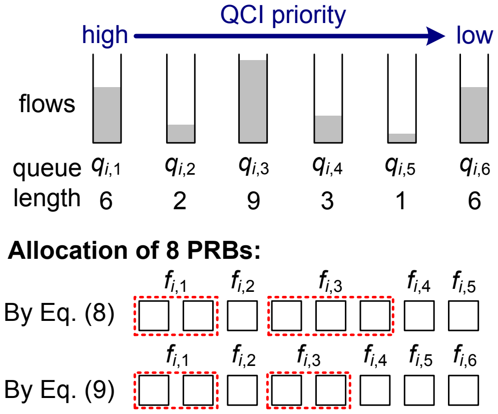


Fig. 3: An example of the PRB allocation in *Stage 1*.

- Case of $n_i \leq |F_i|$: This case implies that some flows of u_i cannot get any PRB, or every flow can be allocated with just one PRB. Thus, UE u_i will give one PRB to each of its flows $f_{i,1}$, $f_{i,2}$, $\cdots$, f_{i,n_i} in this case. Our argument is that high-priority flows each can receive at least one PRB to transmit data in the current TTI, so as to reduce its potential packet delay or dropping.
- Case of $n_i > |F_i|$: In this case, some flows can be given with more than one PRB. Thus, we consider not only the QCI priority but also the queue status of each flow to assign PRBs. In particular, let us denote by $q_{i,j}$ the queue length of flow $f_{i,j}$. Then, each flow of u_i is expected to acquire a number of

$$n_{i,j} = \frac{q_{i,j} \cdot n_i}{\sum_{l=1}^{k} q_{i,l}} \text{ PRBs.} \tag{8}$$

Here, we do not simply add a ceiling function to the above Eq. (8), because it may cause some flows unable to get any PRB (this situation will occur when $\sum_{\forall j} \lceil (q_{i,j} \cdot n_i) / \sum_{l=1}^{k} q_{i,l} \rceil > n_i$). Instead, we set $n_{i,j} = 1$ if $n_{i,j} < 1$, and otherwise round off $n_{i,j}$ to an integer. Then, for each flow $f_{i,j}$, we check whether

$$n_{i,j} + (k - j) \leq n_i. \tag{9}$$

If Eq. (9) becomes true, $f_{i,j}$ can acquire a number of $n_{i,j}$ PRBs, and we decrease n_i by $n_{i,j}$. Otherwise, we give flow $f_{i,j}$ a number of $n_i - (k - j)$ PRBs, and then assign one PRB to each of the remaining flows (i.e., flows $f_{i,j+1} \sim f_{i,k}$).

Notice that all flows of every UE u_i is sorted by their priorities in advance, and we always assign PRBs following the sequence of $f_{i,1}$, $f_{i,2}$, $\cdots$, and $f_{i,k}$ in the above two cases. That means our *Stage 1* does consult the QCI priority to distribute PRBs among flows.

We give an example in Fig. 3 to illustrate the case of $n_i > |F_i|$. Suppose that a UE u_i acquires eight PRBs and it has six flows (that is, $n_i = 8$ and $k = 6$). The queue lengths of these six flows are $q_{i,1} = 6$, $q_{i,2} = 2$, $q_{i,3} = 9$, $q_{i,4} = 3$, $q_{i,5} = 1$, and $q_{i,6} = 6$. Thus, the total data demand of u_i will be $\sum_{l=1}^{6} q_{i,l} = 27$. Then, according to Eq. (8), we can calculate the expected number of PRBs given to each flow of u_i as follows:

$$n_{i,1} = 6 \times 8/27 \approx 1.78 \rightarrow 2,$$
$$n_{i,2} = 2 \times 8/27 \approx 0.59 \rightarrow 1,$$
$$n_{i,3} = 9 \times 8/27 \approx 2.67 \rightarrow 3,$$
$$n_{i,4} = 3 \times 8/27 \approx 0.89 \rightarrow 1,$$
$$n_{i,5} = 1 \times 8/27 \approx 0.30 \rightarrow 1,$$
$$n_{i,6} = 6 \times 8/27 \approx 1.78 \rightarrow 2.$$

However, if we simply use the above allocation, flow $f_{i,6}$ cannot get any PRB, as shown in Fig. 3. Therefore, we adopt Eq. (9) to adjust the PRB allocation as follows:

- Flow $f_{i,1}$: It asks for $n_{i,1} = 2$ PRBs and $n_{i,1} + (k-1) = 2 + (6-1) < n_i = 8$. Thus, we can allocate two PRBs to flow $f_{i,1}$. Then, we update n_i by $8 - 2 = 6$.
- Flow $f_{i,2}$: It asks for $n_{i,2} = 1$ PRB and $n_{i,2} + (k-2) = 1 + (6-2) < n_i = 6$. Thus, we can allocate one PRB to flow $f_{i,2}$. Then, we update n_i by $6 - 1 = 5$.
- Flow $f_{i,3}$: It asks for $n_{i,3} = 3$ PRBs but $n_{i,3} + (k-3) = 3 + (6-3) > n_i = 5$ (i.e., Eq. (9) becomes false). In this situation, we can only allocate $n_i - (k-3) = 5 - (6-3) = 2$ PRBs to flow $f_{i,3}$.
- Other flows: Since we remain only three PRBs, each of flows $f_{i,4}$, $f_{i,5}$, and $f_{i,6}$ can be allocated with just one PRB.

The final allocation of PRBs is shown in Fig. 3. In this way, we can make sure that high-priority flows are able to get more PRBs to satisfy their QoS demands. In addition, low-priority flows can still get one PRB for transmission, so as to prevent them from starvation.

4.2 Stage 2: Queue Trimming and Prediction

When a flow has overdue packets, we need to 'trim' the flow's queue by removing these packets. A packet is called *overdue* if the arrival time at the UE exceeds the delay tolerant time of its flow (mentioned in Section 3). Let $d_{i,j}(t)$ be the HOL packet delay of flow $f_{i,j}$, which is the difference between the current time t and the HOL packet generating time (i.e., when the HOL packet arrived at the queue). Then, the HOL packet will be treated as overdue if

$$d_{i,j}(t) + \xi(p_{i,j}^{HOL}) > \hat{D}_j, \tag{10}$$

where $\xi(p_{i,j}^{HOL})$ is the propagation delay caused by the physical layer to send out the HOL packet $p_{i,j}^{HOL}$ to the destination UE, and $\hat{D}_j$ denotes the delay tolerant time of the flow (referring to Table 1). In this case, the HOL packet should be discarded, and we check the next packet in the queue by Eq. (10), until the new HOL packet is not overdue or the queue becomes empty.

However, the HOL packet of a flow $f_{i,j}$ is not dropped currently but will become overdue in the next TTI. This situation occurs when

$$d_{i,j}(t) + \xi(p_{i,j}^{HOL}) + TTI \geq \hat{D}_j, \tag{11}$$

Practically, we should give $f_{i,j}$ some PRBs to transmit its packets in the current TTI, or it will encounter packet dropping in the following TTI. Therefore, we use an indicator $\text{Drop}_{i,j}$ to mark such flows. In particular, if the HOL packet of a flow $f_{i,j}$ satisfies the condition of Eq. (11) (and thus the eNB predicts that the flow will drop packets in the next TTI), we then set $\text{Drop}_{i,j} = 1$. Otherwise, we set $\text{Drop}_{i,j} = 0$.

We remark that the operations of queue trimming and prediction in this stage is independent of the result from the previous stage. Therefore, the eNB can in fact execute both *Stage 1* and *Stage 2* in parallel (as shown in Fig. 2), so as to save the overall computation time.

4.3 Stage 3: Finding Reallocatable PRBs

After the above two stages, if a flow acquires multiple PRBs, it is possible that the eNB gives the flow 'too many' PRBs. Thus, the eNB will try to ask the flow to return some PRBs (called *reallocatable PRBs*). There are two sources of reallocatable PRBs:

- **Taxed PRBs:** For each flow $f_{i,j}$ whose $\mathrm{Drop}_{i,j} = 0$, the eNB will tax it for a portion of its acquiring PRBs. Here, since flow $f_{i,j}$ will not encounter packet dropping in the next TTI (referring to Eq. (11)), it means that the flow is not urgent for PRBs to send out packets in the current TTI. Therefore, we can actually give flow $f_{i,j}$ fewer PRBs. To tax PRBs from such flows, we use two parameters α and β, where $\alpha \in \mathbb{N}$ and $0 < \beta \leq 1$. Specifically, suppose that a flow $f_{i,j}$ has obtained a number of $n_{i,j} > \alpha$ PRBs from *Stage 1*. The eNodeB then taxes it for a number of

$$\lceil \beta \times (n_{i,j} - \alpha) \rceil \text{ PRBs.} \tag{12}$$

  From Eq. (12), it means that flow $f_{i,j}$ can reserve at least α PRBs for transmission. Then, a β ratio of extra PRBs is expected to be returned to the eNB.
- **Unused PRBs:** *Stage 1* allocates only a 'rough' number of PRBs to each flow, but it does not care about how many data bits can be carried by different PRBs. In fact, when the flow experiences a better channel condition, the eNB can employ a more complex MCS (e.g., 64QAM) to encode data bits in its PRBs. In this situation, the flow can thus use fewer PRBs for communication. To calculate the 'minimum' number of PRBs required by a flow $f_{i,j}$, we adopt the *exponential effective SINR*[2] *mapping (EESM)* scheme [26] to find its effective SINR:

$$\gamma_{eff} = EESM(\gamma, \varepsilon) = -\varepsilon \ln \frac{1}{m} \sum_{k=1}^{m} e^{\frac{-\gamma_k}{\varepsilon}}, \tag{13}$$

  where γ is a vector $[\gamma_1, \gamma_2, \cdots, \gamma_m]$ of the tone SINR value for each subcarrier, m is the number of total subcarriers, and ε is an adjustable parameter (which is usually set to 1). In the LTE implementation, the eNB first uses one PRB to compute γ_{eff} by Eq. (13), and checks whether the PRB is enough to send out the queue content of flow $f_{i,j}$. If not, the eNB iteratively adds one more PRB, recalculates γ_{eff}, and does the above check, until either 1) all $n_{i,j}$ PRBs of $f_{i,j}$ have been used, or 2) the overall data bits of the current PRBs can satisfy $q_{i,j}$, which is the queue length of $f_{i,j}$. In the second case, the residual PRBs of $f_{i,j}$ thus become *unused PRBs*. Apparently, a flow should release all of its unused PRBs, since the flow does not need them to send out packets.

In *Stage 3*, we first use Eq. (12) to compute the taxed PRBs of flows. Then, we check if each flow has unused PRBs. When a flow has unused PRBs, they can be actually counted in the taxed PRBs of that flow. Our reason is that the flow could have better channel quality (so it can use more complex MCS and thus save some PRBs). In this case, we should let the flow reserve more PRBs to improve its data throughput. Therefore, each flow $f_{i,j}$ will return a number of $(\max\{x_{i,j}, y_{i,j}\})$ PRBs to the eNB in *Stage 3*, where $x_{i,j}$ and $y_{i,j}$ respectively denote the number of taxed and unused PRBs.

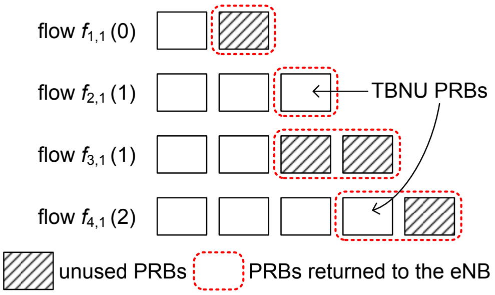


Fig. 4: An example of *Stage 3*, where the figures in brackets denote the number of PRBs taxed from each flow.

We give an example of *Stage 3* in Fig. 4. Suppose that four flows $f_{1,1}$, $f_{2,1}$, $f_{3,1}$, and $f_{4,1}$ are allocated with 2, 3, 4, and 5 PRBs from *Stage 1*. Let $\alpha = 2$ and $\beta = 0.5$. We can thus calculate the number of PRBs that should be taxed from each flow as follows:

Flow $f_{1,1}$: 0 (since it has no more than α PRBs),
Flow $f_{2,1}$: $\lceil 0.5 \times (3-2) \rceil = 1$ PRB,
Flow $f_{3,1}$: $\lceil 0.5 \times (4-2) \rceil = 1$ PRB,
Flow $f_{4,1}$: $\lceil 0.5 \times (5-2) \rceil = 2$ PRBs.

Then, we find unused PRBs from these four flows. According to Fig. 4, flows $f_{1,1}$, $f_{2,1}$, $f_{3,1}$, and $f_{4,1}$ have 1, 0, 2, and 1 unused PRBs, respectively. Therefore, flow $f_{1,1}$ returns its unused PRB to the eNB (even though it is not taxed by any PRB). Flow $f_{2,1}$ is taxed by 1 PRB and thus it returns 1 PRB to the eNB. For flow $f_{3,1}$, it is taxed by 1 PRB but has 2 unused PRBs. Thus, it returns all of the unused PRBs to the eNB (including its tax). Finally, flow $f_{4,1}$ is taxed by 2 PRBs and it has 1 unused PRB. In this case, it not only returns the unused PRB but also gives back one extra PRB to the eNB (to pay the tax). In the example, the data throughput of both flows $f_{1,1}$ and $f_{3,1}$ will not be affected, since they have enough PRBs for transmission. On the other hand, we can avoid significantly decreasing the data throughput of flow $f_{4,1}$, because it actually gives back only one necessary PRB to the eNB.

4.4 Stage 4: PRB Reassignment

According to the flowchart in Fig. 2, if *Stage 3* returns no reallocatable PRB, then the S-DFS algorithm is done in the current TTI. Otherwise, the eNB will execute *Stage 4* to reassign these reallocatable PRBs. Recall that when a flow has unused PRBs, it can pay the tax by returning its unused PRBs to the eNB first. Therefore, we can divide reallocatable PRBs into two categories: *unused* PRBs and *taxed but not unused (TBNU)* PRBs. Fig. 4 gives an example, where we have four unused PRBs and two TBNU PRBs. Apparently, the eNB prefers reassigning unused PRBs first, because TBNU PRBs in fact are required by some other flows (from *Stage 3*). In addition, those flows with $\mathrm{Drop}_{i,j} = 1$ should have a higher priority to obtain the reallocatable PRBs, or their packets will be dropped soon in the next TTI. Based on the above two guidelines, we define the following four rules to reassign the reallocatable PRBs:

- **[Rule 1] The eNB has unused PRBs and there are some flows with $\mathrm{Drop}_{i,j} = 1$:** The eNB first allocates the unused PRBs to those flows with $\mathrm{Drop}_{i,j} = 1$. Flows will use their queue statues and QCI priorities to compete for the unused PRBs. In particular, a flow

[2] Signal-to-interference-plus-noise ratio

(with $\text{Drop}_{i,j} = 1$) that has many small packets in its queue and a higher QCI priority can have precedence over others to get the unused PRBs.

- **[Rule 2] The eNB has unused PRBs but there is no flow with** $\text{Drop}_{i,j} = 1$**:** The eNB allocates these unused PRBs to the flows whose traffic demands have not yet been satisfied. When there are multiple such flows, the eNB gives the unused PRBs to them in a round-robin fashion (from higher QCI priority to lower QCI priority).
- **[Rule 3] The eNB has only TBNU PRBs and there are some flows with** $\text{Drop}_{i,j} = 1$**:** These TBNU PRBs should be also allocated to the flows with $\text{Drop}_{i,j} = 1$, because they are threatened by packet dropping. Similar to *Rule 1*, a flow with more small packets and a higher QCI priority can first get TBNU PRBs for transmission.
- **[Rule 4] The eNB has only TBNU PRBs but there is no flow with** $\text{Drop}_{i,j} = 1$**:** Since no flow will encounter packet dropping in the next TTI, all of the TBNU PRBs are returned to their original owners (in *Stage 3*).

4.5 Rationale of the S-DFS Algorithm

Many flow scheduling schemes (discussed in Section 2) provide a *basic* allocation of PRBs with the goals of increasing transmission efficiency or maintaining UEs' fairness. However, when the eNB has no sufficient PRBs, these schemes may not work well because they ignore the traffic features of different flows. Therefore, by taking into account both the QCI priority and various parameters of each flow, our S-DFS algorithm attempts to further improve the result of PRB allocation by meeting the QoS requirement of GBR flows, while avoiding starving non-GBR flows. To do so, in *Stage 1* we choose the MT scheme to calculate the preliminary PRB allocation, because it can improve transmission efficiency when the network is stable[3]. However, the MT scheme allocates PRBs on the basis of UEs, so we employ a two-phase strategy to assign PRBs to each flow according to their queue lengths. Then, the eNB examines the queue of each flow to find out overdue packets and remove them accordingly by *Stage 2*. In addition, we should identify those flows whose packets will be dropped soon in the next TTI, because they are in urgent need of PRBs. This is done by setting up their indicators $\text{Drop}_{i,j}$ in *Stage 2*.

Then, we try to find out reallocatable PRBs in the network (by *Stage 3*), so as to give them to the flows with $\text{Drop}_{i,j} = 1$. In this way, we can alleviate potential packet dropping of these flows. There are two possible sources for such reallocatable PRBs. First, the eNB can tax those flows which acquire 'too many' PRBs from *Stage 1*. Such a situation may occur when some flows have higher QCI priorities, request a large amount of data transmission, and experience better channel conditions. One representative is a *video* flow (with QCI index = 2). In this case, the preliminary PRB allocation in *Stage 1* could let the video flow acquire a large number of PRBs. Thus, the eNB has to ask these flows to return more PRBs to avoid starving other flows. To do so, the eNB adopts two parameters α and β in *Stage 3* to compute the number of PRBs that should be taxed from the flows whose packets are still safe (i.e., not dropped) in the next TTI. Obviously, Eq. (12) will force a flow with a larger $n_{i,j}$ value to return more PRBs to the eNB, because the flow has acquired a lot of PRBs from *Stage 1*. Second, the eNB can check whether the network remains some resource by carefully calculating the number of data bits actually carried by each PRB. When a flow has good channel quality, the eNB can use a better MCS to encode the data bits in its PRBs. In this case, the flow could spend fewer PRBs to send out its queue content, so the eNB can get back other 'unused' PRBs for later assignment.

TABLE 2: Simulation parameters.

eNB's parameters:	
bandwidth	20 MHz
number of PRBs	100 (12 subcarriers per PRB)
cell range	2 km
frame structure	frequency division duplexing (FDD)
MCS	QPSK, 16QAM, 64QAM
UE's parameters:	
number of UEs	30, 50, 70, 90, 110, 130, 150
mobility model	random direction
moving speed	3 km/h and 120 km/h
GBR flows	VoIP (8.4 Kbps) and H.264 video (242 Kbps)
non-GBR flows	CBR (12 Kbps)
channel's parameters:	
propagation loss	urban macro-cell model
path loss†	$128.1 + 37.6 \log L$
shadowing fading	log-normal distribution with 0 dB mean and 8 dB standard deviation
penetration loss	10 dB
fast fading	Jakes model (for Rayleigh fading)

† L is the distance between the eNB and a UE in kilometers.

Finally, the eNB assigns the reallocatable PRBs to flows by adopting the four rules in *Stage 4*. In particular, we should deal out the unused PRBs first since they are residual resource in the system. Those flows with $\text{Drop}_{i,j} = 1$ should have the precedence to get the reallocatable PRBs to avoid discarding their packets in the next TTI. When there are multiple such flows, we prefer giving PRBs to a flow that possesses a higher QCI priority and its queue contains many small packets (i.e., *Rule 1*). In this case, because a PRB can serve more 'small' packets, we can thus reduce the packet dropping ratio of a high-priority flow by using fewer PRBs. On the other hand, if no flow will suffer from packet dropping in the next TTI (i.e., *Rule 2*), we can employ the round-robin principle to give a relatively fair assignment of PRBs to serve flows. When the eNB has only TBNU PRBs, we should assign them to the flows that will discard packets in the following TTI (i.e., *Rule 3*). However, if there is no flow with $\text{Drop}_{i,j} = 1$, we just return all TBNU PRBs to their original owners (i.e., *Rule 4*), so as to improve their data throughput. The above resource reallocation mechanism helps redistribute a subset of PRBs to those flows in urgent need of resource.

5 Performance Evaluation

In this section, we evaluate the system performance of our S-DFS algorithm by adopting *LTE-Sim* [27], which is an open-source simulator developed for modeling LTE networks. Our experiments aim at investigating the flow scheduling result in an LTE macro-cell, and Table 2 presents the simulation parameters. In every TTI, the eNB is responsible for dealing out 100 PRBs to a number of UEs roaming in the macro-cell. We vary the number of UEs to observe the effect of different amount of traffic loads. UEs will arbitrarily move in the cell according to the random direction model [28]. Both *low-speed* (i.e., 3 km/h) and *high-speed* (i.e., 120 km/h) scenarios are considered in the experiments to measure the impact of

[3] In Section 5.1, we will show that the MT scheme can in fact maximize the cell spectral efficiency when UEs have a lower moving speed.

different moving speeds of UEs. In addition, each UE could generate three types of flows:

1) VoIP flow with data rate of 8.4 Kbps (i.e., 1,000 bits per second). Its QCI index is 1 and thus belongs to GBR flows.
2) H.264 video flow with data rate of 242 Kbps. Its QCI index is 2 and thus belongs to GBR flows.
3) Constant-bit-rate (CBR) flow with data rate of 12 Kbps. Its QCI index is 6 and thus belongs to non-GBR flows.

We compare our S-DFS algorithm with two popular categories of LTE flow scheduling schemes discussed in Section 2:

- **MT-based schemes:** They include both the MT and PPM schemes. For the MT scheme, we use Eq. (10) to help it discard overdue packets and avoid wasting the channel bandwidth. This function is ignored in the original setting of LTE-Sim.
- **PF-based schemes:** They include the M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE schemes. Here, we skip the original PF scheme because the above four schemes are its improved versions by addressing more factors such as the HOL packet delay.

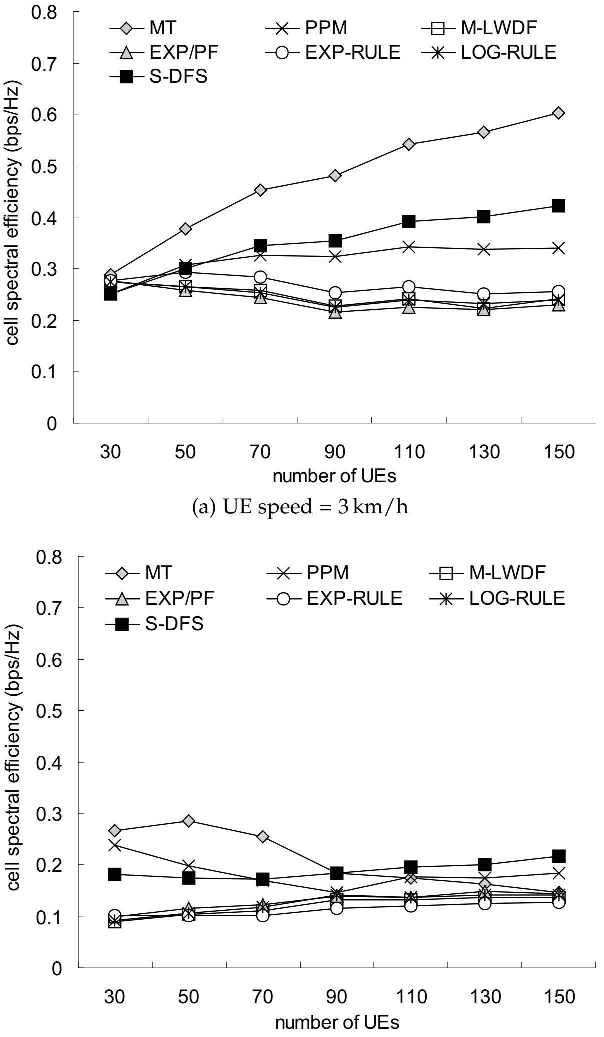
In our S-DFS algorithm, we set $\alpha = 3$ and $\beta = 0.5$ as default. We will further evaluate the effect of these two parameters in Section 5.4. The total simulation time is 180 seconds[4].

5.1 Transmission Efficiency

In order to compare LTE transmission efficiency by different flow scheduling schemes, we use the concept of *cell spectral efficiency*:

$$\psi(t) = \frac{\sum_{\forall i} R_i(t)}{B_{DL}}, \tag{14}$$

where $R_i(t)$ denotes the average data rate of a UE u_i at the measuring time t (i.e., 180 seconds in our simulations), and B_{DL} represents the total bandwidth of the downlink channel. In other words, the cell spectral efficiency is the overall information rate which can be transmitted over a given bandwidth in the LTE system, and it is measured in bits-per-second/hertz (bps/Hz for short). Obviously, a larger $\psi(t)$ value implies that the eNB can utilize the downlink channel more efficiently to serve all UEs, thereby achieving higher transmission efficiency. In this experiment, we thus measure the cell spectral efficiency of MT, M-LWDF, EXP/PF, EXP-RULE, LOG-RULE, PPM, and S-DFS schemes.



(a) UE speed = 3 km/h

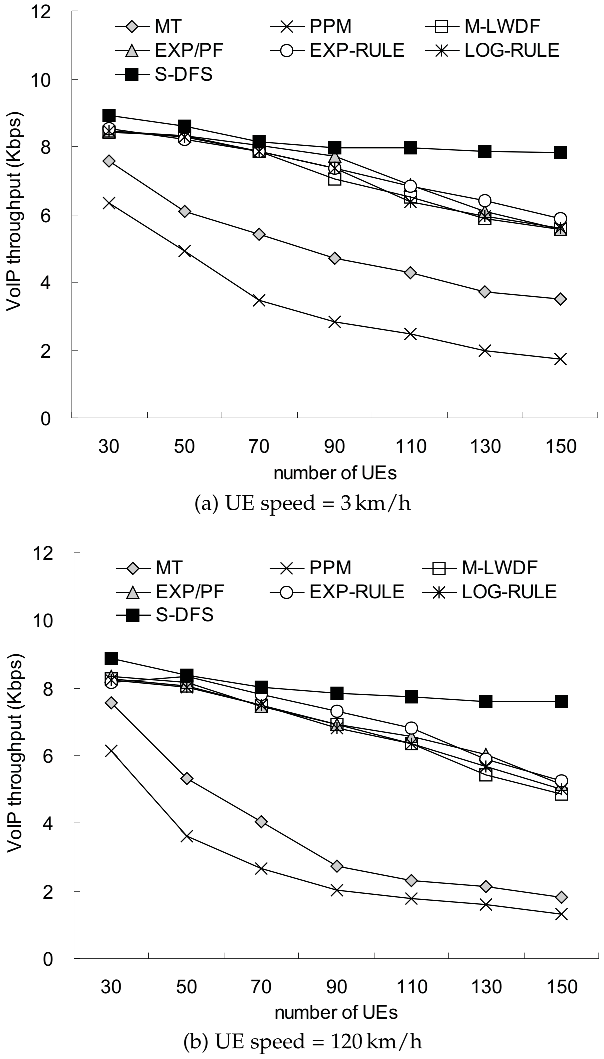
(b) UE speed = 120 km/h

Fig. 5: Comparison on the cell spectral efficiency $\psi(t)$.

Fig. 5(a) compares the cell spectral efficiency $\psi(t)$ in the low-speed scenario, where the average moving speed of UEs is 3 km/h. Since all UEs move slowly, the network becomes relatively stable. Therefore, the MT scheme helps the eNB well utilize the downlink channel by always assigning each PRB to the UE with the best channel quality. When there are more UEs, the eNB can thus increase its $\psi(t)$ value. Both the PPM and S-DFS schemes amend the MT result by reallocating PRBs to some flows, so they can also improve the cell spectral efficiency. However, our S-DFS algorithm works better than the PPM scheme, especially when the number of UEs is more than 50. This verifies that S-DFS can increase LTE transmission efficiency in the case of more UEs. On the other hand, the M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE schemes are improved from the PF scheme, which seek to maintain fair transmissions among UEs. In this case, they could allocate more PRBs to the UEs with bad channel conditions. Therefore, even though the number of UEs grows, their cell spectral efficiency does not change significantly (specifically, $\psi(t) = 0.2\,\text{bps/Hz} \sim 0.3\,\text{bps/Hz}$).

Fig. 5(b) presents the cell spectral efficiency $\psi(t)$ in the high-speed scenario, where the average moving speed of UEs is 120 km/h. In this scenario, the network becomes unstable, which results in lower $\psi(t)$ values for all schemes. In the MT scheme, although the eNB selects a UE with the 'best' channel quality to use each PRB, the channel condition of that UE may degrade very soon due to fast movement. Therefore, as the number of UEs increases, the $\psi(t)$ value of MT scheme decreases drastically. Our S-DFS algorithm allows the eNB to adaptively reallocate some PRBs to other flows based on their queue statues and QCI priorities, so it assists the eNB in better use of the downlink channel. In particular, when the number of UEs exceeds 90, the S-DFS algorithm can result in the largest $\psi(t)$ value. On the other hand, for all PF-based schemes, their $\psi(t)$ values always keep below 0.15 bps/Hz, as they prefer fairly allocating PRBs to flows.

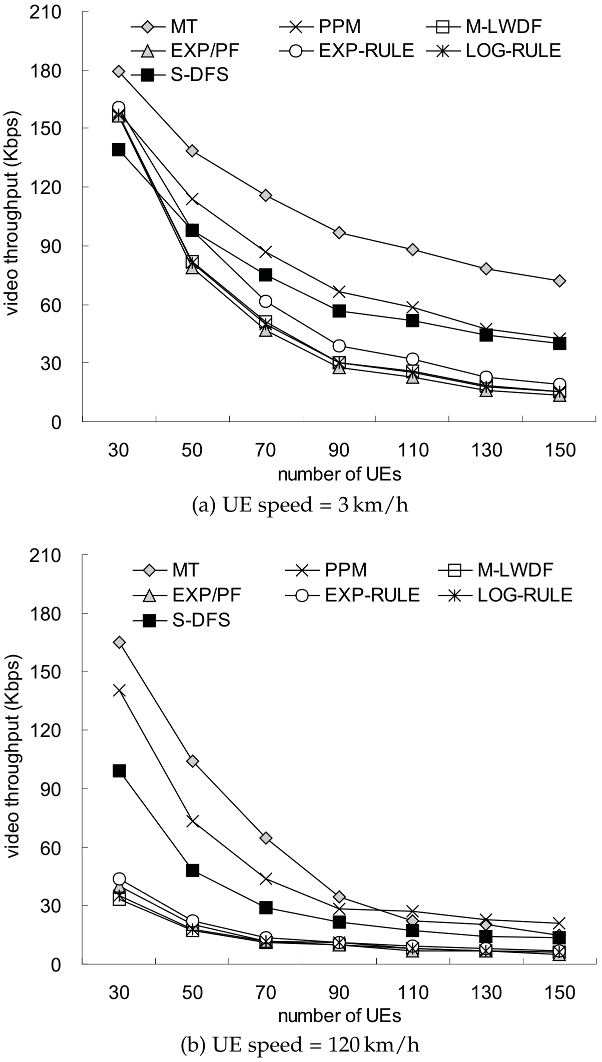
[4] In many studies using LTE-Sim [5], [12], [14], [18], [27], they consider only 100-second simulation time. It thus means that 100 seconds for LTE-Sim simulations is sufficient to prove the performance. However, since the PPM scheme in [21] has the simulation time of 180 seconds, we also extend the simulation time to 180 seconds for fairly comparison.



(a) UE speed = 3 km/h

(b) UE speed = 120 km/h

Fig. 6: Comparison on the average data throughput of VoIP flows.



(a) UE speed = 3 km/h

(b) UE speed = 120 km/h

Fig. 7: Comparison on the average data throughput of video flows.

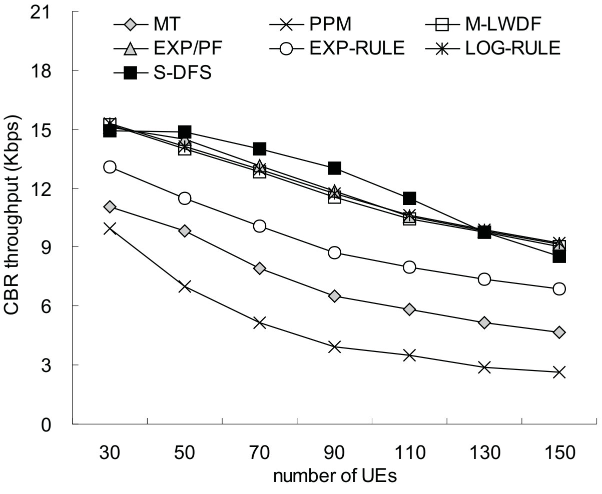
From Fig. 5, we sum up our observations in the experiment:

1) The MT scheme can significantly increase the cell spectral efficiency and it always has the largest $\psi(t)$ value when the network is stable. However, if UEs move in a higher speed, its $\psi(t)$ value degrades fast as the number of UEs grows.
2) Both the PPM and S-DFS schemes can result in higher cell spectral efficiency than PF-based schemes (i.e., M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE). Our S-DFS algorithm can have a larger $\psi(t)$ value than the PPM scheme, especially when there are more UEs competing for the downlink resource.
3) Changing the number of UEs has less impact on the $\psi(t)$ values of the M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE schemes (due to their PF nature).

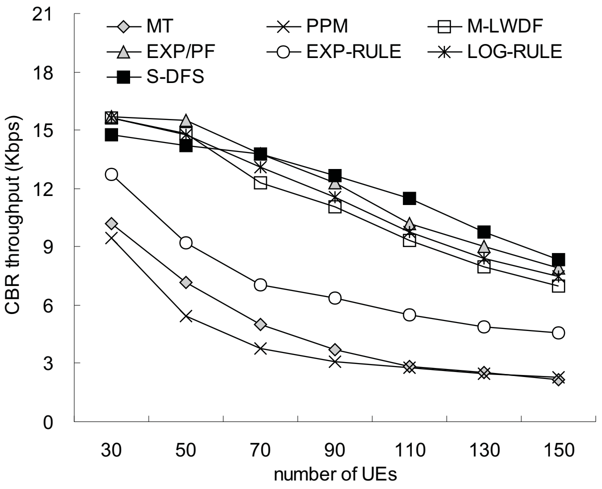
5.2 Data Throughput of Flows

We then measure the data throughput from the viewpoint of flows. Figs. 6, 7, and 8 compare the average data throughput of VoIP, video, and CBR flows by different schemes, respectively. Generally speaking, the data throughput of each flow will decrease when the number of UEs increases, because there are more flows competing for the fixed amount of spectral resource. In addition, increasing the moving speed of UEs has more significant impact on the data throughput of flows with a larger traffic demand (e.g., video flows in Fig. 7). In this case, these flows need to send out large-sized packets and will be affected by the variation of channel condition. Notice that LTE-Sim sets the length of a packet header to 5 bytes but it is not counted in the data rate in Table 2. That is why the VoIP and CBR throughput in Figs. 6 and 8 may exceed 8.4 Kbps and 12 Kbps when the number of UEs is no more than 50, respectively.

The MT scheme prefers allocating PRBs to those UEs with the best channel quality, and these PRBs may be grabbed by their large-demanded video flows. Thus, the VoIP and CBR flows of other UEs with worse channel quality cannot get PRBs for transmission, so they will have lower data throughput. On the other hand, the PPM scheme favors those flows that have a larger amount of data with urgent deadlines. Because video flows satisfy this condition, the PPM scheme will allocate most part of spectral resource to video flows. Thus, both VoIP and CBR flows will be crowded out by these video flows, which results in much lower data throughput. In contrast, the M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE schemes have the goal of maintaining fairness among flows. Because we have traffic demand of video $\gg$ CBR $>$ VoIP, these schemes will result in higher VoIP throughput, lower video throughput, and relatively higher CBR throughput (so as to achieve fair

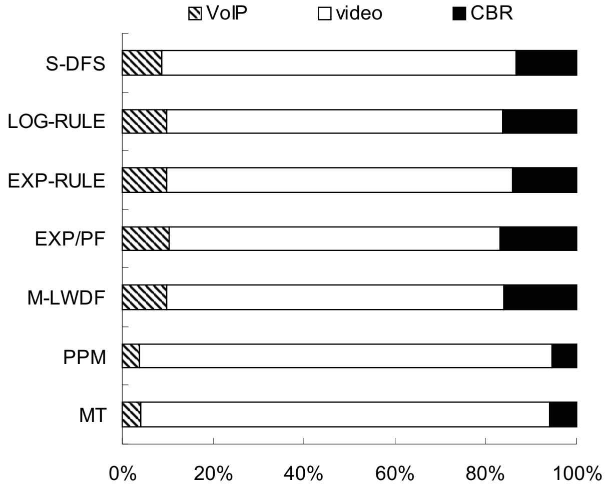


(a) UE speed = 3 km/h

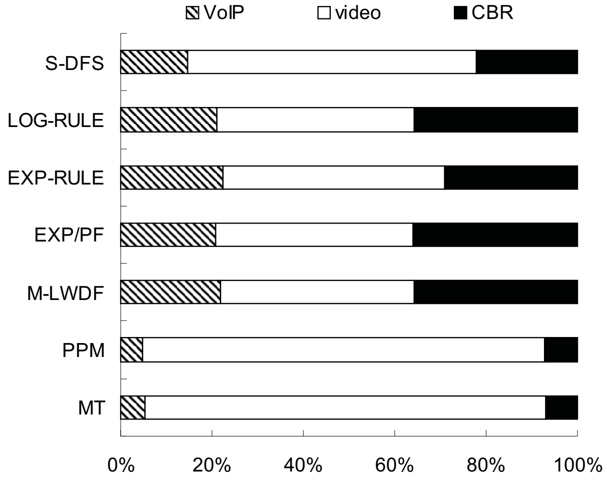


(b) UE speed = 120 km/h

Fig. 8: Comparison on the average data throughput of CBR flows.



(a) UE speed = 3 km/h



(b) UE speed = 120 km/h

Fig. 9: The ratio of data throughput of VoIP, video, and CBR flows by different schemes.

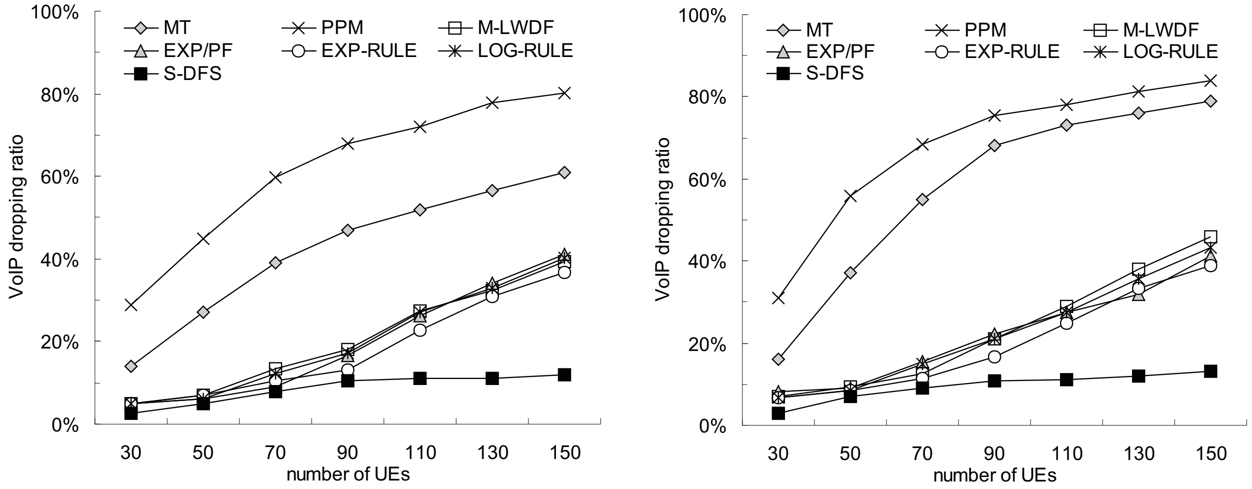
transmission). On the other hand, our S-DFS algorithm takes the MT result as the preliminary PRB allocation but allows the eNB to tax those flows that got too many PRBs in *Stage 1* (e.g., video flows) for reallocatable PRBs. Besides, a flow that will encounter packet dropping in the next TTI can have the precedence over others to acquire such reallocatable PRBs. When there are multiple such flows, a flow with a higher QCI priority and many small packets in its queue (e.g., VoIP flows) can first get these PRBs. Therefore, the S-DFS algorithm can not only maximize the VoIP throughput but also improve the CBR throughput. In addition, compared with the M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE schemes, the S-DFS algorithm can achieve higher data throughput of video flows. The above simulation results demonstrate the effectiveness of our S-DFS algorithm.

Since the amount of spectral resource is constant, there must exist a trade-off between the performance of VoIP, video, and CBR flows. Fig. 9 illustrates the ratio of data throughput of different flows, which takes the average values from Figs. 6, 7, and 8. We discuss the above trade-off by Fig. 9 as follows:

1) For both the MT and PPM schemes, video flows contribute most of the data throughput, which implicitly means that video flows in fact occupy too much resource of the eNB. In this situation, other flows such as VoIP and CBR will be inevitably starved.
2) PF-based schemes seek to 'balance' the data throughput of all flows. This phenomenon is especially obvious when UEs move in a high speed. Although it can achieve more fair transmission among flows, video flows may not receive sufficient resource to meet the QoS requirement.
3) Our S-DFS algorithm allows a video flow to use more resource in order to satisfy its large demand. However, unlike the MT and PPM schemes, the S-DFS algorithm adopts the resource reallocation mechanism (i.e., *Stage 3* and *Stage 4*) to avoid video flows grabbing too much resource. In this way, other flows in urgent need of PRBs (i.e., flows with $\mathrm{Drop}_{i,j} = 1$) can receive resource to transmit data.

5.3 Packet Dropping and Delay of GBR Flows

Then, we focus on investigating the packet transmission behavior of GBR flows. Fig. 10 shows the average packet dropping ratio of VoIP flows. Apparently, when a VoIP flow drops more packets, the corresponding voice quality would degrade. The MT-based schemes (including both MT and PPM) force VoIP flows to drop a large number of packets, because most PRBs are allocated to large-demanded video flows. As the number of UEs grows, VoIP flows will drop more packets. On



(a) UE speed = 3 km/h

(b) UE speed = 120 km/h

UE speed	MT	PPM	M-LWDF	EXP/PF	EXP-RULE	LOG-RULE	S-DFS
3 km/h	42.4%	61.7%	20.4%	19.8%	18.0%	20.1%	8.6%
120 km/h	57.7%	67.7%	23.3%	22.2%	20.1%	22.5%	9.4%
average	50.1%	64.7%	21.9%	21.0%	19.1%	21.3%	9.0%

(c) average packet dropping ratio

Fig. 10: Comparison on the packet dropping ratio of VoIP flows.

the average, the MT and PPM schemes respectively discard 50.1% and 64.7% VoIP packets, so their voice quality will become very bad. On the other hand, the PF-based schemes (including M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE) can help alleviate VoIP packet dropping. However, when there are more than 90 UEs in the LTE cell, the VoIP packet dropping ratio significantly increases (because more flows compete for the limited spectral resource). From Fig. 10(c), we can observe that the average VoIP dropping ratio of the PF-based schemes is around 20%, so the voice quality could be improved.

Because VoIP flows have the highest QCI priority and small packets, it would be favored by our S-DFS algorithm. We can observe from Fig. 10 that the VoIP packet dropping ratio slightly increases as the number of UEs increases. In addition, the average VoIP dropping ratio of our S-DFS algorithm is about 8.6% and 9.4% when the moving speed of UEs is 3 km/h and 120 km/h, respectively, which further reduces more than 11% of VoIP packet dropping compared with the PF-based schemes. Therefore, our S-DFS algorithm is expected to have the best voice quality.

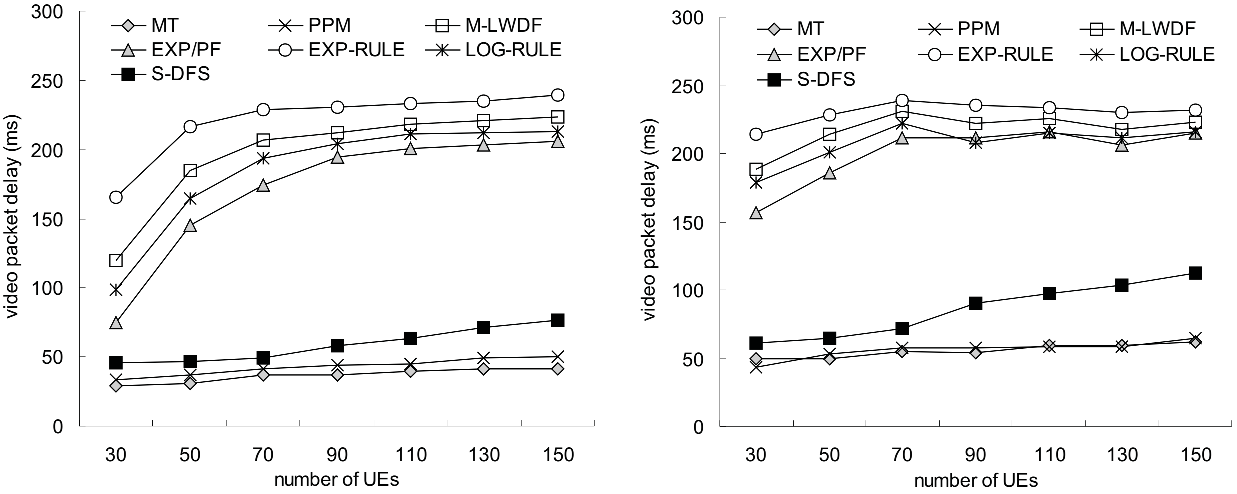
Fig. 11 presents the average packet delay of video flows. As mentioned earlier, the MT-based schemes prefer allocating most of PRBs to the large-demanded video flows. This obviously results in much lower packet delay. On the average, the video packet delay of the MT and PPM schemes is 46.1 ms and 49.6 ms, respectively. On the other hand, the PF-based schemes attempt to keep fair transmissions among flows. Thus, they will not permit video flows to get many PRBs, which causes pretty higher packet delay. From Fig. 11(c), we can observe that the PF-based schemes averagely have around 200 ms of video packet delay. By referring to Table 1, the packet delay budget of video flows (with QCI index = 2) is 150 ms, so the expected delay tolerant time is around 130 ms. Obviously, the video packet delay of the PF-based schemes exceed this tolerant time.

Since video flows have the second largest QCI priority and the largest demand, *Stage 1* of our S-DFS algorithm will give them more PRBs. However, the resource reallocation mechanism will also ask these video flows to return some PRBs to the eNB, in order to avoid allocating too much resource to them. Therefore, the video packet delay of the S-DFS algorithm is between that of the PF-based schemes and the MT-based schemes. In particular, the video packet delay of our S-DFS algorithm is about 58.6 ms and 85.7 ms when the moving speed of UEs is 3 km/h and 120 km/h, respectively, which are quite shorter than the delay tolerant time (i.e., 130 ms). Therefore, our S-DFS algorithm is expected to satisfy the QoS requirement of video flows.

5.4 Effect of Parameters α and β

In *Stage 3* of our S-DFS algorithm, we use two parameters α and β to control the number of PRBs taxed from each flow that has multiple PRBs and $\text{Drop}_{i,j} = 0$. We then evaluate the effect of α and β on the data throughput of different flows. In this experiment, there are 50 and 110 UEs in the LTE cell. Fig. 12 shows the effect of different α values, where we set $\beta = 0.5$. According to Eq. (12), the α value determines how many PRBs can be reserved by each flow. A larger α value means that a flow can keep more PRBs (allocated in *Stage 1*). From Fig. 12, we have the following observations:

- Changing the α value has less impact on the VoIP data throughput. The major reason is that a VoIP flow has the highest QCI priority but smaller amount of traffic demand. Although *Stage 1* first allocates VoIP flows with PRBs, they may not be taxed many PRBs in *Stage 3* (since a VoIP flow can use fewer PRBs to send out its packets). Besides, VoIP flows will be favored in *Stage 4* to get reallocatable PRBs. This phenomenon would counteract the taxing effect by *Stage 3* to some extent.
- Increasing the α value helps improving the video data throughput. Recall that a video flow has the second largest QCI priority and very large amount of traffic demand. Therefore, it is expected to get a lot of PRBs from *Stage 1*. Obviously, if the video flow can keep more



(a) UE speed = 3 km/h

(b) UE speed = 120 km/h

UE speed	MT	PPM	M-LWDF	EXP/PF	EXP-RULE	LOG-RULE	S-DFS
3 km/h	36.6	43.0	198.1	171.1	221.1	185.2	58.6
120 km/h	55.6	56.1	217.6	200.5	230.1	207.6	85.7
average	46.1	49.6	207.9	185.8	225.6	196.4	72.2

(c) average packet delay (ms)

Fig. 11: Comparison on the packet delay of video flows.

PRBs (by increasing the α value) in *Stage 3*, its data throughput will be significantly increased.

- Increasing the α value will decrease the CBR data throughput. Because a CBR flow has the lowest QCI priority, it may not get enough PRBs (or even no PRBs) from *Stage 1*. However, our resource reallocation mechanism in *Stage 3* and *Stage 4* provides an opportunity for CBR flows to receive additional PRBs. Therefore, when the eNB taxes fewer PRBs from flows (due to the increase of parameter α), CBR flows cannot get more PRBs, thereby hurting their data throughput.

On the other hand, Fig. 13 presents the effect of different β values, where we set $\alpha = 3$. According to Eq. (12), the β value determines the ratio of PRBs to be taxed from each flow. A larger β value means that a flow has to return more PRBs to the eNB. Therefore, it is expected that the β value has the opposite effect with the α value, and Fig. 13 proves this argument. In particular, changing the β value has less impact on the VoIP data throughput. In addition, increasing the β value will decrease (respectively, increase) the data throughput of video (respectively, CBR) flows.

5.5 Discussion

TABLE 3: Comparison of different LTE scheduling schemes.

flow	MT-based schemes	PF-based schemes	S-DFS algorithm
VoIP	worst ($>$ 50% dropping)	medium ($>$ 20% dropping)	best ($<$ 10% dropping)
video	best ($<$ 60 ms)	worst ($>$ 200 ms)	medium ($<$ 90 ms)
CBR	worst	medium	best

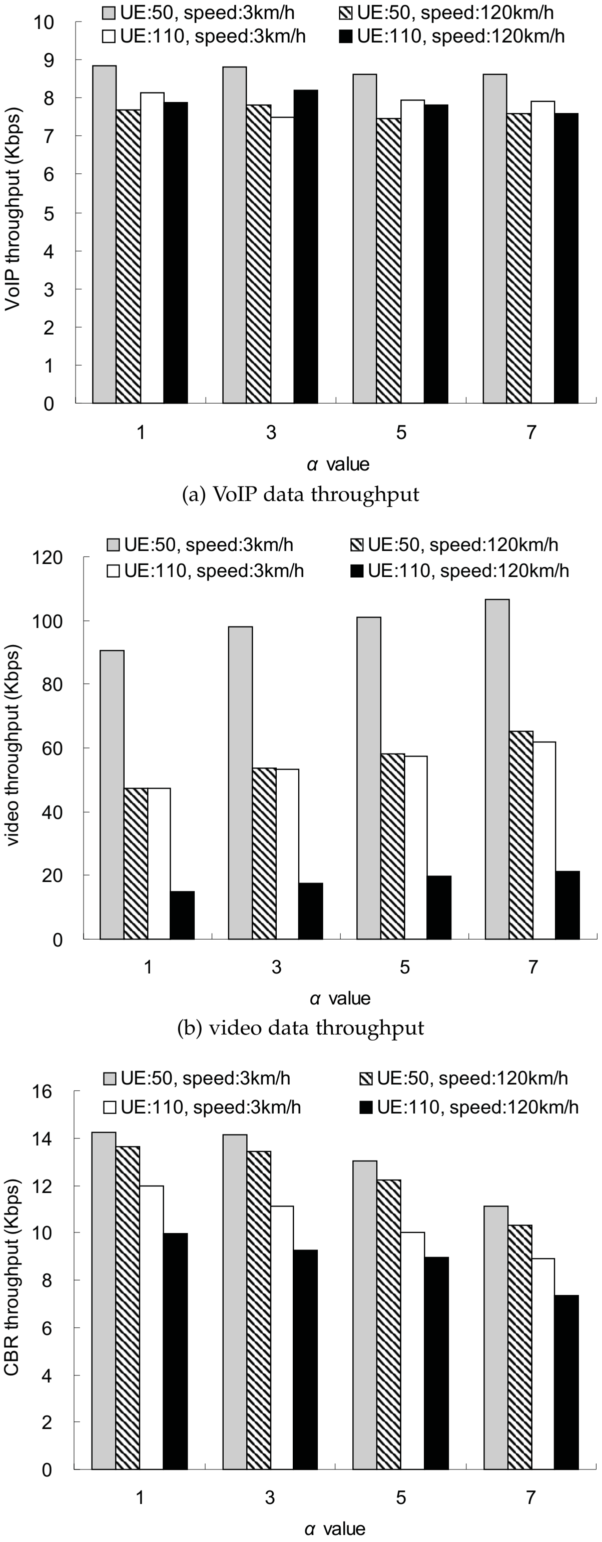
We finally make a discussion on our experiments. From the result in Section 5.1, the MT scheme can achieve the best transmission efficiency when the network is stable. This phenomenon is expectable, because the MT scheme always selects the UE with the best channel quality to get each PRB. Except for the MT scheme, our S-DFS algorithm outperforms other schemes, especially when there are more UEs in the LTE cell. In fact, as UEs move in a higher speed, the S-DFS algorithm can even have better transmission efficiency than the MT scheme, because it can adaptively adjust MT's PRB assignment by the resource reallocation mechanism.

Table 3 summarizes the simulation results in both Section 5.2 and Section 5.3. The MT-based schemes (i.e., MT and PPM) have the best video performance on the cost of other flows. Thus, both VoIP and CBR flows could be starved in the MT-based schemes. On the other hand, the PF-based schemes (i.e., M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE) attempt to keep fair transmissions among flows, so VoIP and CBR flows can have better performance. However, they will significantly increase the average packet delay of large-demanded video flows. Our S-DFS algorithm can result in the best performance of VoIP and CBR flows with the help of the QCI priority and the resource reallocation mechanism. Although it has longer video delay than the MT-based schemes, the S-DFS algorithm still can satisfy the QoS requirement of video flows (according to Table 1, the delay tolerant time of video flows is 130 ms).

In addition, Section 5.4 shows that the data throughput of video and CBR flows can be dynamically changed by adjusting the parameters of our S-DFS algorithm. Specifically, we can further reduce the packet delay of video flows by increasing the α value or decreasing the β value. On the other hand, the data throughput of CBR flows can be improved by decreasing the α value or increasing the β value. This provides certain degree of flexibility to our S-DFS algorithm.

6 Conclusion

Conventional LTE flow scheduling schemes target at either improving the overall transmission efficiency or guaranteeing the fair transmission among flows. However, they do not differentiate flows by their traffic features, which could starve non-GBR flows or may not satisfy the QoS requirement of

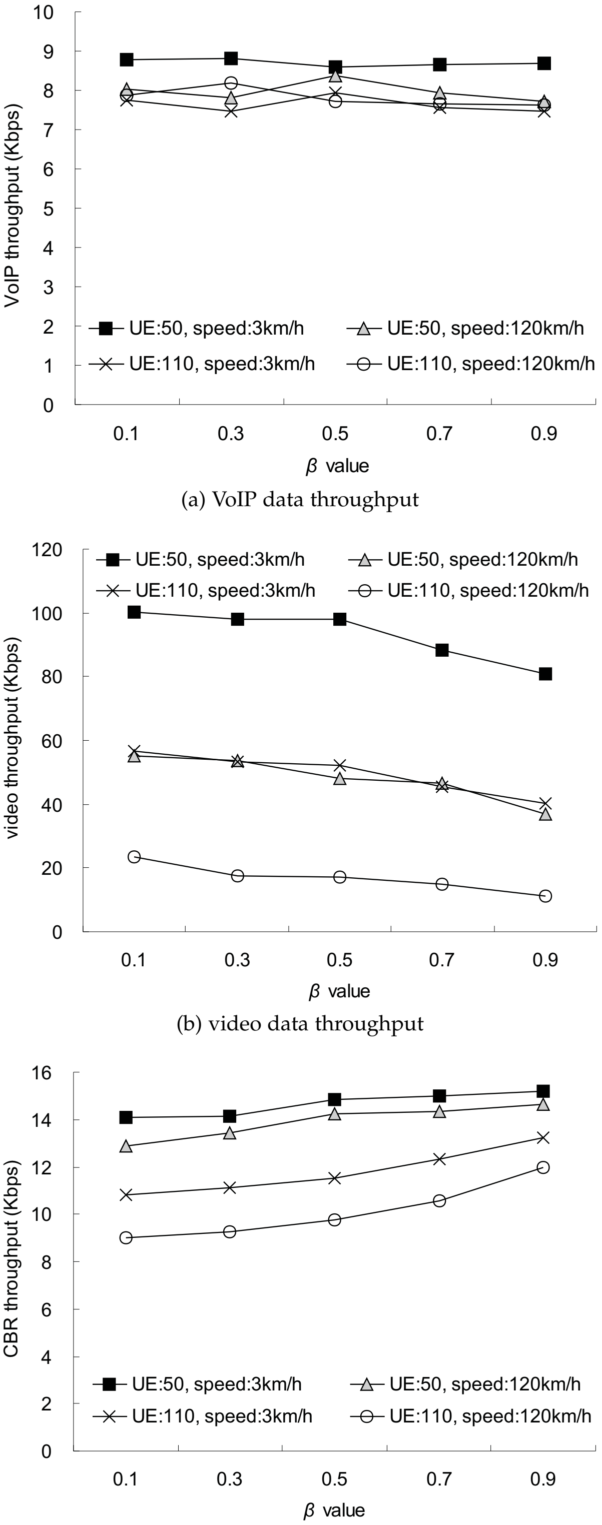


(a) VoIP data throughput

(b) video data throughput

(c) CBR data throughput

Fig. 12: Effect of different α values.



(a) VoIP data throughput

(b) video data throughput

(c) CBR data throughput

Fig. 13: Effect of different β values.

GBR flows. To solve these problems, we develop the S-DFS algorithm that refers to the QCI characteristics and various parameters of flows to allocate PRBs. Furthermore, with the help of the resource reallocation mechanism, the flows in danger of dropping packets can receive additional PRBs to promptly send out their data. By using the LTE-Sim simulator, we compare the proposed S-DFS algorithm with a number of popular LTE flow scheduling schemes, including MT, PPM, M-LWDF, EXP/PF, EXP-RULE, and LOG-RULE. Extensive experimental results have demonstrated the outstanding performance of our S-DFS algorithm, where it can improve the cell spectral efficiency, alleviate the dropping ratio of VoIP packets, reduce the average delay of H.264 video packets, and also keep higher CBR data throughput. In addition, the data throughput of different flows can be dynamically adjusted by changing both the α and β parameters.

REFERENCES

[1] E. Bertin, N. Crespi, and T. Magedanz, *Evolution of Telecomm. Services*, Springer, 2013.

[2] J.G. Andrews, S. Buzzi, W. Choi, S.V. Hanly, A. Lozano, A.C.K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Selected Areas in Comm.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] H. Luo, S. Ci, D. Wu, J. Wu, and H. Tang, "Quality-driven cross-layer optimized video delivery over LTE," *IEEE Comm. Magazine*, vol. 48, no. 2, pp. 102–109, 2010.

[4] Cisco, "Cisco visual networking index: forecast and methodology, 2014–2019 white paper," May 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

[5] F. Capozzi, G. Piro, L.A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: key design issues and a survey," *IEEE Comm. Surveys & Tutorials*, vol. 15, no. 2, pp. 678–700, 2013.

[6] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M.Moisio, "Dynamic packet scheduling performance in UTRA long term evolution downlink," *Proc. IEEE Int'l Symp. Wireless Pervasive Computing*, 2008, pp. 308–313.

[7] H.J. Kushner and P.A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Comm.*, vol. 3, no. 4, pp. 1250–1259, 2004.

[8] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Comm. Magazine*, vol. 39, no. 2, pp. 150–154, 2001.

[9] J.H. Rhee, J.M. Holtzman, and D.K. Kim, "Scheduling of real/non-real time services: adaptive EXP/PF algorithm," *Proc. IEEE Vehicular Technology Conf.*, 2003, pp. 462–466.

[10] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in LTE," *EURASIP J. Wireless Comm. and Networking*, vol. 2009, pp. 1–18, 2009.

[11] H. Luo, S. Ci, D. Wu, J. Wu, and H. Tang, "Quality-driven cross-layer optimized video delivery over LTE," *IEEE Comm. Magazine*, vol. 48, no. 2, pp. 102–109, 2010.

[12] G. Piro, L.A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-level downlink scheduling for real-time multimedia services in LTE networks," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1052–1065, 2011.

[13] K.J. Astrom and B. Wittenmark, *Computer-Controlled Systems: Theory and Design*, Prentice-Hall, 1996.

[14] M. Iturralde, T.A. Yahiya, A. Wei, and A.L. Beylot, "Resource allocation using Shapley value in LTE networks," *Proc. IEEE Int'l Symp. Personal Indoor and Mobile Radio Comm.*, 2011, pp. 31–35.

[15] A.E. Roth, *The Shapley Value*, Cambridge University Press, 1988.

[16] S. Ali and M. Zeeshan, "A utility based resource allocation scheme with delay scheduler for LTE service-class support," *Proc. IEEE Wireless Comm. and Networking Conf.*, 2012, pp. 1450–1455.

[17] D. Niyato and E. Hossain, "A cooperative game framework for bandwidth allocation in 4G heterogeneous wireless networks," *Proc. IEEE Int'l Conf. Comm.*, 2006, pp. 4357–4362.

[18] B. Liu, H. Tian, and L. Xu, "An efficient downlink packet scheduling algorithm for real time traffics in LTE systems," *Proc. IEEE Consumer Comm. and Networking Conf.*, 2013, pp. 364–369.

[19] D. Liu and Y.H. Lee, "An efficient scheduling discipline for packet switching networks using earliest deadline first round robin," *Proc. IEEE Int'l Conf. Computer Comm. and Networks*, 2003, pp. 5–10.

[20] C. Wang and Y.C. Huang, "Delay-scheduler coupled throughput-fairness resource allocation algorithm in the long-term evolution wireless networks," *IET Comm.*, vol. 8, no. 17, pp. 3105–3112, 2014.

[21] W.K. Lai and C.L. Tang, "QoS-aware downlink packet scheduling for LTE networks," *Computer Networks*, vol. 57, no. 7, pp. 1689–1698, 2013.

[22] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-QoS-aware fair scheduling for LTE," *Proc. IEEE Vehicular Technology Conf.*, 2011, pp. 1–5.

[23] M.S. Mushtaq, A. Shahid, and S. Fowler, "QoS-aware LTE downlink scheduler for VoIP with power saving," *Proc. IEEE Int'l Conf. Computational Science and Engineering*, 2012, pp. 243–250.

[24] European Telecommunications Standards Institute, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 13)," Technical Report, 3GPP TS 23.203 V13.3.0, 2015.

[25] European Telecommunications Standards Institute, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 8)," Technical Report, 3GPP TS 23.203 V8.9.0, 2010.

[26] R. Giuliano and F. Mazzenga, "Exponential effective SINR approximations for OFDM/OFDMA-based cellular system planning," *IEEE Trans. Wireless Comm.*, vol. 8, no. 9, pp. 4434–4439, 2009.

[27] G. Piro, L.A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: an open-source framework," *IEEE Trans. Vehicular Technology*, vol. 60, no. 2, pp. 498–513, 2011.

[28] W.H. Yang, Y.C. Wang, Y.C. Tseng, and B.S.P. Lin, "Energy-efficient network selection with mobility pattern awareness in an integrated WiMAX and WiFi network," *Int'l J. Comm. Systems*, vol. 23, no. 2, pp. 213–230, 2010.