A Pricing-Aware Resource Scheduling Framework for LTE Networks

You-Chiun Wang and Tzung-Yu Tsai

Abstract—Long term evolution (LTE) is a standard widely used in cellular networks today. Both resource scheduling and pricing are two critical issues. However, existing studies address them separately, making the goals of improving system performance and increasing operator revenue conflicting. The paper proposes a *pricing-aware resource scheduling (PARS) framework* to conquer this conflict. It classifies users into three levels and has *scheduling* and *pricing* modules, which are installed in a base station and the core network of LTE, respectively. The scheduling module uses three-layer schedulers to assign resource to a flow by considering its packet delay, traffic amount, channel condition, and user level. The pricing module uses *price elasticity of demand* in economics to adaptively adjust the amount of money charged to users. Through experiments by LTE-Sim, we show that PARS achieves a good balance between performance and revenue, and provides quality of service for the flows with strict delay concerns.

Index Terms—cellular network, long term evolution (LTE), pricing, quality of service (QoS), resource scheduling.

1 Introduction

Long term evolution (LTE) has now been operated in many countries to provide 4G service. Comparing with past systems, LTE exploits some efficient techniques, including orthogonal frequency division multiple access (OFDMA) [1], carrier aggregation [2], and heterogenous cells [3], to provide high-speed wireless access. Therefore, people can freely use various broadband applications such as multimedia streaming and video downloads on their mobile phones.

In an LTE cell, the base station called eNodeB (also abbreviated to 'eNB') takes charge of scheduling spectral resource to user equipments (UEs). With OFDMA, the downlink resource is concretized by a 2D array of physical resource blocks (PRBs) in time and frequency domains. Each PRB carries different number of data bits, depending on the channel quality of a UE in respect of that PRB. In general, LTE performance is decided by the way that the eNB allocates PRBs to UEs, which we call *LTE resource scheduling*, and many methods have been developed. They aim at improving *system performance* by, for example, increasing network throughput, keeping fair transmission, or supporting quality of service (QoS) [4].

On the other hand, the pricing policy plays an important role in cellular networks, as it significantly affects *operator revenue*. Most operators classify users into different levels based on the pricing categories. Higher-level users are charged with a higher rate but can enjoy more resource. According to [5], pricing policies are categorized into *static* and *dynamic*. In a static pricing policy, users pay a fixed rate no matter how their traffic loads increase. A dynamic pricing policy charges more money when the user's load exceeds a threshold. In this way, it can help increase the operator's revenue.

Unfortunately, the goals of improving system performance and increasing operator revenue may conflict, especially when network resource is insufficient. Let us consider an example with two-level users. Many high-level users encounter bad channel quality, while most low-level users have good channel quality. To improve performance, one would give more PRBs to low-level users, thereby diminishing revenue. On the contrary, if we give more PRBs to high-level users, the revenue increases but the performance degrades. However, the investigation of resource scheduling and pricing in LTE is independent. Consequently, it motivates us to integrate LTE resource scheduling with a pricing policy, so as to achieve a good balance between performance and revenue.

This paper develops a *pricing-aware resource scheduling (PARS) framework* based on the above observation. Without loss of generality, we classify users into three levels: golden (high), silver (medium), and bronze (low). PARS consists of both *scheduling* and *pricing* modules. The scheduling module employs a three-layer scheduling strategy. It first estimates the amount of resource used to support QoS for guaranteed-bit-rate (GBR) flows, then allocates PRBs to each flow, and finally checks if some PRBs can be reallocated to improve performance. The pricing module follows the *price elasticity of demand (PED)* model [6], where a user's demand is affected by the price. It then adaptively computes the amount of money charged to users, depending on their resource consumption.

Our contributions are threefold. First, this paper indicates that existing studies may face the dilemma of improving performance or increasing revenue, as they solve the problems of resource scheduling and pricing separately. Second, we propose the PARS framework to conquer the dilemma by both scheduling and pricing modules, which work from perspectives of engineering (i.e., resource allocation) and economics (i.e., pricing with PED), respectively. Third, each module will refer to the outcome of the other to make its decision, so the results of scheduling and pricing in PARS will tightly couple. Extensive simulation results exhibit that PARS can increase the operator's revenue, improve spectral efficiency, support QoS for GBR flows, and ensure non-GBR transmissions.

This paper is organized as follows. Section 2 introduces LTE while Section 3 surveys related work. We propose the PARS framework in Section 4 and give some analyses in Section 5. Section 6 evaluates performance and Section 7 concludes the paper. We then summarize acronyms in Table 1.

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; superhawk236@gmail.com

TABLE 1: Summary of common acronyms.

acronym	full name
CBR	constant-bit-rate
CQI	channel quality indicator
FLS	frame layer scheduling
GBR	guaranteed-bit-rate
HOL	head-of-line
LTE	long term evolution
MCS	modulation and coding scheme
M-LWDF	modified largest weighted delay first
NLP	network load based pricing
PARS	price-aware resource scheduling
PCRF	policy control and charging rule function
PED	price elasticity of demand
PF	proportional fairness
P-GW	packet data network gateway
PRB	physical resource block
QCI	QoS (quality of service) class identifier
SCP	subscriber class based pricing
TTI	transmission time interval
UE	user equipment

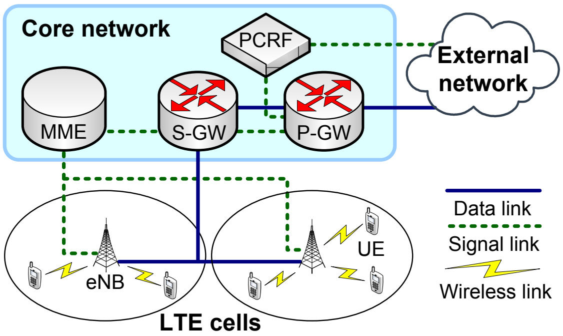


Fig. 1: LTE structure, where we omit some components in the core network.

2 LTE Overview

2.1 Network Structure

LTE network consists of multiple *cells* and a *core network*, as Fig. 1 shows. UEs are served by the eNB in a cell. The core network deals with the management job, and has three main components: 1) *Mobility management entity (MME)* processes signaling between each UE and the core network. 2) *Serving gateway (S-GW)* routes data packets and acts as the mobility anchor when a UE moves among cells. 3) *Packet data network gateway (P-GW)* connects to the external network. It also performs policy enforcement and supports user charging.

Charging control is done by the cooperation of *policy and charging rules function (PCRF)* and P-GW [7]. PCRF is the decision center to manage each flow in P-GW, and checks if the flow's behavior follows its subscription profile. PCRF has an *application function* to provide dynamic charging and QoS data to check flows. LTE supports *offline* and *online* charging. Offline charging provides statistics for event- and session-based charging. Online charging helps P-GW terminate a user's service when certain conditions are met (e.g., when the amount of traffic exceeds the limitation). Based on this structure, PARS's pricing module can be installed in PCRF.

2.2 Downlink Communication

LTE divides resource into non-overlapping PRBs, each with 0.5ms duration and 180kHz bandwidth. PRBs are non-sharable[1], so a PRB cannot be given to multiple UEs. The eNB is responsible for allocating PRBs, in which PARS's scheduling module is installed. The minimum period to allocate PRBs is called a transmission time interval (TTI = 1ms). When the bandwidth of downlink channel is 1.4, 3, 5, 10, 15, or 20 MHz, the eNB can provide 6, 15, 25, 50, 75, or 100 PRBs in a TTI.

Through a *modulation and coding scheme (MCS)*, each PRB carries different number of bits. In general, a more complex MCS allows the PRB to carry more data, but it requires the UE to have better channel condition. To help the eNB select the proper MCS, each UE has to report the *channel quality indicator (CQI)*, which reveals its channel quality in every TTI. A UE can have multiple flows, where each flow has a queue at the eNB to be its packet container. Packets are stamped with arrival time once they are generated, and the eNB sends a queue's packets in a FIFO manner. The head-of-line (HOL) packet delay of a flow is defined by the elapsing time of the first packet in the queue after its arrival.

LTE uses *QoS class identifier (QCI)* to depict the QoS demand of a flow, which includes *packet delay budget* and *packet loss rate*. The packet delay budget is the maximum tolerant time that each packet can be delayed between P-GW and its UE. When the delay of a packet exceeds the budget, the packet is invalid. The packet loss rate limits the maximum probability that a packet is not received by its UE (e.g. due to interference or expiration). LTE categorizes flows into GBR and non-GBR ones. GBR flows mainly support real-time applications with strict delay constraints, such as VoIP, live-streaming video, and on-line games. Non-GBR flows are often used for other service with loose deadlines (e.g., TCP-based service). Thus, GBR flows usually have smaller QCI values and packet delay budgets than non-GBR flows.

3 Related Work

3.1 LTE Resource Scheduling

LTE standards leave the resource scheduling problem to implementers, so various solutions are developed. Capozzi et al. [4] survey some popular solutions below: *Max-CQI* uses a greedy principle to allocate each PRB to the UE with the maximum channel rate r_i. *Proportional fair (PF)* considers the average channel rate r_i^{avg} to support fairness, and it iteratively picks the UE with the largest r_i/r_i^{avg} value to receive resource. *Modified largest weighted delay first (M-LWDF)* adds a weight w_i and HOL packet delay d_i to the PF solution to reduce delay. *Exponential proportional fair* introduces a term $\exp[(w_i d_i - d_{\text{avg}})/(1+\sqrt{d_{\text{avg}}})]$ to the PF solution, where d_{avg} is the average packet latency. Both *LOG-RULE* and *EXP-RULE* refer to the spectral efficiency ψ_i of a UE, and select a flow that has the maximum value of $(\psi_i \log X)$ and $(\psi_i \exp Y)$ to get each PRB, where X and Y are terms defined in LOG-RULE and EXP-RULE, respectively.

Some work adopts a multi-layer strategy to allocate PRBs. Luo et al. [8] develop a cross-layer framework to support video delivery. They refer to the delay requirement, signal distortion, and past rate of each video flow to decide its PRB allocation and coding scheme. In [9], a two-layer scheduler is designed to support multimedia service. One layer computes the amount of data that each flow has to send in a TTI to meet its delay

[1] It occurs when the network uses SISO (single-input single-output) or SU-MIMO (single-user multiple-input and multiple-output) for communication.

demand by the discrete-time linear control theory. Then, the other layer gives PRBs to each flow by using PF for fairness concern. The work of [10] proposes a double-layer scheme to schedule LTE downlink resource. The first layer translates the scheduling problem to a bankruptcy game and then solves it by the Shapley value. Based on the result, the next layer allocates PRBs according to EXP-RULE.

A number of approaches reduce real-time packet dropping by considering their deadlines. The work of [11] applies the earliest-deadline-first method to PF, so as to support fairness while ensure that the packets whose deadlines will expire soon can be sent first. In [12], a *virtual queue* is used to predict the incoming of future packets based on the existing packets in each queue. Then, [12] discards the packets that cannot satisfy their delay demands to avoid unnecessary transmission. The study of [13] divides flows into *urgent* and *non-urgent* ones, where urgent flows are given with a high priority to send their packets. Non-urgent flows, including non-real-time flows and real-time flows whose packets have not expired yet, are given with the same (low) priority for transmission. Wang and Hsieh [14] use max-CQI to compute the preliminary PRB allocation, and tax non-urgent flows with *reallocatable PRBs*. Such PRBs are given to those flows in danger of packet dropping.

Few studies combine resource scheduling with other factors. For example, [15] proposes a scheduling method with power saving. Each UE is assigned with a priority $(F_i(r_g/r_i^{\text{avg}})^2 + Q_i) \times \hat{d}_i\varepsilon_i$ if it is a GBR UE, and $(F_i + Q_i) \times \varepsilon_i$ otherwise, where F_i uses the PF concept, r_g is the average throughput of GBR UEs, Q_i is u_i's queue status, $\hat{d}_i$ is a delay factor, and ε_i is a DRX (discontinuous reception) indicator for power saving. Then, UEs can use their priorities to compete for PRBs.

To the best of our knowledge, none of existing work considers integrating resource scheduling with a pricing policy in LTE. This motivates us to develop the PARS framework with both scheduling and pricing modules, so as to balance between system performance and operator revenue.

3.2 Pricing in Cellular Networks

Past 2G networks use circuit switching for communication, so operators can simply charge each call by its duration. After 2.5G, the technique changes to packet switching, and thus 2G pricing becomes inapplicable [16]. Hence, various pricing policies are developed in response to the technical change [17]: 1) *Fixed price charging* sets a constant rental fee for users, so the operator need not record bandwidth consumption. However, the operator cannot increase its revenue when network traffic grows, and some users may overuse the network. 2) *Metered charging* asks users to pay for network connection on a monthly basis and charge them for metered usage of the service. However, the usage is measured by time, so it is unfair for the users who leave sessions open without sending packets. 3) *Packet charging* computes the expense charged to a user based on the number of packets sent in a session. It provides accurate pricing but relies on a packet counting method, which complicates the billing system. 4) *Expected capacity charging* lets users pay for different amount of money by their expected bandwidth usage, so the price to each user is predictable. However, the operator has to continually monitor the actual bandwidth spent by each user. 5) *Edge pricing* aims at the case when a user stays in two cells such that his packets are relayed by two base stations. This policy simplifies the charging mechanism by making each base station consult local charging information, without exchanging their billing data. 6) *Paris-Metro charging* allows users to assign a preferred class with an associated cost for their different traffic (e.g., business mail is viewed more important than personal mail, so a high class is given to business mail). The policy provides flexility, but it also adds overhead to users for traffic-class decision.

Different pricing methods for 3G and 4G networks are also proposed. The *flat-rate pricing method* [5] works like fixed price charging, where the fee will not change no matter how network traffic grows. The *fixed-PRB pricing method* [18] divides users into golden, silver, and bronze levels. It charges a user by the level l_i and the number n_i of PRBs used:

$$C_i = \mathcal{P}_f(l_i) \times n_i, \tag{1}$$

where $\mathcal{P}_f(\cdot)$ is the fixed charge by levels (in units of PRB). The *network load based pricing (NLP) method* [19] considers both network load $\mathcal{L}$ and QCI. When $\mathcal{L}$ increases, users are charged for more money, so the operator's revenue increases accordingly. Specifically, each user is charged for

$$C_i = \mathcal{P}_v(l_i) \times (\hat{e} - \hat{e}^{-\alpha x}) \times \mathcal{L}, \tag{2}$$

where $\hat{e}$ is the Euler's number and α restricts the QCI value x to [1..9]. In Eq. (2), $\mathcal{P}_v(\cdot)$ is the variable charge and depends on a load threshold δ. When $\mathcal{L} \leq \delta$, $\mathcal{P}_v(l_i)$ is set to a constant $\mathcal{P}_c$, which means that each user is charged fairly if network load is light. Otherwise, we set $\mathcal{P}_v(\mathbf{G}) > \mathcal{P}_v(\mathbf{S}) > \mathcal{P}_v(\mathbf{B})$, where $\mathbf{G}$, $\mathbf{S}$, and $\mathbf{B}$ denote golden, silver, and bronze levels, respectively. The *subscriber class based pricing (SCP) method* [20] also considers three-level users. When $\mathcal{L} \leq \delta$, it charges users by Eq. (1). When $\mathcal{L} > \delta$, SCP charges users by

$$C_i = \begin{cases} (\mathcal{P}_f(\mathbf{G}) + \mathcal{P}_e) \times n_i & \text{if } l_i = \mathbf{G}, \\ (2\mathcal{P}_f(\mathbf{G}) + \mathcal{P}_e) \times n_i & \text{if } l_i = \mathbf{S}, \\ (2\mathcal{P}_f(\mathbf{G}) + \mathcal{P}_f(\mathbf{S}) + \mathcal{P}_e) \times n_i & \text{if } l_i = \mathbf{B}, \end{cases} \tag{3}$$

where $\mathcal{P}_e$ is the extra charging computed by $\kappa/(n_A - n_G)$. Here, κ is a pricing constant, n_A is the number of total PRBs, and n_G is the number of PRBs reserved for golden users. SCP is expected to greatly increase operator revenue when the network becomes overloaded. However, such high pricing in Eq. (3) also degrades users' willingness to use the service, thereby hurting system performance. That is why we propose the PARS framework to integrate resource scheduling with pricing. PARS also adopts PED to avoid overcharging users, so as to improve performance while keeping high revenue.

4 The PARS Framework

This section proposes the PARS framework. We first give the assumption and architecture of our framework. Then, we present both scheduling and pricing modules in PARS, followed by the design rationale. Afterward, we discuss how to extend the PARS framework to the multi-cell environment.

4.1 Assumption and Architecture

We classify users into levels of golden ($\mathbf{G}$), silver ($\mathbf{S}$), and bronze ($\mathbf{B}$), with priorities of $\mathbf{G} > \mathbf{S} > \mathbf{B}$. The eNB assigns PRBs to UEs by referring to these priorities. PCRF then measures the amount of resource spent by each UE and charges its user accordingly, where golden, silver, and bronze users are charged with high, medium, and low unit prices, respectively. Some previous studies restrict the type of flows that a UE can

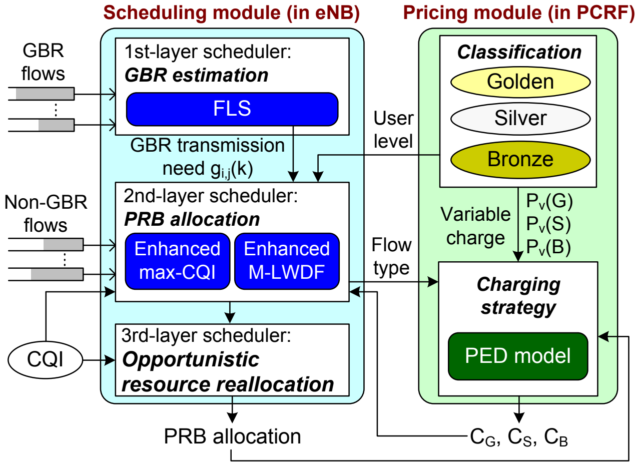


Fig. 2: System architecture of the PARS framework.

TABLE 2: Summary of notations used in PARS.

notation	definition
$d_{i,j}$	HOL packet delay of flow $f_{i,j}$
$f_{i,j}$	the jth flow of a UE u_i
$g_{i,j}(k)$	flow $f_{i,j}$'s transmission need in the kth frame
l_i	user level of UE u_i
$p_{i,j}$	a PRB allocated to flow $f_{i,j}$
$q_{i,j}(k)$	flow $f_{i,j}$'s queue length in the kth frame
r_i, r_i^{avg}	the current and average channel rate of UE u_i
$w_{i,j}$	the weight defined by M-LWDF for flow $f_{i,j}$
C_i	the amount of money charged to UE u_i
E_d	coefficient of price elasticity
$\mathbf{G}, \mathbf{S}, \mathbf{B}$	golden, silver, and bronze levels
$\mathcal{L}$	network load
$\mathcal{P}_c, \mathcal{P}_e$	constant and extra charges
$\mathcal{P}_f(\cdot), \mathcal{P}_v(\cdot)$	fixed and variable charges by the user level
W_i^P	price-based weight for UE u_i
δ	load threshold
ξ_{GBR}, ξ_{NGBR}	two ratios for the portion of reallocated PRBs
$\Psi_{GBR}^R, \Psi_{NGBR}^R$	two sets of reallocated PRBs

use (e.g., [20] does not allow bronze UEs to have GBR flows). In this paper, we assume that each UE can transmit any type of flows for flexibility and practicability.

Fig. 2 gives the system architecture of our PARS framework, which consists of scheduling and pricing modules installed in an eNB and PCRF, respectively. The scheduling module contains three layers of schedulers. The 1st-layer scheduler, *GBR estimation*, measures how much resource that each GBR flow requires to satisfy its delay constraint. We borrow the idea from *frame layer scheduling (FLS)* [9] to do the measurement. The 2nd-layer scheduler, *PRB allocation*, then decides the number of PRBs given to each flow based on the GBR transmission need (from the lst-layer scheduler), user level (from the pricing module), and CQI. It enhances the max-CQI and M-LWDF methods [4] to cope with PRB allocation for GBR and non-GBR flows, respectively. The 3rd-layer scheduler, *opportunistic resource reallocation*, finally checks whether it is possible to exchange the usage of some PRBs by different levels of UEs, so as to improve system performance. It also involves extra charge to some users who acquire additional PRBs by this scheduler, and such information will feed back to the pricing module for calculation.

The pricing module provides variable charges for different levels of users. The *charging strategy* is the core of this module. It takes the PED model into consideration, where the user demand for service will depend on its price and flow type. Then, the charging strategy adaptively adjusts the fee charged to each flow according to the information of flow type and PRB allocation from the scheduling module. Apparently, the amount of money charged to a user will be the sum of fees on all flows that his UE uses. Next, we present the detailed design of both scheduling and pricing modules. Table 2 summarizes the notations used in the PARS framework.

4.2 Scheduling Module

4.2.1 GBR Estimation

The 1st-layer scheduler considers only GBR flows, as shown in Fig. 2. We adopt FLS to evaluate the amount of data that should be transmitted for a GBR flow to satisfy its delay requirement (called *transmission need*). Specifically, let $q_{i,j}(k)$ and $g_{i,j}(k)$ be the queue length of a GBR flow $f_{i,j}$ and its transmission need in the kth frame, respectively, where $0 \leq q_{i,j}(k) \leq q^{\max}$, $g_{i,j}(k) \geq 0$, and $q^{\max}$ denotes the maximum size of a queue. Then, the variation in queue length can be described by the following equation:

$$q_{i,j}(k+1) - q_{i,j}(k) = \phi_{i,j}(k) - g_{i,j}(k), \tag{4}$$

where $\phi_{i,j}(k)$ is the amount of newly generated data to $f_{i,j}$'s queue in the kth frame, and $\phi_{i,j}(k) \geq 0$. To calculate the transmission need of $f_{i,j}$, FLS defines a control rule by

$$g_{i,j}(k) = h_{i,j}(k) * q_{i,j}(k), \tag{5}$$

where '$*$' denotes the discrete-time convolution and $h_{i,j}(k)$ is a pulse-response function. According to [9], we can derive $g_{i,j}(k)$ by combining Eqs. (4) and (5) as follows:

$$\begin{aligned} g_{i,j}(k) = \; & q_{i,j}(k) + \sum_{m=2}^{M_{i,j}} \hat{c}_{i,j}(m) \times (q_{i,j}(k-m+1) - \\ & q_{i,j}(k-m+2) - g_{i,j}(k-m+1)). \end{aligned} \tag{6}$$

In Eq. (6), $M_{i,j}$ is the sampling interval for the flow. We can set $M_{i,j} = d_{i,j}^{\max} - 1$, where $d_{i,j}^{\max}$ is the maximum tolerant delay of $f_{i,j}$ (in frames). Thus, once the eNB transmits at least $g_{i,j}(k)$ amount of data for a GBR flow during every frame k, we can guarantee that the flow's packets will never be dropped due to expiration. On the other hand, $\hat{c}_{i,j}(m)$ is a coefficient that satisfies two conditions [21]:

$$\begin{aligned} 0 \leq & \hat{c}_{i,j}(m) \leq 1 \quad \forall m \in \mathbb{Z}_0^+, \\ & \hat{c}_{i,j}(m) \geq \hat{c}_{i,j}(m+1), m \geq 1 \text{ with } \hat{c}_{i,j}(m) \in \mathbb{R}. \end{aligned} \tag{7}$$

The coefficient $\hat{c}_{i,j}(m)$ is a real number between zero and one, and it monotonically decreases as the parameter m increases (when $m \geq 1$). One possible way to satisfy the conditions in Eq. (7) is to set $\hat{c}_{i,j}(0) = 0$, $\hat{c}_{i,j}(1) = 1$, and $\hat{c}_{i,j}(m+1) = \hat{c}_{i,j}(m)/2$, for $m = 1, 2, \cdots, M_{i,j} - 1$. In other words, we have $\hat{c}_{i,j}(2) = 1/2$, $\hat{c}_{i,j}(3) = 1/4$, $\hat{c}_{i,j}(4) = 1/8$, and so on. This setting will be also used by our simulations in Section 6.

We give an example in Fig. 3 to demonstrate FLS, where $M_{i,j}$ = 10 frames. In the kth frame, an amount $\phi_{i,j}(k)$ = 1000 bits of data comes to $f_{i,j}$'s queue. Based on the above setting of coefficient $\hat{c}_{i,j}(m)$, we can spread the enqueued data over the kth, $(k+1)$th, $\cdots$, and $(k+9)$th frames to 500, 250, $\cdots$, and 1 bits, respectively. Then, supposing that $\phi_{i,j}(k+1)$ = 2000 bits and $\phi_{i,j}(k+2)$ = 0 bit (i.e., no data generated in the $(k+2)$th frame), we can calculate the amount of data spread over $M_{i,j}$ observing frames accordingly. Based on Eq. (6), we eventually

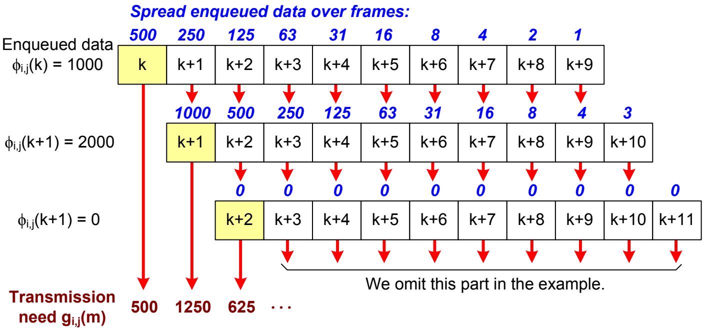

Spread enqueued data over frames:
500 250 125 63 31 16 8 4 2 1
Enqueued data
ϕi,j(k) = 1000
k k+1 k+2 k+3 k+4 k+5 k+6 k+7 k+8 k+9
1000 500 250 125 63 31 16 8 4 3
ϕi,j(k+1) = 2000
k+1 k+2 k+3 k+4 k+5 k+6 k+7 k+8 k+9 k+10
0 0 0 0 0 0 0 0 0 0
ϕi,j(k+1) = 0
k+2 k+3 k+4 k+5 k+6 k+7 k+8 k+9 k+10 k+11
We omit this part in the example.
Transmission need gi,j(m)
500 1250 625 ···

Fig. 3: Calculate the transmission need of a GBR flow by FLS.

obtain $g_{i,j}(k)$ = 500 bits, $g_{i,j}(k+1)$ = 1250 bits, $g_{i,j}(k+2)$ = 625 bits, and so on, as shown in Fig. 3.

4.2.2 PRB Allocation

From Fig. 2, the 1st-layer scheduler passes the transmission need $g_{i,j}(k)$ of each GBR flow to the 2nd-layer scheduler, and then the 2nd-layer scheduler determines PRB allocation for both GBR and non-GBR flows. Specifically, it obeys two priority rules to allocate PRBs:

- **[Priority rule 1]** GBR flows are given precedence over non-GBR ones, as they have stringent delay constraints.
- **[Priority rule 2]** Golden UEs can acquire network resource first, followed by silver and bronze UEs.

Consequently, the eNB first allocates PRBs to GBR flows to meet rule 1. Since there may not be sufficient resource to serve all GBR flows, we thus modify M-LWDF to select a GBR flow to receive each PRB. In order to apply rule 2, we define a *price-based weight* for each UE u_i by

$$W_i^P = \frac{C_{l_i}}{C_{\mathbf{G}} + C_{\mathbf{S}} + C_{\mathbf{B}}}, \tag{8}$$

where $C_{\mathbf{G}}$, $C_{\mathbf{S}}$, and $C_{\mathbf{B}}$ denote the average amount of money charged to a golden, silver, and bronze user, respectively, and $C_{l_i} \in \{C_{\mathbf{G}}, C_{\mathbf{S}}, C_{\mathbf{B}}\}$. Such information can be obtained from the pricing module (discussed in Section 4.3). Then, we select a flow $f_{i,j}$ to obtain the PRB as follows:

$$f_{i,j} = \arg\max_{i,j} \left[\left(w_{i,j} d_{i,j} \times \frac{r_i}{r_i^{\text{avg}}} \right) \times W_i^P \right], \tag{9}$$

where $w_{i,j}$ is the original weight defined by M-LWDF. Specifically, $w_{i,j} = -\log \beta_{i,j}/\sigma_{i,j}$, where $\beta_{i,j}$ is the maximum probability of packet dropping (i.e., $d_{i,j} > d_{i,j}^{\max}$), and $\sigma_{i,j}$ denotes the expected delay of $f_{i,j}$. In Eq. (8), the price-based weight W_i^P is limited to (0, 1], and a higher-level UE will have a larger W_i^P value[2]. Thus, there is a higher possibility to pick its GBR flow to receive the PRB by Eq. (9). The eNB then iteratively uses Eq. (9) to allocate each PRB, until either 1) all PRBs have been consumed or 2) the transmission need $g_{i,j}(k)$ of every GBR flow is satisfied in the current TTI.

Afterward, if there still remain PRBs, the eNB distributes them among non-GBR flows. We enhance max-CQI by introducing the price-based weight, so as to increase the overall throughput of non-GBR flows while come up to rule 2:

$$u_i = \arg\max_{i} \left(r_i \times W_i^P \right). \tag{10}$$

[2] The case of $W_i^P = 1$ occurs when all UEs have the same user level.

TABLE 3: CQI table defined in LTE.

CQI value	MCS level	code rate	efficiency (×1024)	bits carried by a PRB[3]
1	QPSK[1]	78	0.1523	12.79
2	QPSK	120	0.2344	19.69
3	QPSK	193	0.3770	31.67
4	QPSK	308	0.6016	50.53
5	QPSK	449	0.8770	73.67
6	QPSK	602	1.1758	98.77
7	16QAM[2]	378	1.4766	124.03
8	16QAM	490	1.9141	160.78
9	16QAM	616	2.4063	202.13
10	64QAM	466	2.7305	229.36
11	64QAM	567	3.3223	279.07
12	64QAM	666	3.9023	327.79
13	64QAM	772	4.5234	379.97
14	64QAM	873	5.1152	429.68
15	64QAM	948	5.5547	466.59

[1] QPSK: Quadrature phase shift keying
[2] QAM: Quadrature amplitude modulation
[3] This is an average value.

The eNB then iteratively uses Eq. (10) to allocate the remaining PRBs, until either 1) all PRBs have been allocated or 2) the traffic demand of each non-GBR flow is satisfied.

Next, we give two remarks for our PRB allocation in the 2nd-layer scheduler. Remark 1 discusses how to determine the current rate r_i of each UE, which is used by both Eqs. (9) and (10). Then, Remark 2 addresses how to estimate the number of PRBs actually required by a flow.

Remark 1 (*Determining the current rate r_i of a UE*)**.** Both max-CQI and M-LWDF methods are developed for general wireless networks [4], so the current rate r_i of each UE can be easily determined if there is only one downlink channel. However, LTE adopts OFDMA for downlink communication, where PRBs may locate in different subchannels that encounter frequency selective fading [22]. In other words, r_i may not be necessarily the same across all PRBs. Therefore, to make both Eqs. (9) and (10) function well in the LTE environment, the rate r_i can be defined by the data rate supported by the current PRB for a UE u_i. In fact, the LTE standard [23] defines a CQI table shown in Table 3 to determine the relationship between efficiency and MCS (including the code rate) for each CQI value. Through the CQI table, we can calculate the average number of data bits carried by each PRB when a UE has a certain CQI value for that PRB. In this way, we can also determine its current rate r_i (for the PRB) accordingly. □

Remark 2 (*Calculating the number of PRBs used by a flow*)**.** As mentioned in Remark 1, a flow may have different channel quality across its allocated PRBs. To find the effective SINR (signal-to-interference-plus-noise ratio) γ_{eff} on these PRBs, we can employ the *exponential effective SINR mapping (EESM)* approach [24] as follows:

$$\gamma_{\text{eff}} = \text{EESM}(\Gamma, \varepsilon) = -\varepsilon \ln \frac{1}{s} \sum_{k=1}^{s} e^{-\gamma_k/\varepsilon}, \tag{11}$$

where Γ is a vector $[\gamma_1, \gamma_2, \cdots, \gamma_s]$ of the tone SINR value for each subchannel, s is the number of subchannels, and ε is a tunable parameter (usually set to one). In the LTE implementation, the eNB first uses one PRB to calculate γ_{eff} by Eq. (11), and then checks if this PRB has sufficient capacity (by consulting Table 3) to satisfy the traffic demand of flow

$f_{i,j}$ in the current TTI. If not, the eNB iteratively adds the next PRB, recalculates γ_{eff}, and repeats the above check, until $f_{i,j}$'s demand becomes satisfied or there is no available PRB. In this way, we can estimate the number of PRBs actually used by each flow. □

4.2.3 Opportunistic Resource Reallocation

By introducing the price-based weight W_i^P to Eqs. (9) and (10), we allow golden UEs to acquire PRBs first (and followed by silver and bronze UEs). However, such PRB allocation may hurt system performance, especially when high-level UEs encounter bad channel condition. In this case, their PRBs can only use simple MCS and carry quite few data bits (in other words, the spectral resource is wasted). To deal with the problem, the 3rd-layer scheduler adopts opportunistic resource reallocation, whose idea is to allow a small portion of PRBs to be 'reallocated' to low-level UEs according to their channel rates.

Let Ψ_{GBR} and Ψ_{NGBR} be the sets of PRBs allocated to GBR and non-GBR flows by the 2nd-layer scheduler, respectively. We have $\Psi_{GBR} \cap \Psi_{NGBR} = \emptyset$ and $\Psi_{GBR} \cup \Psi_{NGBR} \subseteq \Psi$, where Ψ denotes the set of available PRBs in the current TTI. Also, we define two ratios ξ_{GBR} and ξ_{NGBR} to respectively control the portion of PRBs to be reallocated in Ψ_{GBR} and Ψ_{NGBR}. Below, we separate our discussion into GBR and non-GBR cases.

Let us denote by $p_{i,j}$ the PRB allocated to a flow $f_{i,j}$ of UE u_i. For the GBR case, we sort all $p_{i,j}$ in Ψ_{GBR} by its UE's channel rate r_i in an increasing order. Then, we create a subset $\Psi^R_{GBR} \subseteq \Psi_{GBR}$ of candidate PRBs that can be reallocated. Initially, we have $\Psi^R_{GBR} = \emptyset$. Then, for each $p_{i,j} \in \Psi_{GBR}$, we add it to Ψ^R_{GBR} if the user level $l_i \in \{\mathbf{G}, \mathbf{S}\}$ (i.e., golden or silver UEs), until $|\Psi^R_{GBR}|$ reaches to $\lceil |\Psi_{GBR}| \times \xi_{GBR} \rceil$, where '$|\cdot|$' denotes the number of elements in a set and '$\lceil \cdot \rceil$' is the ceiling function. Then, we consider two reallocation rules:

- **[Reallocation rule 1]** $l_i = \mathbf{G}$:
  If the following condition satisfies
  $$r_i < \max\{r_{i'} \mid l_{i'} \in \{\mathbf{S}, \mathbf{B}\} \text{ and } f_{i',j'} \text{ is GBR}\}, \quad (12)$$
  which means that another GBR flow $f_{i',j'}$ owned by a lower-level UE actually has a higher channel rate to PRB $p_{i,j}$, we thus reallocate $p_{i,j}$ to $f_{i',j'}$ to improve system performance. If Eq. (12) is violated, we remove $p_{i,j}$ from Ψ^R_{GBR} because there is no gain to reallocate the PRB.
- **[Reallocation rule 2]** $l_i = \mathbf{S}$:
  If the following condition satisfies
  $$r_i < \max\{r_{i'} \mid l_{i'} = \mathbf{B} \text{ and } f_{i',j'} \text{ is GBR}\}, \quad (13)$$
  which implies that GBR flow $f_{i',j'}$ owned by a bronze UE has a higher channel rate to $p_{i,j}$ than its original flow $f_{i,j}$, we reallocate $p_{i,j}$ to $f_{i',j'}$ to increase network throughput. When Eq. (13) is violated, we remove $p_{i,j}$ from Ψ^R_{GBR} as there is no need to reallocate the PRB.

After the above examination, Ψ^R_{GBR} will remain only the PRBs that have been reallocated to other flows. The scheduling module then passes Ψ^R_{GBR} to the pricing module for extra charge (discussed in Section 4.3).

We deal with the non-GBR case following the above two reallocation rules. Then, the set Ψ^R_{NGBR} is also passed to the pricing module to calculate the extra charge for those low-level users who get additional resource by opportunistic resource reallocation. Remark 3 discusses the effect of both parameters ξ_{GBR} and ξ_{NGBR} on PRB allocation.

Remark 3 (*Impact of ξ_{GBR} and ξ_{NGBR}*)**.** Both ξ_{GBR} and ξ_{NGBR} are the ratios of GBR and non-GBR PRBs that will be considered to be reallocated, respectively. Based on the two reallocation rules in Eqs. (12) and (13), a PRB will be reallocated if we can find a lower-level UE that has better channel condition (i.e., larger CQI value) to that PRB. In other words, when both ξ_{GBR} and ξ_{NGBR} increase, the overall throughput could improve as more PRBs can be given to the UEs with the best channel condition. Nevertheless, higher-level UEs may be forced to give up more PRBs. Let us consider an extreme case where $\xi_{GBR} = \xi_{NGBR} = 1$ and bronze UEs have better channel condition than others. In this case, the 3rd-layer scheduler reallocates each PRB to the bronze UE that has the largest CQI value. Consequently, the result of PRB allocation will be the same with that of the max-CQI method. However, both golden and silver UEs will receive no PRB, which violates the principle of UE classification.

The design of opportunistic resource reallocation is to improve network throughput under the prerequisite that UEs are given with resource based on their priorities (i.e., levels). Obviously, the values of ξ_{GBR} and ξ_{NGBR} should not be set too large. Therefore, we suggest setting $\xi_{GBR} \leq 0.1$ and $\xi_{NGBR} \leq 0.1$ so that no more than 10% of PRBs will be reallocated. In this way, the price-based weight W_i^P can have dominating effect on PRB allocation. We will also set both ξ_{GBR} and ξ_{NGBR} to 0.1 in our simulations. □

4.3 Pricing Module

The pricing module refers to the user level and the PRB allocation from the scheduling module to charge each user, as shown in Fig. 2. It also consults PED to model the reaction of user demand to the change of price. Here, we adopt the PED-related equation in [25], which is used to analyze the effect of price $\tilde{P}$ on user demand $\tilde{D}$ in wireless networks:

$$\tilde{D} = \lambda \tilde{P}^{-E_d}, \quad (14)$$

where λ is a scaling constant to represent the demand potential[3], and E_d is the coefficient of price elasticity. From Eq. (14), we can derive E_d by

$$\tilde{D}_2/\tilde{D}_1 = \left(\tilde{P}_1/\tilde{P}_2\right)^{E_d} \;\Rightarrow\; E_d = \frac{\ln(\tilde{D}_2/\tilde{D}_1)}{\ln(\tilde{P}_1/\tilde{P}_2)}. \quad (15)$$

In general, a larger E_d value implies that the user demand is relatively elastic. In other words, when the price increases, the user demand will decrease more significantly, and vice versa. According to [26], [27], VoIP and video applications have dominated the revenue of most telecommunications operators. It implies that VoIP and video flows should have smaller E_d values, as people usually use these applications. Therefore, we set E_d to 1.3, 1.7, and 2.1 for VoIP, video, and non-GBR flows, respectively, based on the suggestion in [25].

To compute the amount of money charged to a flow $f_{i,j}$ based on its consumption of network resource, we improve Eq. (2) as follows:

$$C_{i,j} = [\mathcal{P}_v(l_i) + \mathcal{P}_e] \times (\hat{e} - \hat{e}^{-y}) \times \mathcal{L}, \quad (16)$$

[3] λ can be set to equal to the value of $\tilde{D}$ when $\tilde{P} = 1$.

where the variable charge is defined by

$$\mathcal{P}_v(l_i) = \begin{cases} \mathcal{P}_c & \text{if } \mathcal{L} \leq \delta, \\ \mathcal{P}_v(\mathbf{G}) & \text{if } \mathcal{L} > \delta \text{ and } l_i = \mathbf{G}, \\ \mathcal{P}_v(\mathbf{S}) & \text{if } \mathcal{L} > \delta \text{ and } l_i = \mathbf{S}, \\ \mathcal{P}_v(\mathbf{B}) & \text{if } \mathcal{L} > \delta \text{ and } l_i = \mathbf{B}, \end{cases} \tag{17}$$

y is flow $f_{i,j}$'s QCI[4], and the network load is defined by

$$\mathcal{L} = \frac{\text{the number of used PRBs}}{\text{the number of available PRBs}}. \tag{18}$$

In Eq. (16), $\mathcal{P}_e$ is the extra charge incurred when $f_{i,j}$ uses PRBs in Ψ^R_{GBR} or Ψ^R_{NGBR}. With the PED model, we define

$$\mathcal{P}_e = \gamma - E_d, \tag{19}$$

where $\gamma \geq \max\{E_d, \forall f_{i,j}\}$. For example, we can set $\gamma = 2.5$, so $\mathcal{P}_e$ will be 1.2, 0.8, and 0.4 for VoIP, video, and non-GBR flows. Such setting is feasible due to two reasons. First, VoIP service has the lowest price elasticity, so we can charge more money to increase operator revenue without significantly degrading user demands. Second, since non-GBR service has the highest price elasticity, we can reduce its extra charge to encourage users to utilize such service. Notice that if flow $f_{i,j}$ does not use any PRB in $\Psi^R_{GBR} \cup \Psi^R_{NGBR}$, its user need not pay for such extra charge. Then, the amount of money charged to a user u_i will be the sum of charges to all its flows:

$$C_i = \sum_{\forall f_{i,j} \in u_i} C_{i,j}. \tag{20}$$

In addition, the average amount of money charged to a golden user can be derived by

$$C_{\mathbf{G}} = \begin{cases} \frac{\sum_{u_i}\{C_i \mid l_i = \mathbf{G}\}}{N_{\mathbf{G}}} & \text{if } N_{\mathbf{G}} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

where $N_{\mathbf{G}}$ is the number of golden users. Similarly, we can compute $C_{\mathbf{S}}$ and $C_{\mathbf{B}}$ (i.e., the average amount of money charged to a silver and bronze user, respectively) following Eq. (21). As illustrated in Fig. 2, the parameters $C_{\mathbf{G}}$, $C_{\mathbf{S}}$, and $C_{\mathbf{B}}$ will change depending on the number of PRBs allocated to different levels of UEs (determined by the scheduling module), and they are necessary to calculate the price-based weight W_i^P in Eq. (8) used by the 2nd-layer scheduler. This relationship exhibits that both scheduling and pricing modules can tightly couple with each other.

4.4 Design Rationale

Most studies discussed in Section 3 independently cope with the resource scheduling and pricing problems in LTE. They aim at either improving system performance or increasing operator revenue. However, the two objectives may conflict with each other if we do not take both of them into consideration. Specifically, when we simply allocate most resource to the flows with better channel quality to improve performance, the priority of high-level users would be omitted by the scheduler. On the contrary, if we want to increase revenue by giving most resource to high-level users, performance may degrade when their UEs encounter worse channel condition.

To conquer this dilemma, our PARS framework proposes two tightly-coupled modules to handle PRB allocation and user charge, as shown in Fig. 2. The scheduling module relies on the information of user level and charge from the pricing module to calculate the price-based weight W_i^P, which plays a critical role in assigning PRBs for transmission. On the other hand, the pricing module estimates user charge based on PRB allocation and two sets Ψ^R_{GBR} and Ψ^R_{NGBR} outputted from the scheduling module. In this way, the PARS framework can balance between system performance and operator revenue.

In the scheduling module, there are three special designs to support QoS for GRB flows:

- We employ FLS in the 1st-layer scheduler to estimate the amount of data transmission required to meet the delay constraint of each GBR flow. Thus, the eNB can try its best to satisfy the transmission needs by the 2nd-layer scheduler. Notice that we modify FLS to compute only the transmission amount used to avoid a GBR flow dropping its packets. When some users request a large amount of GBR traffic, non-GBR users can still have network resource to use their service.
- For the priority rules in Section 4.2.2, we give rule 1 precedence over rule 2 to guarantee GBR transmissions of low-level users. Let us consider an example where golden users send a huge amount of non-GBR data. If we simply satisfy their demands first, silver and bronze users will never have a chance to receive network resource (i.e., starvation), even though they have delay-critical GBR traffic. To solve this problem, the eNB first satisfies the 'necessary' GBR demands (i.e., rule 1), and then gives the remaining PRBs to non-GBR flows. When dealing with the GBR and non-GBR cases, rule 2 can ensure that golden users have a higher priority to receive resource.
- GBR and non-GBR flows are scheduled in different ways. We use the enhanced M-LWDF method to schedule GBR flows, which takes care of the urgent degree (i.e., packet delay) of each flow. Non-GBR flows are scheduled by the enhanced max-CQI method for performance concern.

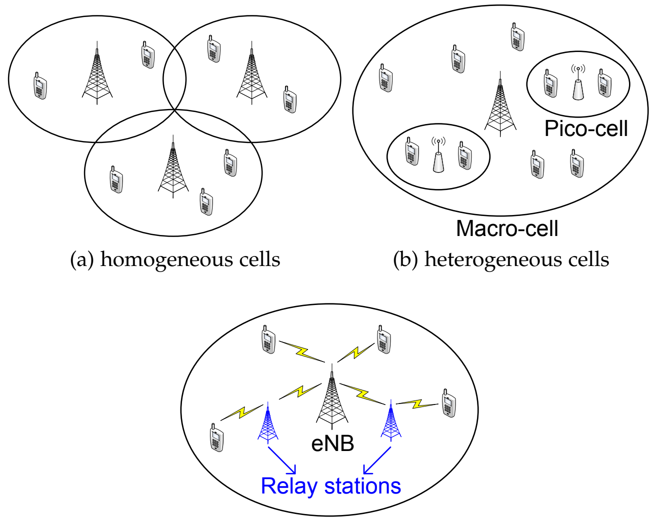
Furthermore, we propose the mechanism of opportunistic resource reallocation to adjust the scheduling result by replacing some 'bad' PRB allocation. In particular, we pick those PRBs assigned to high-level UEs that encounter worse channel condition, and check if each of such PRBs can be reallocated to another UE with a lower user level but better channel quality. In this case, these low-level users have to pay for extra fee (i.e., $\mathcal{P}_e$ in Eq. (16)) to receive the reallocating PRBs by the pricing module. Moreover, our pricing module uses Eq. (19) to adjust the extra fee according to different service types, which considers the price elasticity defined by the economic PED model. It thus helps increase operator revenue without significantly degrading user demands.

4.5 Extending to the Multi-cell Environment

Till now, the discussion of our scheduling module aims at a single-cell environment. However, it can be easily extended to a multi-cell environment. Below, we consider three types of LTE multi-cell networks, as shown in Fig. 4:

- *Homogeneous cells:* This is the simplest case. Each eNB independently manages the spectral resource in its cell without affecting other eNBs. Therefore, we can directly apply our scheduling module to each individual eNB.

[4] There were originally nine QCIs defined in LTE Release-8 standard [28]. However, LTE Release-13 standard [7] adds four new QCIs, 65, 66, 69, and 70, for some special applications such as mission critical data. We thus restrict y to range between 1 and 9 in Eq. (16) for backward compatibility.



(a) homogeneous cells (b) heterogeneous cells

(c) cells with relay stations

Fig. 4: Three types of LTE multi-cell networks.

- *Heterogeneous cells:* A large macro-cell may contain several small pico-cells. In this case, they may cause signal interference with each other. To conquer this problem, LTE adopts the technique of *enhanced inter-cell interference coordination (eICIC)* [29]. In eICIC, the macro-cell eNB will select some slots (called *almost blank subframe*, abbreviated to 'ABS') to transmit only low-power signals. Thus, pico-cell eNBs can send their data without interference in ABS slots. To apply our scheduling module to such networks, the macro-cell eNB will allocate PRBs only in non-ABS slots. On the other hand, pico-cell eNBs can allocate PRBs in ABS slots.
- *Cells with relay stations:* Each UE can choose to receive data directly from the eNB or via a relay station. In this network, the eNB and all relay stations share the same spectral resource. Therefore, the eNB will partition PRBs into three groups: 1) PRBs for the data from the eNB directly to UEs, 2) PRBs for the data from the eNB to relay stations, and 3) PRBs for the data from relay stations to UEs. How to calculate the number of PRBs in each group can refer to our previous work [30]. After the partition, both the eNB and relay stations can separately use our scheduling module to allocate PRBs in the corresponding groups.

On the other hand, since we install the pricing module in PCRF (referring to the LTE structure in Fig. 1), there is no need to modify the module when we switch from the single-cell environment to a multi-cell network. The major reason is that each piece of data will be associated with the destination eNB, so it is easy to allow PCRF to calculate the amount of data received by each UE in the network.

5 Theoretical Analysis

In this section, we give analysis on performance and complexity of the PARS framework. For performance analysis, we aim at whether the scheduling module can guarantee delay bound of GBR flows (i.e., QoS support). In particular, our scheduling module uses FLS to calculate the GBR transmission need in the 1st layer. The objective of FLS is to support *bounded-input, bounded-output (BIBO)* stability [21], where the output of a system remains bounded in amplitude, provided that the input is also bounded. In other words, the eNB will never seek to allocate an infinite bandwidth due to the reason that the system input (i.e., the incoming data rate) is bounded in amplitude as any practical application cannot produce an infinite packet rate. Lemma 1 shows that FLS satisfies BIBO stability and also indicates its delay bound for GBR flows.

Lemma 1. *FLS is BIBO stable and ensures that the queuing delay of any GBR flow $f_{i,j}$ is smaller than $M_{i,j}+1$, where $M_{i,j}$ is the flow's sampling interval.*

Proof: The proof can be found in Theorem 1 of [9]. □

With Lemma 1, the following theorem then proves that the scheduling module in PARS can meet the delay requirement of GBR flows when the eNB has sufficient downlink resource.

Theorem 1. *Given R_k PRBs supported by an eNB in the kth frame, where the minimum number of data bits carried by each PRB is kept above $b_{\min}$, then the scheduling module in PARS can guarantee that there is no packet dropping of GBR flows due to expiration if the following equation holds for any k:*

$$R_k \times b_{\min} \geq \sum\nolimits_{\forall f_{i,j}} g_{i,j}(k). \tag{22}$$

Proof: The scheduling module uses FLS in its 1st-layer scheduler and sets the sampling interval $M_{i,j}$ of each GBR flow $f_{i,j}$ to $d^{\max}_{i,j}-1$, where $d^{\max}_{i,j}$ is the maximum tolerant delay. According to Lemma 1, if the eNB can allocate enough resource to meet the transmission demand $g_{i,j}(k)$ of $f_{i,j}$ by Eq. (6), each of $f_{i,j}$'s packet must be delivered before the deadline $d^{\max}_{i,j}$. In fact, Eq. (22) indicates the condition of whether the eNB has sufficient PRBs, where the left term is the total number of data bits that can be sent out in the kth frame, while the right term is the amount of overall GBR transmission need that should be satisfied in the frame. Based on priority rule 1 in Section 4.2.2, GBR flows can always obtain PRBs for transmission first in the 2nd-layer scheduler. This rule implies that the eNB must allocate enough PRBs for each GBR flow to avoid dropping its packets, no matter there exist non-GBR flows. Then, the 3rd-layer scheduler deals with GBR and non-GBR PRB reallocation independently. Thus, there is no possibility that some GBR PRBs will be reallocated to non-GBR flows. Therefore, the scheduling module can meet the delay requirement of every GBR flow if Eq. (22) holds for any k, thereby proving this theorem. □

On the other hand, our pricing module enhances the NLP method discussed in Section 3.2, which employs the linearity factor to estimate the amount of money charged to users:

$$f_L(x) = A \times (\hat{e} - \hat{e}^{-Bx}), \tag{23}$$

where A decides the base level of price while B adjusts the deduction of price when using high bit rate or volume transfers. It has been shown in [19] that the linearity factor of NLP charging increases operator revenue while considering price elasticity. Moreover, Theorem 2 shows that our pricing module can further improve revenue than the NLP method.

Theorem 2. *With the same result of PRB allocation, the pricing module in PARS can receive revenue no less than NLP.*

Proof: By comparing the pricing equations of NLP and the pricing module in Eq. (2) and Eq. (16), it is apparent that the pricing module will ask users for extra charge $\mathcal{P}_e$. Such charge occurs only when a user receives PRBs from the opportunistic resource reallocation method. In other words, if there is no PRB reallocation, our pricing module will compute the same amount of money with NLP. Otherwise, it can further improve revenue comparing with NLP due to extra charge $\mathcal{P}_e$. □

We then analyze the computational complexity. Lemma 2 discusses the complexity to run the scheduling module on each eNB, while Lemma 3 shows the complexity to conduct the pricing module by PCRF. Theorem 3 finally gives the overall complexity of our PARS framework.

Lemma 2. *Let N_l^{GBR} and N_l^{NGBR} be the number of GBR and non-GBR flows in the lth cell, respectively. Then, the worst-case complexity of the scheduling module for the lth cell is $O(N_l^{\mathrm{GBR}}(D_l^{\max}-1))+O(R_l\cdot\max\{N_l^{\mathrm{GBR}},N_l^{\mathrm{NGBR}}\})$, where $D_l^{\max}$ and R_l are the maximum tolerant delay of flows (in frames) and the number of PRBs (in a TTI) in the cell.*

Proof: The scheduling module conducts the three-layer schedulers in sequence (referring to Fig. 2), so we analyze the computational complexity of each scheduler separately. Specifically, the 1st-layer scheduler adopts FLS to compute the GBR transmission need, which relies on Eq. (6) to do the computation. Obviously, the computation of $g_{i,j}(k)$ for a flow $f_{i,j}$ requires $(M_{i,j}-1)$ multiplications and $(3(M_{i,j}-1)+1)$ sums, so the complexity of Eq. (6) is $O(M_{i,j})$. Because there are N_l^{GBR} GBR flows, so we have to repeat Eq. (6) for N_l^{GBR} times, which spends time of $O(N_l^{\mathrm{GBR}}\cdot M_{\max})$, where $M_{\max}=\max\{M_{i,j}\}$. As mentioned in Section 4.2.1, we set $M_{i,j}=d_{i,j}^{\max}-1$, so $M_{\max}=D_l^{\max}-1$. Thus, the complexity of the 1st-layer scheduler will be $O(N_l^{\mathrm{GBR}}(D_l^{\max}-1))$.

The 2nd-layer scheduler first uses the enhanced M-LWDF scheme to allocate PRBs to GBR flows. From Eq. (9), it takes $O(N_l^{\mathrm{GBR}})$ time to allocate each PRB because we have to check every GBR flow. If there still remain PRBs, we use the enhance max-CQI scheme to distribute PRBs among non-GBR flows. Based on Eq. (10), it requires $O(N_l^{\mathrm{NGBR}})$ time to assign a PRB since we should examine every non-GBR flow. When $N_l^{\mathrm{GBR}}\geq N_l^{\mathrm{NGBR}}$, the worst case occurs if all PRBs are allocated to GBR flows. If $N_l^{\mathrm{GBR}}<N_l^{\mathrm{NGBR}}$, the worst case occurs when the resource is given to only non-GBR flows. Thus, the complexity of the 2nd-layer scheduler will be $O(R_l\cdot\max\{N_l^{\mathrm{GBR}},N_l^{\mathrm{NGBR}}\})$.

In the 3rd-layer scheduler, two cases are considered. For the GBR case, the eNB reallocates at most $\xi_{GBR}\cdot R_l$ PRBs to silver and bronze UEs by Eqs. (12) and (13), so the complexity is $O(\xi_{GBR}\cdot R_l\cdot(N_l^{\mathrm{GBR,S}}+N_l^{\mathrm{GBR,B}}))$, where $N_l^{\mathrm{GBR,S}}$ and $N_l^{\mathrm{GBR,B}}$ respectively denote the number of GBR flows owned by silver and bronze UEs. Similarly, the non-GBR case will spend time of $O(\xi_{NGBR}\cdot R_l\cdot(N_l^{\mathrm{NGBR,S}}+N_l^{\mathrm{NGBR,B}}))$, where $N_l^{\mathrm{NGBR,S}}$ and $N_l^{\mathrm{NGBR,B}}$ are the number of non-GBR flows owned by silver and bronze UEs, respectively.

Thus, the total complexity is $O(N_l^{\mathrm{GBR}}(D_l^{\max}-1))+O(R_l\cdot\max\{N_l^{\mathrm{GBR}},N_l^{\mathrm{NGBR}}\})+O(\xi_{GBR}\cdot R_l\cdot(N_l^{\mathrm{GBR,S}}+N_l^{\mathrm{GBR,B}}))+O(\xi_{NGBR}\cdot R_l\cdot(N_l^{\mathrm{NGBR,S}}+N_l^{\mathrm{NGBR,B}}))$. Because $\xi_{GBR}\leq 1$, $\xi_{NGBR}\leq 1$, $N_l^{\mathrm{GBR,S}}+N_l^{\mathrm{GBR,B}}\leq N_l^{\mathrm{GBR}}$, and $N_l^{\mathrm{NGBR,S}}+N_l^{\mathrm{NGBR,B}}\leq N_l^{\mathrm{NGBR}}$, we can simplify the complexity to $O(N_l^{\mathrm{GBR}}(D_l^{\max}-1))+O(R_l\cdot\max\{N_l^{\mathrm{GBR}},N_l^{\mathrm{NGBR}}\})$, thereby proving the lemma. □

Lemma 3. *Suppose that N is the total number of flows in an LTE network. Then, the computational complexity of the pricing module is $O(N)$ in the worst case.*

TABLE 4: Simulation parameters.

eNB-related parameters:	
bandwidth	20MHz
number of PRBs	100 (12 subcarriers per PRB)
cell range	1500 meters
frame structure	frequency division duplexing (FDD)
MCS	QPSK, 16QAM, 64QAM
UE-related parameters:	
number of UEs	36, 48, 60, 72, 84, 96, 108
mobility model	random direction
moving speed	3km/h
GBR flows	VoIP (8.4kbps) and H.264 video (242kbps)
non-GBR flows	CBR (400kbps)
channel-related parameters:	
propagation loss	urban macro-cell model
path loss	$128.1+37.6\log L$, where L is the distance between the eNB and a UE in kilometers
shadowing fading	log-normal distribution with 0dB mean and 8dB standard deviation
penetration loss	10dB
fast fading	Jakes model (for Rayleigh fading)
pricing-related parameters: (price unit: mu/PRB)	
flat-rate method	$\mathcal{P}_c=30$
fixed-PRB method	$\mathcal{P}_f(G)=11,\mathcal{P}_f(S)=6,\mathcal{P}_f(B)=4$
NLP method	$\alpha=1,\mathcal{P}_c=2.6$ $\mathcal{P}_v(G)=0.9,\mathcal{P}_v(S)=0.7,\mathcal{P}_v(B)=0.5$
SCP method	$\kappa=520$ $\mathcal{P}_f(G)=9,\mathcal{P}_f(S)=8,\mathcal{P}_f(B)=4$
PARS framework	$\gamma=2.5,\mathcal{P}_c=2.6$ $\mathcal{P}_v(G)=0.9,\mathcal{P}_v(S)=0.7,\mathcal{P}_v(B)=0.5$

Proof: The pricing module uses Eq. (16) to compute the fee of each flow, where the variable charge $\mathcal{P}_v(l_i)$, extra charge $\mathcal{P}_e$, and QCI index y can be determined by the flow itself. Moreover, the network load can be calculated once by Eq. (18). Thus, it takes $O(N)$ time to compute the charges for all flows in the network by Eq. (16). Then, calculating the fee to each UE requires $O(N)$ time in Eq. (20), because each flow belongs to only one UE. Due to the same reason, it also spends $O(N)$ time to find the values of $C_{\mathbf{G}}$, $C_{\mathbf{S}}$, and $C_{\mathbf{B}}$ by Eq. (21). Therefore, the worst-case complexity of the pricing module will be $O(N)+O(N)+O(N)=O(N)$. □

Theorem 3. *Given N^{GBR} GBR flows and N^{NGBR} non-GBR flows in an LTE network, the computational complexity of the PARS framework is $O(N^{GBR}(D-1))+O(R\cdot\max\{N^{GBR},N^{NGBR}\})+O(N^{GBR}+N^{NGBR})$, where D is the maximum tolerant delay of all flows (in frames) and R is the maximum number of PRBs supported by a cell (in a TTI) in the worst case.*

Proof: Based on the discussion in Section 4.5, each eNB conducts the scheduling module independently. Let $\mathcal{L}$ denotes the set of all cells in the LTE network. By Lemma 2, the overall complexity to conduct the scheduling module will be

$$\sum_{l\in\mathcal{L}}O(N_l^{\mathrm{GBR}}(D_l^{\max}-1))+O(R_l\cdot\max\{N_l^{\mathrm{GBR}},N_l^{\mathrm{NGBR}}\})$$
$$=O(N^{\mathrm{GBR}}(D-1))+O(R\cdot\max\{N^{\mathrm{GBR}},N^{\mathrm{NGBR}}\}).\quad(24)$$

By combining Eq. (24) with Lemma 3, where N is replaced by $(N^{GBR}+N^{NGBR})$, we can thus prove the theorem. □

6 Performance Evaluation

We use LTE-Sim [31] to verify the efficiency of the PARS framework. Table 4 lists the parameters of our simulations. We consider an LTE macro-cell where the eNB distributes 100 PRBs among UEs in every TTI. Each UE follows the random direction model [32] to move in the cell with a velocity of 3km/h to imitate human walking. We also vary the number of UEs to evaluate the effect of different network loads, where the number of golden, silver, and bronze UEs are equal.

Each UE has two GBR flows: 8.4kbps VoIP flow (QCI = 1) and 242kbps H.264 video flow (QCI = 2). It also has a non-GBR flow: 400kbps constant-bit-rate (CBR) flow (QCI = 6). In each single experiment, the number of golden/silver/bronze UEs and their flows do not change. However, we apply the PED model to reflect the relationship between price and traffic demand. Specifically, we make a UE adjust the demand of its flows in each period (= 100 TTIs). Then, the UE will use Eq. (14) to recalculate the demand of its flows based on the average price that it has to pay in the previous period.

We compare our PARS framework with the flat-rate, fixed-PRB, NLP, and SCP methods discussed in Section 3.2. Since these four methods do not have resource scheduling mechanism, we use the two-level scheme in [9] to be their scheduling solution, which first adopts FLS to find the amount of GBR transmission need and then uses the PF solution to allocate PRBs. For NLP, SCP, and PARS, the load threshold δ is 50%. Also, we set both ξ_{GBR} and ξ_{NGBR} to 0.1 in PARS. For the PED model, the scaling constant λ in Eq. (14) is set to 2×10^5 according to [25]. The basic unit for price is called *monetary unit* (abbreviated to 'mu'). Remark 4 gives a discussion on our simulation setting.

Remark 4 (*Setting in simulations*)**.** LTE-Sim is a popular open-source simulator to model LTE behavior. It encompasses critical aspects of LTE networks, including E-UTRAN (evolved universal terrestrial radio access) and EPS (evolved packet system) [33]. Thus, LTE-Sim can provide sophisticated simulations for not only the wireless communication but also core network in LTE. Hence, we choose to use LTE-Sim for performance evaluation.

Moreover, to provide a general simulation environment, we set parameters in Table 4 as follows:

- The eNB-related parameters are determined based on the common setting of an LTE macro-cell eNB [34].
- We select GBR and non-GBR flows by consulting the LTE standard [23].
- The channel-related parameters are set according to the LTE specification in [35].
- All of the pricing methods (including PARS) use the FLS scheme for scheduling, so they can allocate PRBs based on the same amount of GBR transmission need.
- We set the pricing-related parameters of the flat-rate, fixed-PRB, NLP, and SCP methods by following their original simulation setting in [20], [36].

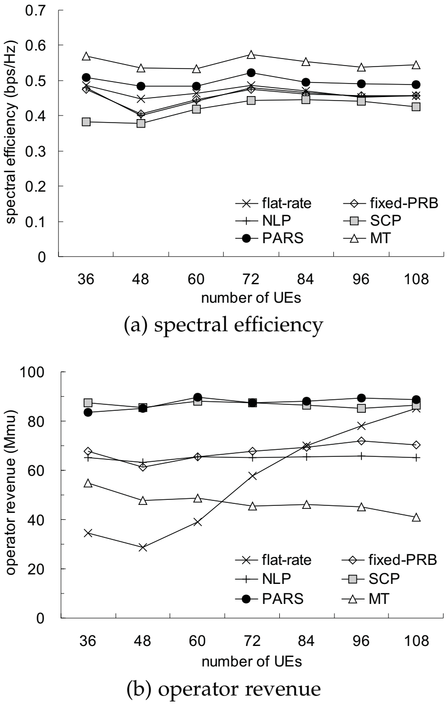
Based on the above parameter setting in our simulations, we can provide more fair comparison among different methods. □

6.1 System Performance versus Operator Revenue

We first measure both system performance and operator revenue by different methods. To evaluate system performance, we employ the concept of *spectral efficiency*:

$$\psi(t) = \frac{\sum_{\forall i} r_i^{\text{avg}}(t)}{\mathcal{B}_d}, \tag{25}$$

where $r_i^{\text{avg}}(t)$ is the average data rate of a UE u_i at the measuring time t (in particular, 100 seconds in our simulation), and $\mathcal{B}_d$ is the downlink channel bandwidth. Specifically, the spectral efficiency is the total information rate that can be sent over a given bandwidth in the LTE network, and its unit is 'bps/Hz'. Apparently, higher efficiency $\psi(t)$ indicates that the eNB can use the downlink channel more efficiently to transmit data, thereby achieving better system performance.



(a) spectral efficiency

(b) operator revenue

Fig. 5: Comparison on system performance and operator revenue.

To better illustrate the tradeoff between system performance and operator revenue, we develop a method, called *maximum throughput (MT)*, which has a bias in favor of the spectral efficiency. In particular, MT uses the max-CQI scheme for resource allocation, where each PRB is always given to the UE with the best channel condition. Besides, MT adopts the fixed-PRB method as its pricing mechanism, so it also classifies users into golden, silver, and bronze levels.

Fig. 5(a) presents the result of spectral efficiency. It is expected that the MT method will result in the highest efficiency $\psi(t)$, because it only picks the UE for each PRB that can achieve the highest data rate. On the other hand, users will be charged with much more money when the network load $\mathcal{L}$ exceeds the threshold δ in the SCP method, no matter what type of service is used. It will prevent low-level UEs from using more network resource. In this case, SCP inevitably incurs the lowest efficiency $\psi(t)$ among all methods. Except for MT, our PARS framework can have higher spectral efficiency than other methods due to three reasons. First, it flexibly uses the pricing-based M-LWDF and max-CQI methods to schedule resource for GBR and non-GBR flows, respectively. Second, PARS employs the opportunistic resource reallocation mechanism to adaptively adjust the usage of some PRBs to improve channel utilization. In particular, when high-level UEs encounter bad channel condition, their PRBs can be given to other UEs that have better channel quality. Third, by taking the price elasticity into consideration, PARS can encourage users to well utilize such additional resource. These three designs help significantly improve system performance of the PARS framework.

Fig. 5(b) shows the result of operator revenue, where the

TABLE 5: Improvement ratio by our PARS framework.

item	flat-rate	fixed-PRB	NLP	SCP	MT
performance	6.0%	8.6%	8.4%	15.5%	-10.8%
revenue	35.7%	22.6%	25.6%	0.9%	46.3%

unit is 10^6 mu (denoted by 'Mmu'). Although achieving the highest spectral efficiency, the revenue of MT is kept quite low, especially when there are more UEs. The major reason is that MT does not allocate PRBs based on user level. When high-level UEs has 'slightly' bad channel condition, they may be starved as most resource will be given to bronze UEs whose users pay less money. The flat-rate method employs a fixed charging policy, so the only way to increase the revenue is to increase the number of users. However, it would also congest the network. Both SCP and PARS result in the highest revenue, since they will charge users for extra money. However, SCP 'punishes' low-level users for using PRBs when $\mathcal{L} > \delta$ according to Eq. (3). It will discourage bronze UEs to well utilize network resource even when golden and silver UEs do not have much traffic demand. On the contrary, our PARS framework adaptively adjusts the extra charge $\mathcal{P}_e$ in Eq. (19) by considering the PED model, and such charge occurs only when low-level UEs use additional PRBs originally assigned to high-level UEs. That is why PARS can have higher spectral efficiency than SCP in Fig. 5(a), even though the values of their revenue Fig. 5(b) are close to each other.

We summarize the improvement ratio by PARS in Table 5 and give our observations in this experiment as follows:

- The MT method optimizes system performance by greedily matching each PRB with the UE that achieves the highest data rate. However, without considering user level and pricing, it will greatly decrease operator revenue.
- The SCP method seeks to maximize operator revenue by asking low-level users to pay for more money when the network load becomes heavy. Nevertheless, it ignores the case where most of the network load is contributed by silver or bronze UEs, thereby hurting system performance.
- Our PARS framework considers both scheduling and pricing modules to balance between system performance and operator revenue. The scheduling module works on the basis of user level but allows exchanging the usage of some PRBs to improve performance. The pricing module takes the PED model into account and asks users for extra charge depending on their service types. Thus, PARS can keep high revenue (almost the same with that of SCP) at the expense of degrading a small amount of throughput (i.e., around 10% of performance loss than MT).

6.2 Network Throughput by Flows

We then evaluate the throughput of different flows, as shown in Fig. 6. Since all methods use FLS to estimate the amount of GBR transmission need and let GBR flows receive resource first, VoIP and video throughput increases as the number of UEs grows. However, video flows have quite larger demand (242kbps) than VoIP flows (8.4kbps), so video throughput starts decreasing when there are more than 72 UEs. On the other hand, because non-GBR flows can receive resource only after the eNB has met GBR demand, CBR throughput will significantly decrease when the number of UEs increases.

Since the PED model reflects the relationship between demand and price, the traffic demand of flows would dynamically interact with the behavior of different methods. From Fig. 6, we have some observations about such interaction:

- **VoIP flows:** VoIP throughput of each method is similar due to two reasons. First, VoIP has the smallest price elasticity $E_d = 1.3$ in Eq. (14), which means that users may not significantly reduce VoIP traffic even when the price is raised. Second, a VoIP flow has light traffic load, so it is easy to satisfy VoIP demand.
- **Video flows:** Comparing with VoIP, video flows have larger price elasticity ($E_d = 1.7$) and very heavy traffic load. For the SCP method, since it substantially raises the (extra) price when the network load becomes heavy, users (especially for low-level ones) are thus inclined to decrease their video demand. On the contrary, our PARS framework takes the PED model into account, so it does not burden users with much money on receiving video service. It therefore can improve video throughput.
- **CBR flows:** With the largest price elasticity ($E_d = 2.1$) and traffic load, CBR throughput decreases drastically when there are more UEs. Such effect is more obvious in SCP. On the other hand, PARS lets low-level UEs use the PRBs allocated to high-level UEs if they have better channel quality, so it could increase CBR throughput.

To sum up, our PARS framework can achieve higher video and CBR throughput compared with the flat-rate, fixed-PRB, NLP, and SCP methods in the experiment.

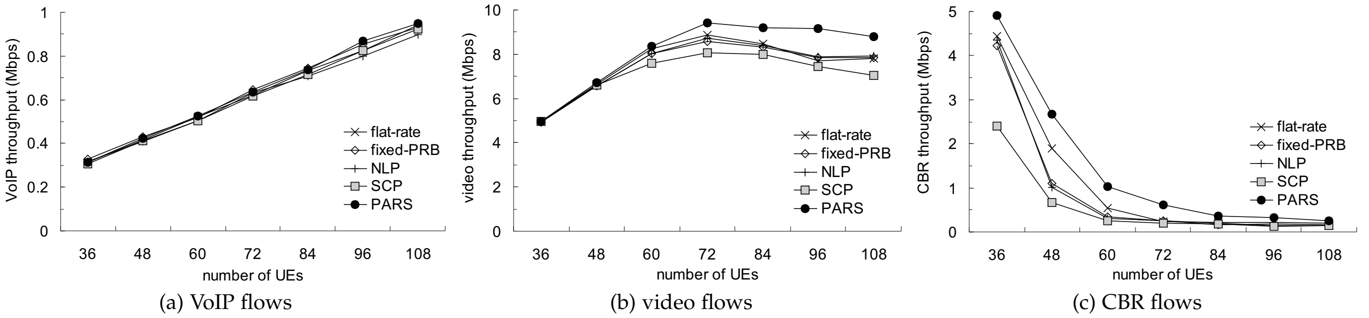
6.3 Network Throughput by Users

Next, we study the throughput of different UEs, as given in Fig. 7. Two scenarios are considered: *light-load* (48 UEs) and *heavy-load* (96 UEs). In the light-load scenario, all methods have similar VoIP and video throughput (referring to Fig. 7(a) and (b)), because they allocate PRBs to GBR flows first and the eNB has sufficient resource to satisfy GBR demand. Thus, the effect of user level becomes insignificant. For CBR service in Fig. 7(c), golden UEs have higher throughput than others due to their high priority. Since PARS achieves higher spectral efficiency in Fig. 5(a), there remain more PRBs for CBR transmission. Such PRBs are given to UEs by their price-based weight W_i^P, so PARS can have much higher CBR throughput for both golden and silver users than other methods.

In the heavy-load scenario, user level has less impact on VoIP throughput in Fig. 7(d), as VoIP service has small traffic demand. On the contrary, video service requests a large amount of data transmission, so golden UEs are allocated with more PRBs for video transmission, followed by silver and bronze UEs, as shown in Fig. 7(e). However, our PARS framework allows low-level UEs to get additional PRBs when high-level UEs incur bad channel condition, thereby resulting in much higher video throughput for bronze users than other methods. On the other hand, since the limited network resource is competed by many UEs, non-GBR CBR service can only receive few PRBs for transmission, as shown in Fig. 7(f).

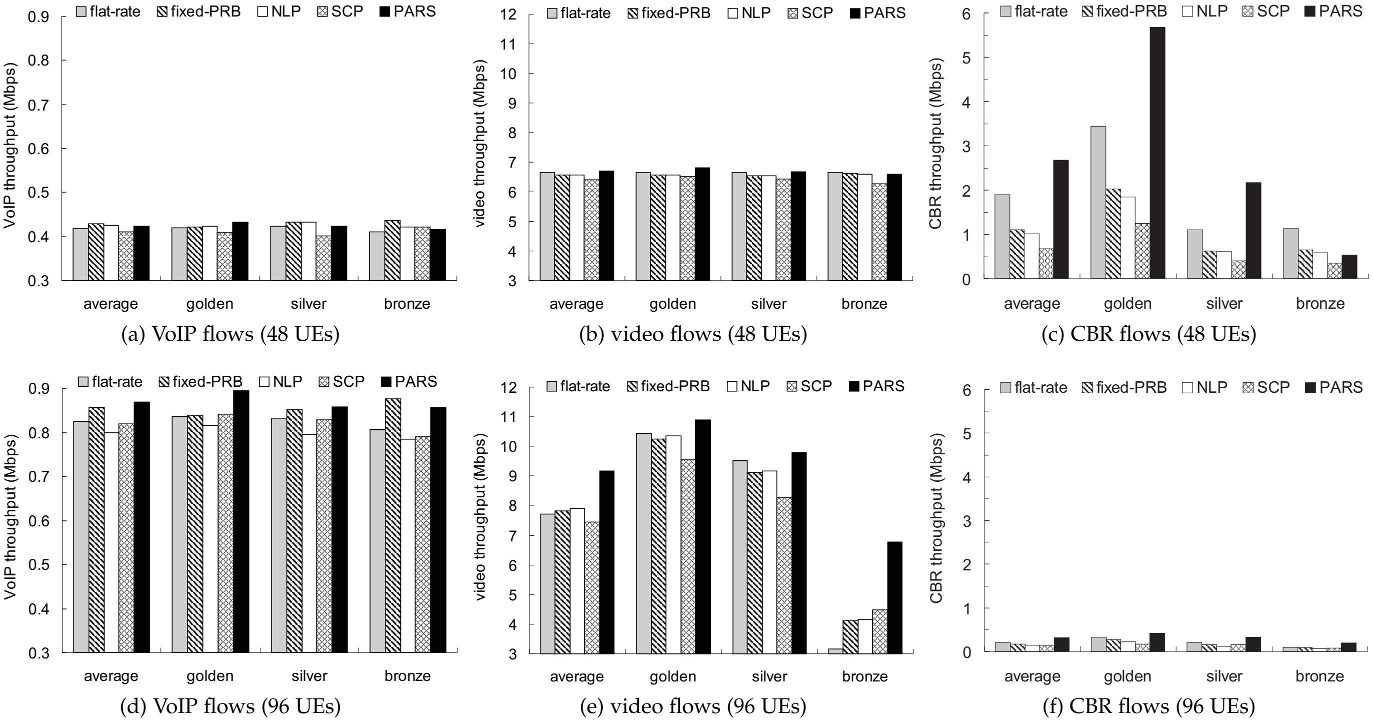
6.4 QoS Support for GBR Flows

We finally investigate QoS support for GBR flows by different methods. Here, we measure the packet loss rate of VoIP and



(a) VoIP flows (b) video flows (c) CBR flows

Fig. 6: Comparison on network throughput by different types of flows.



(a) VoIP flows (48 UEs) (b) video flows (48 UEs) (c) CBR flows (48 UEs)

(d) VoIP flows (96 UEs) (e) video flows (96 UEs) (f) CBR flows (96 UEs)

Fig. 7: Comparison on network throughput by different levels of users.

video flows due to expiration, as shown in Fig. 8. Based on the LTE standard [7], when the latency of a VoIP/video packet exceeds 100ms/150ms, it is dropped. Obviously, a lower loss rate implies that the method well supports QoS for GBR flows.

As discussed earlier, VoIP flows have small demand, so their loss rate can be kept low in Fig. 8(a). However, when the number of UEs exceeds 72, the VoIP loss rate starts increasing in flat-rate, NLP, and SCP. On the other hand, video flows incur a significantly high loss rate when the number of UEs grows, as it is not easy for the eNB to satisfy their large demand. Since the scheduling module of our PARS framework can better utilize PRBs for transmission and allow UEs to exchange the usage of PRBs by their channel quality, PARS thus greatly reduces the video loss rate, as shown in Fig. 8(b).

7 Conclusion

Both issues of resource scheduling and pricing are important in LTE research. However, previous studies deal with them independently and face a difficult choice between improving system performance or increasing operator revenue. To conquer this difficulty, we develop the PARS framework with two tightly-coupled modules. The scheduling module adopts a three-layer strategy to apply priority rules to PRB allocation based on user level and flow type. It also allows the eNB to exchange the usage of some PRBs to improve channel utilization. The pricing module keeps the PED model in mind and charges users for extra money depending on the service they use, so as to increase revenue without significantly degrading user demand. Through LTE-Sim experiments, we compare PARS with the flat-rate, fixed-PRB, NLP, and SCP methods. The results demonstrate that PARS strikes a good balance between performance and revenue, and also supports QoS for VoIP and video service.

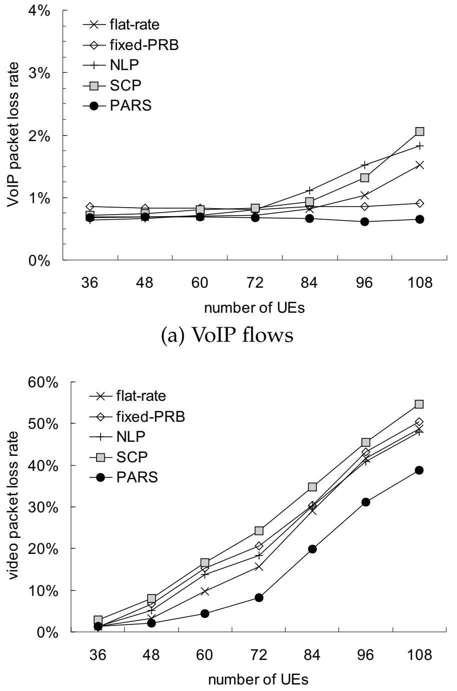
References

[1] S. Srikanth, P.A. Murugesa Pandian, and X. Fernando, “Orthogonal frequency division multiple access in WiMAX and LTE: a comparison,” *IEEE Comm. Magazine*, vol. 50, no. 9, pp. 153–161, 2012.

[2] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger, and F. Xu, “Overview of 3GPP LTE-advanced carrier aggregation for 4G wireless communications,” *IEEE Comm. Magazine*, vol. 50, no. 2, pp. 122–130, 2012.

[3] Y.C. Wang and C.A. Chuang, “Efficient eNB deployment strategy for heterogeneous cells in 4G LTE systems,” *Computer Networks*, vol. 79, pp. 297–312, 2015.

[4] 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.



(a) VoIP flows

(b) video flows

Fig. 8: Comparison on GBR packet loss rate.

[5] L.A. DaSilva, "Pricing for QoS-enabled networks: a survey," *IEEE Comm. Surveys & Tutorials*, vol. 3, no. 2, pp. 2–8, 2000.

[6] N.G. Mankiw, *Principles of Economics, 7th Edition*, Boston: Cengage Learning, Inc., 2014.

[7] European Telecommunications Standards Institute, "Policy and charging control architecture (release 13)," 3GPP TS 23.203, 2015.

[8] 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.

[9] 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.

[10] M. Iturralde, A. Wei, T. Ali-Yahiya, and A.L. Beylot, "Resource allocation for real time services in LTE networks: resource allocation using cooperative game theory and virtual token mechanism," *Wireless Personal Comm.*, vol. 72, no. 2, pp. 1415–1435, 2013.

[11] 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.

[12] 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.

[13] 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.

[14] Y.C. Wang and S.Y. Hsieh, "Service-differentiated downlink flow scheduling to support QoS in long term evolution," *Computer Networks*, vol. 94, pp. 344–359, 2016.

[15] M.S. Mushtaq, S. Fowler, A. Mellouk, and B. Augustin, "QoE/QoS-aware LTE downlink scheduler for VoIP with power saving," *Journal of Network and Computer Applications*, vol. 51, pp. 29–46, 2015.

[16] F. Khan and N. Baker, "Charging data dimensioning in 3G mobile networks," *Proc. IEE Int'l Conf. 3G Mobile Comm. Technologies*, 2004, pp. 183–187.

[17] J. Cushnie, D. Hutchison, and H. Oliver, "Evolution of charging and billing models for GSM and future mobile Internet services," *Quality of Future Internet Services*, Springer, 2002, pp. 312–323.

[18] P. Marbach, "Analysis of a static pricing scheme for priority services," *IEEE/ACM Trans. Networking*, vol. 12, no. 2, pp. 312–325, 2004.

[19] E. Wallenius and T. Hamalainen, "Pricing model for 3G/4G networks," *Proc. IEEE Int'l Symp. Personal, Indoor and Mobile Radio Comm.*, 2002, pp. 187–191.

[20] U. Mir and L. Nuaymi, "LTE pricing strategies," *Proc. IEEE Vehicular Technology Conf.*, 2013, pp. 1–6.

[21] K.J. Astrom and B. Wittenmark, *Computer Controlled Systems: Theory and Design, 3rd Edition*, New York: Dover Publications Inc., 2012.

[22] H. Lee, S. Vahid, and K. Moessner, "A survey of radio resource management for spectrum aggregation in LTE-Advanced," *IEEE Comm. Surveys & Tutorials*, vol. 16, no. 2, pp. 745–760, 2014.

[23] European Telecommunications Standards Institute, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 3GPP TS 36.213, 2012.

[24] 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.

[25] S. Lanning, D. Mitra, Q. Wang, and M. Wright, "Optimal planning for optical transport networks," *Philosophical Trans. of the Royal Society of London A*, vol. 358, no. 1773, pp. 2183–2196, 2000.

[26] Global Knowledge, "Using voice over IP (VoIP) in mobile networks," 2011. [Online]. Available: http://www.globalknowledge.nl/

[27] Ericsson, "Voice and video calling over LTE," 2014. [Online]. Available: http://www.ericsson.com/

[28] European Telecommunications Standards Institute, "Policy and charging control architecture (release 8)," 3GPP TS 23.203, 2010.

[29] B. Soret, H. Wang, K. I. Pedersen, and C. Rosa, "Multicell cooperation for LTE-advanced heterogeneous network scenarios," *IEEE Wireless Comm.*, vol. 20, no. 1, pp. 27–34, 2013.

[30] J. M. Liang, Y. C. Wang, J. J. Chen, J. H. Liu, and Y. C. Tseng, "Energy-efficient uplink resource allocation for IEEE 802.16j transparent-relay networks," *Computer Networks*, vol. 55, no. 16, pp. 3705–3720, 2011.

[31] LTE simulator. [Online]. Available: http://telematics.poliba.it/index.php/en/lte-sim

[32] 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 Journal on Comm. Systems*, vol. 23, no. 2, pp. 213–230, 2010.

[33] 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.

[34] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, Elsevier, 2013.

[35] European Telecommunications Standards Institute, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) requirements for LTE Pico Node B," 3GPP TR 136 931, 2011.

[36] A. Belghith, S. Trabelsi, and B. Cousin, "Realistic per-category pricing schemes for LTE users," *Proc. IEEE Int'l Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, 2014, pp. 429–435.