Madi *et al. EURASIP Journal on Wireless Communications and Networking* (2018) 2018:180
https://doi.org/10.1186/s13638-018-1185-3

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Delay-based and QoS-aware packet scheduling for RT and NRT multimedia services in LTE downlink systems

Nadim K. M. Madi[1*], Zurina Mohd Hanapi[1], Mohamed Othman[1,2] and Shamala K. Subramaniam[1]

Abstract

Guaranteeing Quality of Service (QoS) for heterogeneous traffic is a major challenge in the Fourth Generation (4G) mobile networks. Therein, the absence of sophisticated resources allocation process at the base station jeopardizes QoS in terms of latency data transfer. It has been observed from the literature that low delay bounds might be ensured, however, at the expense of other QoS aspects; for example, throughput and data loss. Therefore, in this article, we propose an effective Delay–based and QoS–Aware Scheduling (DQAS) scheme with a low complexity overhead as an efficient solution for the resource allocation issue in LTE Medium Access Control (MAC) layer. The ultimate aim of DQAS is to minimize delay for Real-Time (RT) traffic while still offering a good level of QoS. Complying with QoS of different traffic types, we effectively analyze the queue buffer of each user flow by developing an algorithm called Efficient Delay Control (EDC) that weights each flow priority in terms of delay. Then, this weight is utilized as a principle for the scheduling decision on the attending flows. Furthermore, the Least Delay Increase (LDI) algorithm is developed to tune the scheduler behavior to maintain a balance between delay and system throughput. Simulation results considering different user mobility levels reveal that DQAS significantly guarantees a low end-to-end delay trend that is independent of increased RT load, and moreover, a reasonable throughput and data drop levels compared to other existing schedulers.

Keywords: 4G LTE networks, Downlink packet scheduling, Resource allocation, Delay, QoS awareness, QoS balancing

1 Introduction

Long Term Evolution (LTE) is recently the most promising mobile technology which allows various multimedia applications to be transferred with a high network capacity and utility [1]. LTE employs Orthogonal Frequency–Division Multiple Access (OFDMA) as a radio access technology at the downlink channel; this grants more flexibility by contiguously utilizing portions of the spectrum to maximize the network performance.

QoS provisioning has been defined as a major objective in 4G LTE radio access networks. Therein, MAC layer scheduling, which is a gist function in radio resource management (RRM) entity of LTE network architecture, presents an immanent challenging issue that seeks effective and realistic solutions to conform with the variety of data traffic evolution.

In channel–aware scheduling [2–5], a trivial scheduling principle is based on the users' reported channel information. This may return a reasonable data rate level. However, in scenarios of multi–traffic types, channel–aware scheduling concept by itself is not sufficient to guarantee a good network QoS performance, particularly on RT applications. According to [6], the maximum tolerated delay for RT applications is defined to be less than 0.1 s and 0.3 s for Voice-over-IP (VoIP) and video flows, respectively; otherwise, traffic QoS is deteriorated. Commonly, queue–related parameters are adopted beside channel rate to allow obtaining a delay–awareness trait. For example, buffer delay with a maximum bound has been adopted in many works (see for instance [7–9]) to assign delay–oriented scheduling weights to different flows. A straightforward QoS improvement might be

*Correspondence: nadim.k.madi@gmail.com
[1]Department of Communication Technology and Networks, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia
Full list of author information is available at the end of the article

realized, the performance, in this case, is compromised during high offered load though. The reason here is that these scheduling rules are designed based on a single-dimension QoS consideration, hence this precludes the scheduler to tweak its behavior to adopt changes of the network load states [10].

In addition to that, minimum delay is observed to be guaranteed only over a single traffic type and at the expense of low QoS on other types. For example, in [11, 12], the algorithms are observed to reduce delay only on burst RT video, this indeed causes a high data loss ratio though. Basically, these schemes deliberately trigger a dropping procedure during the network load congestion states against certain flows in order to alleviate delay and improve throughput on other flows with good channel conditions. Such an excessive dropping event usually leads to a frequent data retransmissions which eventually deteriorates delay and QoS on delay–sensitive and small flows like VoIP.

Motivated by the aforementioned scheduling issues, we remark that maximizing system throughput is not always the main goal, but rather, guaranteeing on–time and full–bits service delivery is the efficient principle for long–term QoS provisioning [13]. Therefore, in this article, we address the problem of radio resources allocation for heterogeneous traffic scenarios. The outcome contributions of this paper are as follows:

- Proposing a novel delay–derived packet scheduling scheme, DQAS, for the downlink LTE MAC channel to minimize the end–to–end delay for RT traffic. DQAS comprises two algorithms to model the MAC scheduling process. In the first mechanism (EDC), different RT flows are analyzed and assigned with a priority weight that is derived from a delay–based formula. The second mechanism (LDI) handles flows with relatively low weights from EDC by examining their impact to the overall delay to possibly be transmitted with a reasonable throughput.
- Evaluating and discussing the performance of proposed DQAS against some recent and well-known schemes over different mobility scenarios with an emphasis on ensuring low delay pattern for RT traffic and near-to-ideal QoS balancing.

It is remarkable that, implementing DQAS as a MAC scheduler returns a minor overhead computational complexity. In addition, it provides a robust and consistent behavior on sustaining low end–to–end delay that is independent of the increased network traffic load.

The remainder of the paper is organized as follows: Section 2 elaborates on the related works of the relevant scheduling issues. In Section 3 the involved downlink system model is elaborated. Section 4 describes DQAS as a MAC scheduler and its components with the complexity analysis. Afterward, Section 5 demonstrates the performance evaluation and the simulation experiments. The numerical results of the simulation are then discussed in Section 6. Finally, conclusion and the future direction of this work are presented in Section 7.

2 Related Works

Traditionally, packet scheduling techniques in wireless systems are classified as elastic–based and non–elastic–based. Such a classification tends to be not applicable for 4G mobile networks due to the diverse QoS characteristics of the emerged multimedia applications. Instead, a reasonable classification of the scheduling schemes should be according to their QoS specific target. The implemented scheduling schemes in LTE can be grouped into two sets that are, throughput–driven, and delay–driven which are discussed in the following context.

2.1 Throughput-driven schedulers

An unpretentious rule of maximizing throughput, known as Maximum Throughput (MT), favors flows with the best channel qualities [2]. Further investigations by [3] and [4] were involved on MT considering system capacity and/or complexity. MT is known for its unfair service. Therefore, the Proportional Fairness (PF) scheduler in [5] has been vastly adopted to guarantee a fair service beside utility maximization under variable channel condition. This is usually realized by providing a prioritized service to the flow minding its historical rates. Since then, PF has been extended for throughput optimization in many proposals, for example in [14], the scheduler behavior is tuned by means of predefined tunable parameters to keep throughput-to-fairness balancing relation. Besides, in another proposal, [15], PF was modified in a way to produce a soft decrease of throughput against the increased network load. These proposals are able to show a reasonable performance, however only on Non-RT flows. Scheduling algorithms in [16] and [17] were designed by adopting queue-state information such as Channel Quality Indicator (CQI), queue size, or arrival rates where system throughput improvement might be seen to some extent. Furthermore, authors in [18] formulated the packet scheduling problem under Markov Decision Process (MDP) and introduced a value iteration algorithm to improve throughput on RT traffic. Besides, authors in [19] leveraged the carrier aggregation concept to improve the system throughput, whereby users are assigned to component carriers based on their channel gains.

2.2 Delay-driven schedulers

Schedulers under this category are mostly developed for delay–sensitive or RT applications. A basic and common

delay–aware scheduling rule named Modified–Largest Weighted Delay First (M–LWDF) by [8] was initially proposed for CDMA mobile systems, and yet, has been vastly adopted in LTE. The idea behind M–LWDF depends on flows delay that is measured from the flow Head of Line (HoL) delay. Besides, it adopts the PF behavior to enforce a fair service level between different flows. In fact, like M–LWDF, many other works also rely on PF concept's properties[1] to invoke their scheduling decisions. For instance, authors in [20] extended the PF concept to a delay–based scheduler where buffer status in current and previous scheduling interval is accounted. The scheduling rule emphasizes on fair delay distribution instead of actually minimizing delay time. In [21], PF is joined with a user–defined delay threshold below the maximum bound and a ratio of allocated resource blocks are utilized to design a scheduling rule. Although low delay can be established, ensuring a long–term low delay during the network overload states using the statically determined parameters tends to be an elusive goal, especially, in case of burst Variable Bit Rate (VBR) traffic where long buffers are common. Besides, in another proposal [22], flows urgency level is defined using thresholds of minimum and maximum delay bounds; this allows the delay–based algorithm with a scheduling decision that favors highly urgent flows.

The known EXPonential–PF (EXP–PF) in [23] shows a decent behavior of QoS in both RT and NRT traffic. EXP–PF benefits from the PF principle to keep an opportunity for NRT to be scheduled, meanwhile maintaining low delay bounds on certain RT traffic[2] by employing an exponential function of HoL delay. The exponential term was also used by [7] to introduce a delay–driven scheduling principle (EXP–Rule), therein flows of long buffers are selected to be scheduled in order to maintain a queue balancing. A variant algorithm of M–LWDF was introduced by analyzing queue HoL delay to decrease the probability of exceeding the maximum delay bound [24]. The proposal nonetheless causes a high amount of data drop because it imposes a hard dropping probability that grows exponentially on RT traffic. Furthermore, in [25], M–LWDF was manipulated with a location–based parameter to improve QoS on the cell–edge UEs. Authors in [26] remarked that, under a simple Markov decision process, the delay optimization problem can be proved by radial sum monotonicity. Accordingly, they proposed a delay–based scheduler named "Log–Rule" that serves users in a way that de–emphasizes on queue balancing to enable more delay–based behavior decision, however on the expense of system throughput.

Dwelling further, the Time Domain (TD) scheduling mechanism in [11] estimates the pattern of incoming packets to the queue. A Delay threshold is then calculated based on the number of expired packets. Nevertheless, QoS of small flows is severely deteriorated by this mechanism. This is due to the assumption that traffic always comes in burst, and thereby data loss is tolerated for alleviating the delay. Depending on this assumption indeed drive other small flows to suffer from high delay whereas burst flows may benefit from the channel resources. With the aim of managing different flows types, recent research attempts, for example, [27] employ the concept of Active Queue Management (AQM) in Transmission Control Protocol (TCP). Therein, delay is still defined as a major issue unless the traditional TCP architecture is enhanced to balance between both reliability and latency.

On the other hand, the literature reveals contributions of resource allocation schemes (i.e. [28–30]) in which the issued of delay is implicitly figured out by controlling the energy consumption at the link layer. Therein, a concept known as "Discontinuous Reception" (DRX) was adopted by [28, 29] and is debated to minimize delay by controlling DRX cycles UE sleep mode (normally shorter periods returns lower delay). In [30], the authors utilized a duty cycle control technique to manage the active states on UEs complying with a derived transmission policy which is believed to possibly reduce end-to-end delay.

Based on the several investigations handled on the aforementioned studies, the following observations can be remarked:

- The literature is still seeking more effective solutions for packet scheduling which cultivate QoS provisioning on both RT and NRT traffic in LTE. Some of the existing algorithms, i.e. [7] [11], only rely on the assumption that traffic is always offered to the network in burst and thereby data loss is a common trend. In fact, this assumption may not be practical in multi–traffic scenarios, since flows such as VoIP does not tolerate heavy data loss which eventually increases its delay when being transmitted with other burst traffic. On the other hand, adopting a relaxed delay threshold as in [22], or statically assigning resource blocks quota as in [21] leads to QoS deterioration and high delay for VBR flows which is obviously noticed at the traffic congestion states.
- With the variety of traffic types, multiple dimensions of QoS such as data rate, data loss, and delay should be satisfied. Practically, it is impossible to maximize all of the three aspects, although a reasonable tradeoff can be established among the metrics. The optimal QoS situation is approached by balancing these aspects, hence improving the performance on different applications is obtained complying with their QoS characteristics. Up to this point, we notice that majority of the related works do not actually emphasize on this QoS premise as a long–term goal

to be achieved by their different proposed scheduling schemes.

By introducing our proposed scheduling model (DQAS), a heterogeneous traffic scenario is considered where RT flows are examined in terms of their delay specifications. In addition, QoS of NRT flows is also reckoned by pledging a sufficient data rate. Without the loss of generality, the core principle in developing DQAS cultivates enhancing multimedia performance in a way that does not severely deteriorate that balanced state between QoS metrics. Within the next section, the downlink system model intended for the proposed packet scheduling scheme is thoroughly introduced.

3 Downlink System Model

The downlink channel in LTE adopts OFDMA as an access technology. This allows empowering the system with many features, for example, enormous transmission rates, scalable and efficient bandwidth planning, and high consistency against multi–path fading [31]. According to OFDMA channel model in LTE Release 10 [32], a wide range of radio spectrum up to 20 MHz is supported, which allows a high order of available radio resources to carry user's data. The basic radio resource unit that can be transferred in LTE channel is known as a Resource Block (RB). The RB structure is depicted in Fig. 1, where each RB is composed of 12 consecutive sub–carriers and occupies a size of 180 kHz in Frequency Domain (FD). In TD, the RB sustains for 0.5 ms length and is corresponding to 7 OFDM symbols. Basically, radio resources are usually assigned to the scheduled flows every Time Transmission Interval (TTI) that lasts for 1 ms. Therefore, the resources allocation procedure in TD/FD circles over the available sub-channels every TTI wherein a Physical Resource Block (PRB) (consisted of two RBs) is mapped to each attended flow.

At the typical LTE network scenario, each connected User Equipment (UE) reports CQI via the uplink channel to evolved NodeB (eNB) as an estimation of its link efficiency. Thereafter, eNB may decide whether to use the reported CQI or adjusting it based on the service requirements. It is important to note that all the resources allocation and scheduling functions are managed and handled by the eNB. Thereupon, the downlink packet scheduling process is initiated after eNB receives the instantaneous CQI feedback from the involved UEs. These feedbacks are usually triggered periodically with intervals of several tens of TTIs. At the eNB side, the scheduling process takes place associated with the CQI and QoS-related parameters. For each selected PRB to UE flow, the Signal–Interference–plus–Noise–Ratio (SINR) calculated using reported UE's CQI. Then the Adaptive Modulation and Coding (AMC) exploits the SINR to select the proper MCS for the allocated PRB; this procedure actually ensures low channel errors during the transmission process. Using the MCS, Transport Block Size (TBS) can be decided for the UE's payload to be carried with the allocated PRB. Information of RBs mapping, as well as the determined MCS, are reported back to the respective UE via the Physical Downlink Control Channel (PDCCH). Eventually, the UE decodes the PDCCH payload and checks if it is permitted to be scheduled by the eNB and possibly access the Physical Downlink Shared Channel (PDSCH) to receive its requested flow.

4 The Proposed Delay–based and QoS–Aware Scheduling Scheme

Basically, DQAS is a delay–based scheduling scheme that operates on both TD and FD. The major objective of DQAS is to efficiently reduce latency on RT traffic while not compromising QoS of different RT and NRT flows that are sharing the eNB available resources. DQAS comprises two mechanisms in the MAC scheduler: Delay–oriented flows analysis via EDC algorithm, and QoS–aware procedure for throughput–oriented flows using LDI algorithm. A general model of DQAS components with their relations is illustrated in Fig. 2. Intuitively, each TTI, a complete scheduling event occurs at MAC layer. Wherein, the list of flows to be scheduled is updated and assigned with the specific amount of bandwidth by means of upper MAC level functions. Then, flows are selected to admit at the lower MAC level where DQAS scheme is located for flows analysis and RBs allocation process.

In DQAS, RBs allocation is decided on the selected flows via delay–derived rules to meet QoS requirements of LTE RT traffic [6]. As the first part of DQAS, EDC algorithm is developed to efficiently examine different flow types and adhere tag them with identical weights. Accordingly, flows with the highest weights are handled by an FD RBs allocation. Meanwhile, flows with low metric weights are treated with LDI algorithm hence a scheduling procedure with throughput emphasis is imposed as long as low impact on delay experienced by each flow. With these efficient procedures, it is evident that delay is minimized on delay–sensitive traffic while other QoS aspects are maintained in a good level. In the following context, DQAS scheme is discussed in further details.

4.1 Data Flows QoS Analysis: The EDC Algorithm

EDC is proposed as a component in DQAS scheme to alleviate severe delay on RT traffic, particularly, on flows with a high level of delay sensitivity (*i.e.* VoIP) as defined in [6]. EDC assigns a metric weight to each selected flow by the upper MAC level. Although the task may sound primitive, it becomes more tedious when traffic

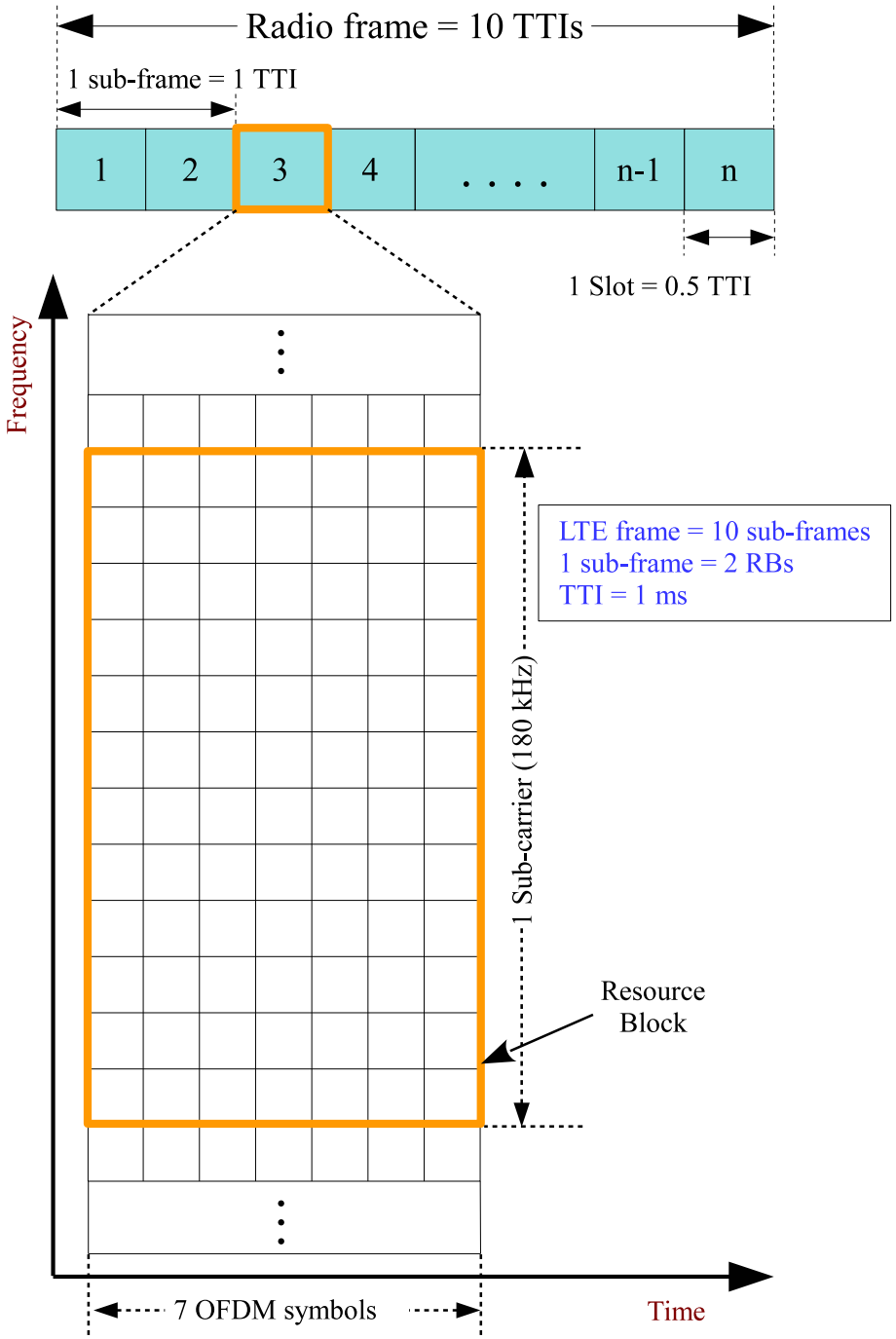

Fig. 1 Illustration of LTE frame and RB structure in FD and TD

with assorted QoS specifications are transmitted using a shared channel. It is common that burst traffic dominates the channel capacity leaving other small flows with low service. This sparks the significance of EDC by controlling delay on burst RT traffic flows, so that small RT flows like VoIP are able to be transmitted with low latency and high rates. NRT flows are handled with the minimum acceptable service rate. It is worth mentioning that EDC is developed as a service–based and lightweight mechanism in which QoS of different RT flows is analyzed based on their class features. For more understanding, Algorithm 1 explains the procedures handled in this scheduling phase. In addition, for clarity, Table 1 defines the control parameters and notations used throughout DQAS components.

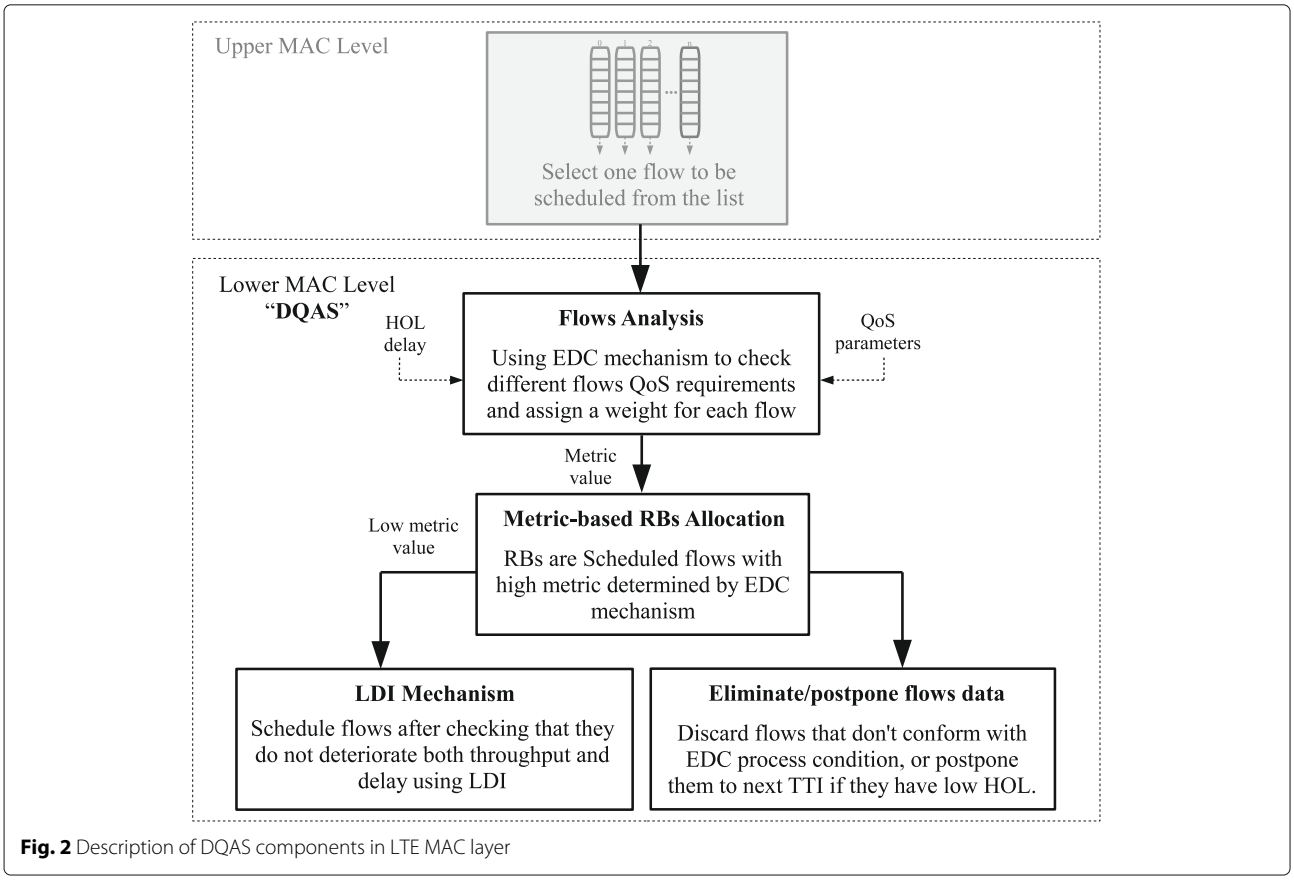

Fig. 2 Description of DQAS components in LTE MAC layer

Table 1 Control parameters and notations used in DQAS

Variable	Indication
TTI	Incremented scheduling interval value
$M[j][i]$	Priority metric of flow j on RB i
$D_{voip,j}$	delay term for flow j, $\forall j \in$ VoIP
$D_{video,j}$	delay term for flow j, $\forall j \in$ VBR Video
$maxM[j][i]$	maximum value of $M[j][i]$ in TTI
$D_{HOL,j}$	Head of line delay on flow j (buffer delay)
$\overline{D_{HOL}}$	Average HOL delay for available flows
D_{max}	Max delay bound (set to 0.1 sec [6])
α	Delay-throughput optimality parameter
δ	Slope coefficient
γ	$D_{voip,j}$ balancing factor
σ_j	Delay-drop stabilizing index
μ_j	Data rate of user flow j
$\overline{\mu}$	Average data rate for a user flows
Thr_{dec}	Decrease on average throughput
D_{inc}	Increase on buffer delay
$I[min(D_{inc})]$	index of minimum delay increase
$I[min(Thr_{dec})]$	Index of minimum throughput decrease

The state–of–art principle in designing EDC is to transmit RT flows with minimum delay values that are independent of either the increased load or the channel variability. Therefore, considering the buffer delay $D_{HOL,j}$ such that j belongs to RT VoIP class, the metric weight is determined as,

$$M[j][y] = D_{voip,j}.\mu_j \tag{1}$$

$$D_{voip,j} = \frac{\delta}{\exp\left(\alpha.\frac{\overline{D_{HOL}} - D_{HOL,j}}{-\ln(D_{HOL,j}).\sqrt{D_{HOL,j}}}\right)} + \gamma \tag{2}$$

hence,

$$\alpha = \frac{-\ln(\sigma_j)}{D_{max}} \tag{3}$$

In general, VoIP is characterized as lightweight traffic with packets of fixed size [33]. With such a low traffic variability, the variation in flows' delay distribution is comparatively small. It's important to remark that, in this study, it is assumed that UEs always have data to be transmitted every TTI. Adhere, buffer delay values on different flows are mostly greater than zero.

The rationale in Eqs. 1, 2, and 3 is motivated by the traditional scheduling rule in [23]. Wherein, the rule logic is reformulated in this study to guarantee the minimal delay for small RT flows such as VoIP. In details, according to Eq. 2, a tight delay control is realized on VoIP flows as long as the exponential term of $D_{voip,j}$ is obtained in small values. This requires that the difference between $D_{HOL,j}$ and $\overline{D_{HOL}}$ to be kept low, and limited by $(\ln(D_{HOL,j}).\sqrt{D_{HOL,j}})$ at any time instance. If $D_{HOL,j}$ is less than $\overline{D_{HOL}}$ by maximum the value of $(-\ln(D_{HOL,j}).\sqrt{D_{HOL,j}})$, then flow j is granted a high priority. This intuition can be expressed by the following condition.

$$\overline{D_{HOL}} - D_{HOL,j} < -\ln(D_{HOL,j}).\sqrt{D_{HOL,j}} \tag{4}$$

$$\frac{1}{K}\sum_{j\in K}(D_{HOL,j}) - D_{HOL,j} < -\ln(D_{HOL,j}).\sqrt{D_{HOL,j}} \tag{5}$$

Assume that at an instant TTI, there is a number of VoIP flows, $K > 1$, whose delay values are closed to each other and upper bounded by D_{max}. This case makes the following relation holds true.

$$\overline{D_{HOL}} - D_{HOL,j} < D_{HOL,j} \tag{6}$$

By definition, it is valid that $[x < -\ln(x).\sqrt{x}\ \forall\ x \in [0, 0.1]]$. This denotes that the change rate on the right–side of relation 4 is always influenced by a logarithmic increase over all the occurrences $(x_1, x_2, .., x_K)$. Therefore, from relation 5 we have,

$$D_{HOL,j} < -\ln(D_{HOL,j}).\sqrt{D_{HOL,j}} \tag{7}$$

Based on relations 6 and 7 it is obvious that relation 4 is valid when $0 < D_{HOL,j} < D_{max}$. This implies that the exponential term in Eq. 2 is often kept in a small range and controlled by the parameters δ and γ (details on the impact of these parameters is elaborated in Section 4.1.1 below) in order to obtain high $D_{voip,j}$ weights. As demonstrated by the curves in Fig. 3, the behavior of $D_{voip,j}$ function increases with an exponent trend. Therein, VoIP scheduling decision enforces the highest concentration on minimal delay transmissions compared with other existing flows types in the scenario. Besides, the significance of μ_j raises at the phases of high traffic congestion to ensure assigning high weights values to VoIP flows so that they are prioritized for scheduling side to side with burst traffic.

RT video can be characterized as a VBR with burst traffic application in LTE networks traffic [6]. This means that the generated flows are of different sizes and have a vacillated buffer delay. With that, it is essential to control the buffer delay in an efficient way to suit the variability nature of such a traffic type without compromising its QoS. In EDC, the priority weight for VBR RT video flows is formalized considering a dynamic bound for buffer delay. The metric is expressed as,

$$M[j][y] = D_{video,j}.\mu_j \tag{8}$$

$$D_{video,j} = \alpha.(\overline{D_{HOL}} - D_{HOL,j}) \tag{9}$$

where α is calculated as in Eq. (3); besides its influence on the scheduling decision is generously described in Section 4.1.1. We comply with the following Remark as the principle decision to schedule the flows of burst VBR type.

Remark 1 *During the network overload states, delay on burst traffic is significantly improved by considering a dynamic threshold such as* $\overline{D_{HOL}}$ *to control high values of* $D_{HOL,j}$ *in a way to deemphasize on buffers balancing.*

Note that the notion of providing QoS by deemphasizing on different queues balancing has been previously followed by [26] to ensure throughput optimality over small traffic flows. In this work, we rather manipulate this principle in a unique scheduling rule (seen by Eq. 8) to obtain minimal–delay transmissions on bust traffic.

To reasonably explain the impact of the priority weight expressed by Eq. 9 based on the above-stated Remark, assume a scenario in which two VBR video flows (i.e. j_1, and j_2) are waiting for the service to be scheduled. Let j_1 has a bigger buffer size than j_2, and both has a similar channel quality index. If the scheduling decision is solely constructed on the buffer delay value $D_{HOL,j}$, then j_1 will always be prioritized over j_2, and the calculated metric increases by the value of $D_{HOL,j}$. This eventually lefts no chance for j_2 to be hereinafter selected by the scheduler. Therefore, by sticking to Remark 1 above, the scheduler's behavior is induced to favor flows with relatively small buffers by controlling $D_{HOL,j}$ values up to the order of $\overline{D_{HOL}}$ for K flows in a specific TTI. In other words, deemphasizing on queues balancing, in this case, allow controlling the scheduling decision to extensively rely on minimizing delay rather than providing a fair service.

The impact of the delay term ($D_{video,j}$) on the scheduling decision under escalated $D_{HOL,j}$ is demonstrated in Fig. 4. Note that the delay improvement area is closely affected by the $D_{HOL,j}$ values of the attended flows. If for instance, the majority of flows have relatively low $D_{HOL,j}$ values, then delay improvement area for $D_{video,j}$ grows up to the limit of $\overline{D_{HOL}}$ in a certain TTI.

In details, for K flows belong to RT Video application in a specific TTI and the measured $D_{HOL,j}$ for each flow j,

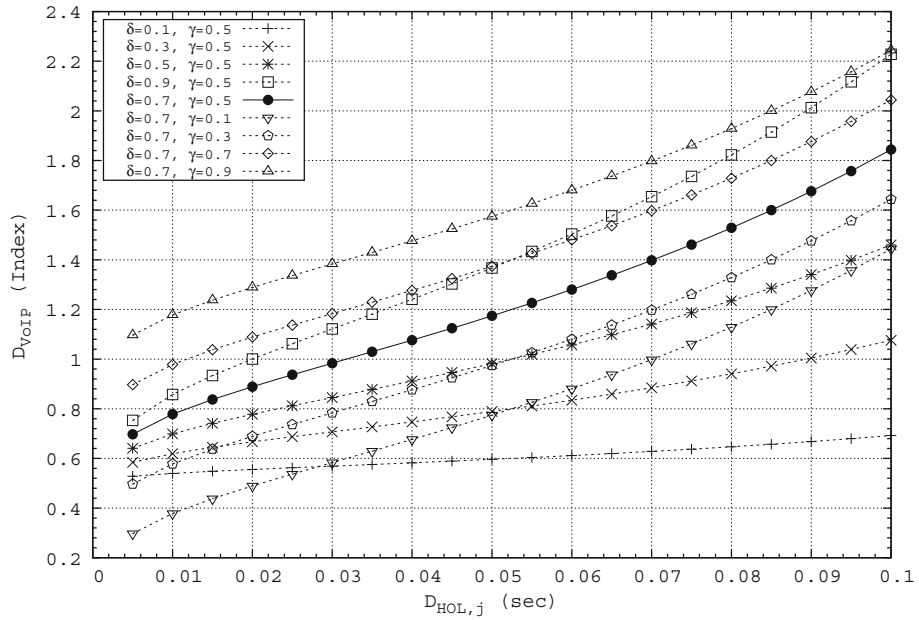

Fig. 3 The behavior of $D_{voip,j}$ against the increased $D_{HOL,j}$ and based on different settings of δ and γ

$\overline{D_{HOL}}$ is calculated as

$$\overline{D_{HOL}} = \frac{1}{K}.\sum_{j=0}^{K} D_{HOL,j} \qquad \forall j \in K, K \subseteq \mathbb{N} \tag{10}$$

Equation 10 implies that a proportional relation is established between $D_{HOL,j}$ and $\overline{D_{HOL}}$. This relation is usually bounded by the diversity range of $D_{HOL,j}$ values within the set of K. With $K > 1$, the distribution of $D_{HOL,j}$ values over different time scales is mostly lower than $\overline{D_{HOL}}$ orders; this means that,

$$\lim_{K\to\infty} \frac{D_{HOL,j}}{\overline{D_{HOL}}} \leq 1 \tag{11}$$

Equation 11 reveals that the ratio of $D_{HOL,j}$ to $\overline{D_{HOL}}$ is always maintained less than or equal to 1 as the amount of K flows offered to the network channel grows linearly. Moreover, it is indicated that the value of $\overline{D_{HOL}}$ increases proportionally as the value of $D_{HOL,j}$ incremented on each certain flow j in the system. Therefore, $\overline{D_{HOL}}$ can be adopted as an effective and dynamic delay threshold to define the tolerated delay bound for flows each TTI. This directly strikes a tight delay control to restrict different queues buffers from growing extensively, especially during traffic congestion phases. Consequently, low delay values are still maintained for small flows as they share the channel resources with other heavy flows.

It is important to remark that, α plays an essential rule in weighing the metric value shown in Eq. 9 such that the scheduling decision balances the prioritization basis towards either throughput (gained by (μ_j)) or delay ($\overline{D_{HOL}} - D_{HOL,j}$). Therein, σ_j should be properly selected for RT traffic so as to avoid excessive packet discarding by the time not violating D_{max} bound. Numerical demonstrations on this regard are provided in Section 6.2.

FunctionFunction end EventEvent doend

Algorithm 1: The EDC algorithm in DQAS.

Initialization:
  define N as list of RBs at TTI;
  define K as list of selected flows to be scheduled at TTI;
  set $M[j][i] \leftarrow 0, maxM[j][i] \leftarrow 0$;
  set $\overline{D_{HOL}} \leftarrow D_{HOL,j}$;
  set $TTI \leftarrow 0$;
Event *On TTI* **do**
  for $i \leftarrow 1$ **to** N **do**
    for $j \leftarrow 1$ **to** K **do**
      Compute: α base on Equation (3);
      Update: $\overline{D_{HOL}} \leftarrow \frac{1}{K}.\sum_{j=0}^{K-1} D_{HOL,j}$;
      if $(j \in VoIPClass)$ **then**
        Update: $M[j][i]$ base on Equation (1);
      else if $(j \in VideoClass)$ **then**
        Update: $M[j][i]$ base on Equation (8);
      else
        Update: $M[j][i] \leftarrow \mu_j$;
      end
      if *j not scheduled in TTI)* **then**
        if $(M[j][i] > maxM[j][i]$ **then**
          Update: $maxM[j][i] \leftarrow M[j][i]$;
          Schedule flow j on RB i;
        else
          Invoke LDI Mechanism;
        end
      end
    end
  end
  Update: $TTI \leftarrow TTI + 1$;
end

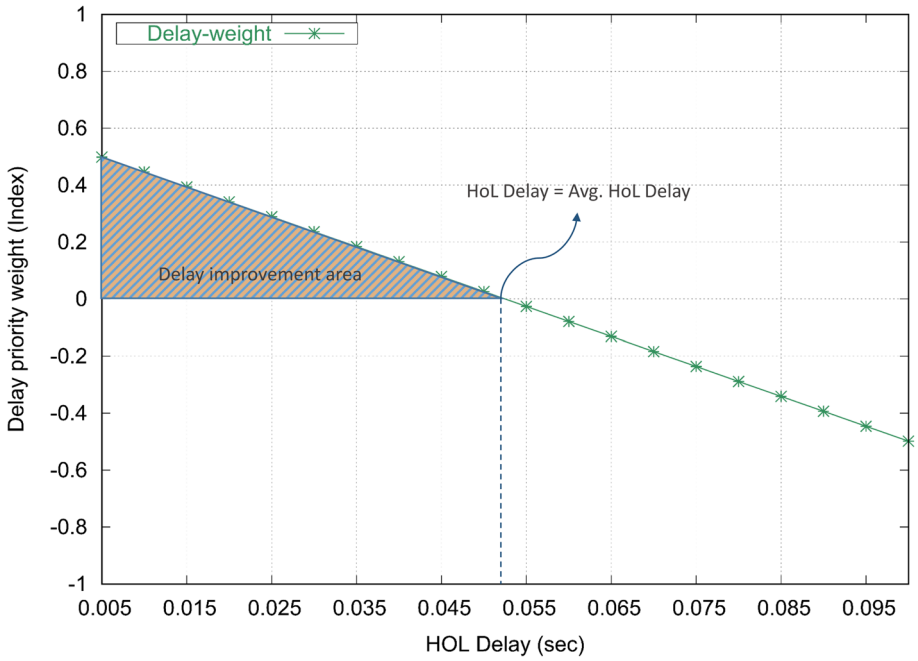

Fig. 4 Impact of $D_{video,j}$ on Video metric weight with the increased $D_{HOL,j}$

For NRT traffic, delay is not a crucial parameter, hence QoS on such a type can be satisfied by exploiting the channel quality condition. Therefore, for simplicity, the scheduling rule, in this case, advocates a channel–aware decision which depends on the channel achievable rate of the UE μ_j. This allows flows to be transmitted in high data rates and thus, fewer errors are guaranteed.

4.1.1 Choice of Parameters

In Eq. 2, the tunable parameters δ and γ are adopted to keep the scheduling function in a proper scale and their values range is defined between $\{0, 1\}$. In specific, δ controls the slope of $D_{voip,j}$ function, whereas γ is a stabilizing factor that anchors $D_{voip,j}$ within a certain range so that it does not interfere with RT video metric function domain. The impact of these two parameters on the behavior of $D_{voip,j}$ slope is clearly demonstrated by Fig. 3. Based on our observations, setting δ and γ to $[0.7, 0.5]$, respectively, is found to ensure an effective delay–based influence on the priority weight as subsequently appears within the simulation results section. The reason behind recommending these values is that the VoIP function, in this case, has the minimum probability to intersect with the slop of video priority weight function ($D_{video,j}$), and thereby allowing more VoIP flows to be flexibly scheduled based on their delay budgets.

On the other hand, α is denoted as a dropping sensitivity factor that balances the behavior of EDC algorithm between maintaining low delay and data drop. In other words, α moderates the distribution of flows delays. Increasing the value of this parameter on particular flows enforces a hard delay reduction however at the expense of other flows. To inherit a QoS–derived impact, α is determined from Eq. 3 as a relation between the logarithmic value of σ_j and D_{max} hence $0 < \sigma_j < 1$. Configuring different values of σ moves the biased point of α to influence the scheduler behavior between minimizing delay and increasing data drop. For purpose of simplicity, in this manifest that all users are assumed to require a similar delay QoS, so, the assigned σ and D_{max} values are same for all users in the system. In this work, D_{max} is set as low as 0.1 sec which is the delay bound that satisfies most of RT applications while a value of $\sigma = 0.35$ is utilized in order to obtain a tight delay control while still holding a reasonably low data drop level. Furthermore, the impact of σ on QoS metrics is generously demonstrated with numerical results in Section 6.2 under DQAS performance evaluation.

4.2 The LDI Mechanism

According to the exhibited analysis carried on EDC mechanism within the previous Subsection, it is evident that latency on different RT flows can be minimized by enforcing efficient priority rules on various service types. As a matter of fact, due to traffic variability, guaranteeing low delay while achieving a good level of data rate is realized with a certain level of tradeoff. Therefore, in scenarios where burst RT traffic like the video is involved, it is

valuable to allow flows that are able to enhance the system throughput to be scheduled as long as they do not severely deteriorate the overall delay. In this essence, the LDI mechanism is proposed to grant flows aborted by EDC algorithm a chance of being scheduled on a different basis. It is important to highlight that, the attained flows by LDI mechanism may have relatively high buffer delay values comparing with those selected by EDC. Nevertheless, some of these flows have a good channel quality hence they can be transmitted with high throughput. The significance of LDI is to improve the QoS of RT and NRT flows relaying on the intuition that data rate is a major characteristic for all flow types. Therein, the following remark is availed of as a principle scheduling decision in LDI.

Remark 2 *In LDI, flows with relatively low* $M[j]\,[y]$ *values still can be scheduled considering their data rates as long as they have the least impact on delay increase.*

It's noteworthy that the generic idea of Remark 2 is inherited from the scheduling scheme in [11], wherein the flows which have the least effect on throughput degradation are scheduled to be transmitted. However, in LDI mechanism, the idea has been further extended and significantly enhanced in a way to grant flows that do not heavily contribute to the overall delay a chance to be selected for scheduling.

The control flow diagram of LDI mechanism is demonstrated in Fig. 5. The process starts by reckoning for the set of remaining flows L which have not been scheduled yet, having that flow j at this instance does not return the highest metric value based on EDC as well. A flow $l \in L$ is then picked from the set of available flows defined above, such that, $l \neq j$. The average throughput for UE with j, or l ($\overline{\mu}$) is then obtained as stated in [34] by the formula,

$$\overline{\mu} = (1-\beta).\overline{\mu} + \mu.\beta \tag{12}$$

Where β is an efficiency constant parameter that is set to 0.2, and μ refers to the actual data rate on flow j or l. This achievable data rate depends on the channel health (usually determined by SINR) between the UE and the eNB, and can be estimated using the well–known Shannon capacity formula [35],

$$\mu = b.\log_2(1+\tau) \tag{13}$$

where τ denotes the SINR of the UE channel and b is the bandwidth size of the subchannel, (i.e. RB). Now, the decrease in throughput (Thr_{dec}) caused by flow j is determined as,

$$Thr_{dec} = \overline{\mu}(l) - \overline{\mu}(j) \tag{14}$$

Complying with the principle by Remark 2, if the throughput decrease by Eq. 14 is less than the minimum throughput decrease index ($I[\,min(Thr_{dec})]$), then j is said to have the minimum decrease on throughput. Thereupon, $I[\,min(Thr_{dec})]$ is updated by the value of Thr_{dec}. To maintain a balanced relation between delay and achievable throughput, flow j should also be ensured to severely increase the delay. Subsequently, the minimum increase on delay (D_{inc}) is calculated between j and l as,

$$D_{inc} = D_{HOL,l} - D_{HOL,j} \tag{15}$$

Likewise, if D_{inc} is less than the minimum delay increase index ($I[\,min(D_{inc})]$), then j is said to add a minor budget to the overall traffic delay. Herewith, $I[\,min(D_{inc})]$ is updated by D_{inc} value and flow j is declared to be scheduled by assigning it to an RB for its data transmission. It's apparently noticed that the two indices $I[\,min(Thr_{dec})]$, $I[\,min(D_{inc})]$, which are frequently updated, restrict flow j to be scheduled unless it obtains the minimal deterioration on throughput and delay with respect to any $l \in L$. If in case j does not has the minimum D_{inc}, it is allowed to be compared with the rest of flows in the list L. Flow l is presumably dropped from MAC and RLC[3] layers at the end of the TTI if it is unable to provide the minimum throughput decrease.

Based on the narrated discussions and analysis about the two components of DQAS scheme, it is obvious that diverse QoS requirements for delay–sensitive and throughput–targeted applications are tightly considered. Moreover, the concept of scheduling flows in two different bases distinguishes DQAS as a robust scheduler to provide an effective low level of delay and good throughout which is independent of the variable traffic nature offered to the system. In the following subsection, the overhead complexity is examined on DQAS components.

4.3 Complexity Analysis

In this context, the overhead complexity analysis of DQAS scheme is demonstrated based on the allocation time per TTI. Assume that at an instant TTI there is a number of K UEs' flows from different traffic sources are imposed to the scheduler seeking to be assigned to RBs for transmission, hence the total available RBs is N. Mind that, incoming flows from the upper level of MAC layer (as depicted in Fig. 2) are of different types. At the lower MAC level where RBs allocation process takes place, these flows are handled sequentially and for once over the DQAS procedures. This means that DQAS procedures (with its two mechanisms) are triggered once every TTI. The DQAS overhead computational complexity, T_{DQAS}, can be determined from both of its mechanisms (EDC, and LDI) as,

$$T_{DQAS} = T_{EDC}.T_{LDI} \tag{16}$$

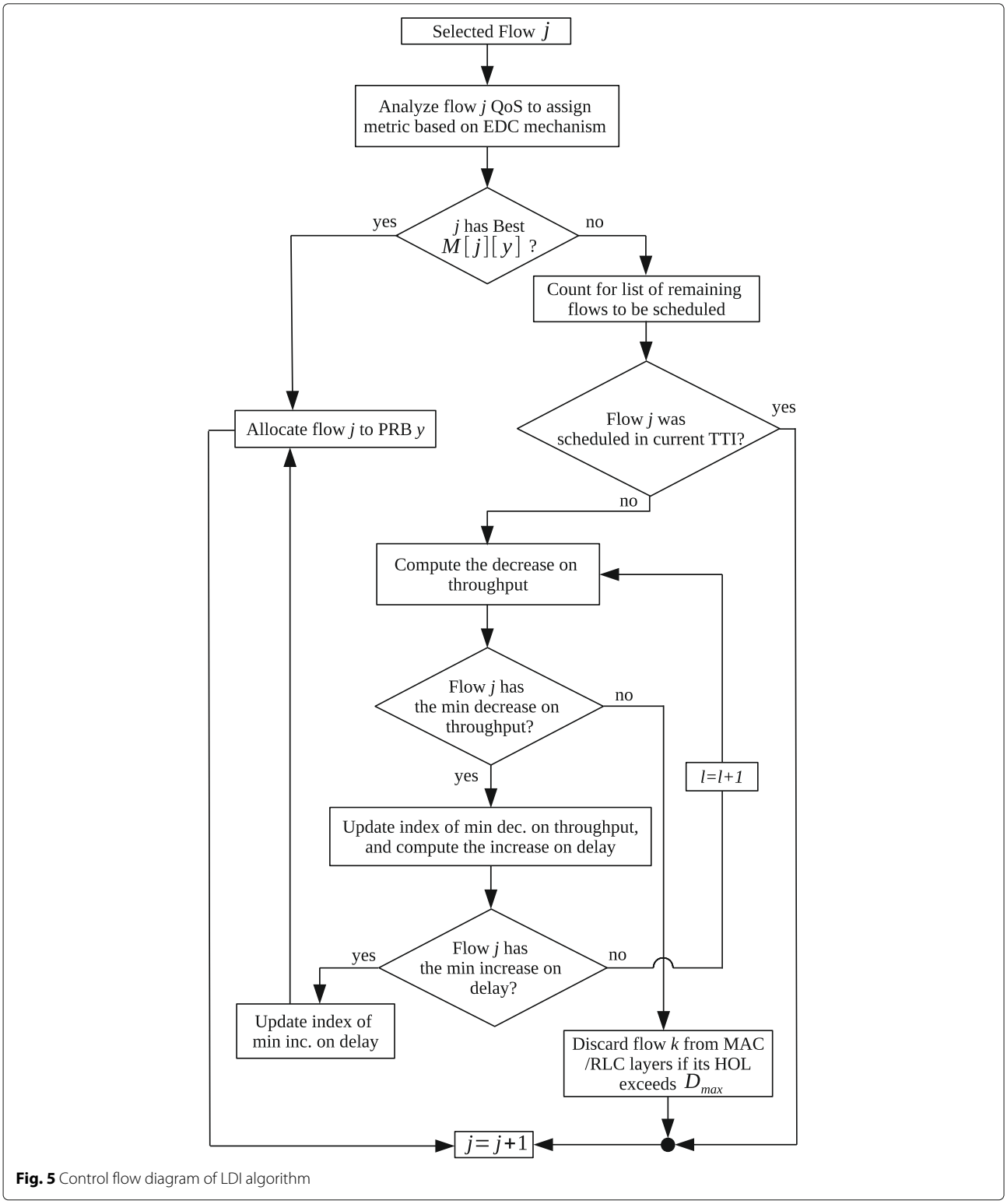

Fig. 5 Control flow diagram of LDI algorithm

EDC mechanism is executed as the beginning of scheduling event, whereby metric weights of all K on the RB $i \in N$ are determined and stored in $M[j][i]$. Although there are different formulas to handle various flows QoS, $\overline{D_{HOL}}$ is calculated only once for all the available flows at the current TTI to determine $M[j][i]$. So, this produces imposes $O(1)$ time complexity. After that, flows are compared with each other to select the flow $j \in K$

with the highest metric $M[j][i]$. The possible overhead, in this case, is $O(N \log K)$ time. As a result, the overhead complexity to execute EDC in a *TTI* is

$$\begin{aligned} T_{EDC} &= O(N \log K).O(1) \\ &= O(N \log K) \end{aligned} \tag{17}$$

On the other hand, LDI mechanism is invoked simultaneously with EDC hence the set of flows (L) with low metric weight is accounted. Then flow j is compared with a flow l, such that $\{j, l\} \in L$ and $j \neq k$ to decide either a scheduling or a dropping event on j. Similar to EDC, in LDI one flow l is only selected from the set L and compared with the flow j to conclude the scheduling decision. This means that the computational overhead produces $O(\log L)$ time. Besides, note that Eqs. (14) and (15) are calculated once each time LDI mechanism is triggered. This adds an overhead of $O(1)$ time. Eventually, computational overhead on LDI mechanism is expressed as,

$$\begin{aligned} T_{LDI} &= O(\log L).O(1) \\ &= O(\log L) \end{aligned} \tag{18}$$

It is important to remark that, by performing EDC, there should be at least 1 flow $j \in K$ with the highest metric weight, i.e $L \subset K$, hence the size of L is always less than K's size. This indicates that the computational overhead for $O(\log L)$ is less than $O(\log K)$ which makes LDI contributes with a minor overhead compared to EDC. Nevertheless, in both mechanisms, the overhead is shown to be limited by a logarithmic scale. By substituting Eqs. (17) and (18) in Eq. (16), the computational overhead of DQAS in a *TTI* duration is obtained as,

$$T_{DQAS} = O(N \log K).O(\log L) \tag{19}$$

Based on the above complexity analysis, it is obvious that DQAS has a minor overhead effect on the overall scheduling process that is limited between logarithmic to linear behavior in big$-O$ notation. This enables DQAS to be a possible MAC layer scheduler that can be implemented in eNB within real network scenarios.

5 Simulation Experiments

To reveal the effectiveness of DQAS as a downlink MAC scheduler, a performance evaluation using simulation experiments is carried out with respect to recent and standard scheduling strategies that are designed for different RT and NRT traffic. The involved scheduling algorithms are PPM [11], EXP-Rule [7], and EXP/PF [23]. To be aligned with the main objective of the proposed work, the performance evaluation deliberately focuses on the capability of the scheduler to reduce the end–to–end delay. This indeed should be linked with maintaining a reasonable QoS level, for example, high throughput and low data drop ratio on different flows types.

5.1 Scenario Configuration

In this work, the LTE downlink system model depicted in Fig. 6 is considered for the simulation scenario. A single macro–cell eNB[4] is deployed at the center point of the LTE cell area. UEs are created such that a direct connection is maintained with the eNB. Moreover, the UEs are randomly distributed within the eNB transmission range. UE mobility is considered in this scenario, where a random motion within the eNB range involving pedestrian and vehicular speeds of 3 and 120 km/h, respectively, is designed and modeled using "Random Direction" mobility.

The experiments are carried out using an object–oriented and open–source system level simulation tool namely "LTE–Sim" [34]. In fact, LTE–Sim is an appropriate and detailed framework tool that models the whole LTE protocol stack, with more concentration on MAC layer functions. Whereby, it supports resources allocation over both time and frequency domains. Further description of other important simulation parameters is included in Table 2.

On the other hand, the physical layer at the downlink channel is modeled with carrier frequency band of 2.1 GHz. It contains a number of sub-carriers with 15 kHz spacing for each. The power transmission rate at eNB is configured to 43 dBm and it is equally distributed among subchannels. To cope with LTE standards, in this work propagation loss of the channel is implemented on a macro–cell urban area model in which it operates by combining four different modules (multi–path, shadowing, path loss, and penetration). The path loss is calculated based on [36] as following,

$$\rho_L = 128.1 + 37.6 \log d \tag{20}$$

where d is the distance in meters between eNB and UE. In multipath module, Rayleigh fast fading is implemented using Jakes' model [37], and a multiple paths number is uniformly selected from the set {6, 8, 10, 12}. In addition, the penetration loss is set to 10 dB while shadowing is modeled by log–normal distribution (standard deviation = 8dB, with mean of 0dB) according to [34].

5.2 Traffic Models

During the simulation scenario, three traffic types (RT Video, RT VoIP, and NRT application) are involved. The load of these traffic sources is imposed to the network in such a way that 40% of UEs are using RT Video, 40% of UEs are using VoIP, and the rest of 20% are using NRT application. The RT Video application is implemented

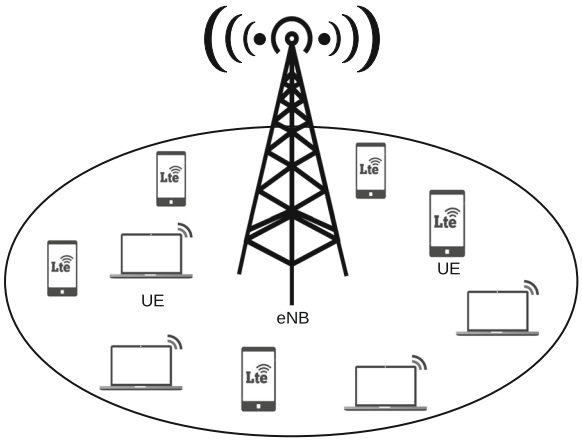

Fig. 6 Simulation system model topology

by a trace–based generator. It sends packets based on realistic trace files that are available in [38]. Video data sequences are encoded using H.264 standard to generate VBR streams at a rate of 242 Kbps. On the other hand, VoIP application provides RT and lightweight flows that use the voice type of G.729 which has been declared as an ITU standard [39]. VoIP is normally modeled by ON/OFF Markov chain. The ON period is exponentially distributed with a mean value of 4 sec, whereas the OFF period has an abbreviated exponential Probability Density Function (PDF) with an upper boundary of 6.9 sec and an average value of 3 sec [40]. During the ON period, the application source transmits packets of 20 bytes size every 20 ms. While for the OFF period, no transmission occurs assuming the presence of voice activity detector. Finally, for NRT application (i.e. buffered video streams), traffic flows are generated in a constant bit rate where packet size and their inter–arrival time are fixed to return a data rate of 20 Kbps.

Table 2 Descriptions of simulation parameters

Parameter	Description
Bandwidth	10 MHz (50 PRBs per TTI)
eNBs in cell	1 eNB
Simulation time	120000 ms
Max delay bound	100 ms
Frame structure	FDD
eNB transmission radius	1 km
PRBs allocation time	1 ms
UEs applications rates	242 kbps video, 9 kbps VoIP, and 20 kbps NRT
MCSs	QPSK, 16QAM, 64QAM
QoS parameters for radio bearer	Default in LTE-Sim QoSParameters object
RLC ARQ of UEs	Activated with max 5 retransmissions
Number of UEs	10-100 with period of 10 UEs

6 Numerical Results and Discussions

During the performance evaluation discussion, the central concentration is on the effectiveness of DQAS in minimizing latency when transmitting UE flows. Besides, robustness in maintaining a good QoS level at high network loads is another potential aspect to be examined. With this evaluation criteria, a closer insight can be revealed on the proposed scheme applicability for real implementation in mobile systems.

6.1 Performance Evaluation Results

Figures 7, 8, and 9 present the end–to–end delay D_{E2E} results. The delay in this scenario is expressed as the time duration starts when a flow is generated by a traffic source, processed by all LTE protocol layers, and transmitted through the channel until it reaches the application layer of the UE. D_{E2E} can be calculated as,

$$D_{E2E} = D_{queue} + D_{prop} + D_{trans} \tag{21}$$

Whereby, D_{queue} is the queue delay at the MAC/RLC layer and normally this part has the dominant impact on D_{E2E}, especially, on burst traffic. Besides, D_{prop}, D_{trans} are the propagation delay captured at the physical layer and the transmission delay caused by the wireless medium between eNB and UE, respectively.

In Fig. 7, the average D_{E2E} is presented for RT VoIP flows. DQAS has a steady pattern of low delay over the increased load comparing with PPM and EXP-Rule. This

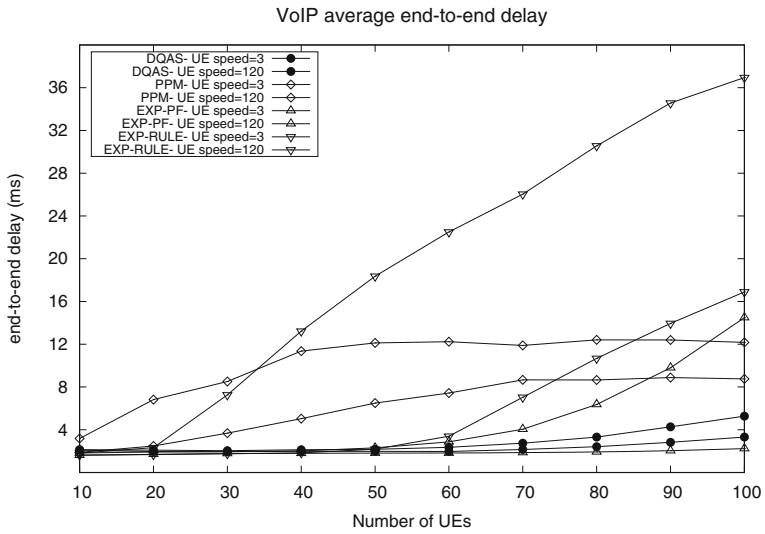

Fig. 7 Average D_{E2E} on RT VoIP flows

improvement is imposed by the effective delay–based decision formulated in EDC mechanism. In details, the metric weight formula allows VoIP flows to be prioritized as long as the buffer size and channel quality condition are relatively high. A slight increase in delay is demonstrated by DQAS against EXP-PF in case of low UE speed (3 km/h), a reduced D_{E2E} pattern of 43.62% lower than EXP–PF is still maintained by DQAS when the cell has more than 50 UEs in high mobility.

Although EXP–PF plots an acceptable D_{E2E} behavior at the low UE mobility compared to EXP–Rule, D_{E2E} is exponentially increased during high UE mobility. This is because both schemes employ a relaxed delay threshold (D_{max}) which restricts high data delivery in congestion stages. The situation seems relatively better in case of PPM, as it keeps a steady delay when more UEs involve the network. At normal loads, VoIP flows suffer a service shortage as PPM emphasizes on flows with heavy buffers.

The results in Fig. 8 for RT VBR Video flows uphold the above discussions in each algorithm's attitude to restrict delay. The results interestingly exhibit a significant delay reduction by DQAS with respect to other scheduling schemes over both UE mobility levels. As a matter of fact, due to the high variability and density of this application,

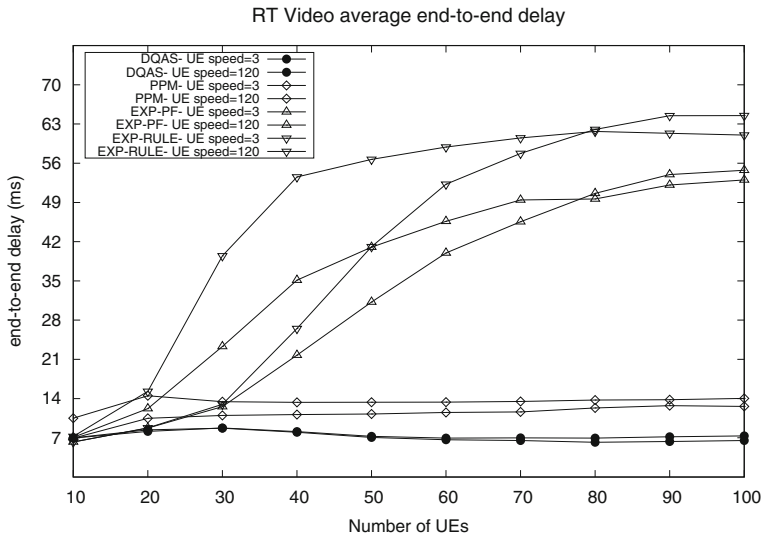

Fig. 8 Average D_{E2E} on RT VBR Video flows

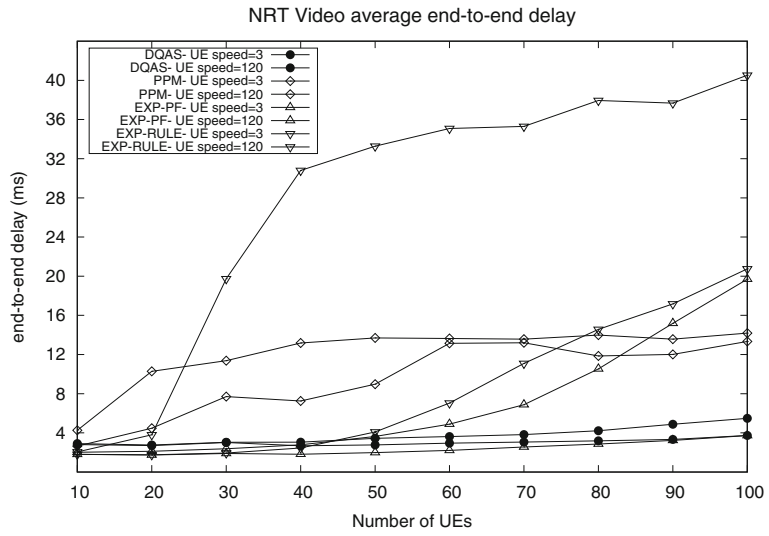

Fig. 9 Average D_{E2E} on NRT Video traffic

preserving low D_{E2E} is a tedious task for most of MAC scheduling schemes. Even though PPM plots steady delay pattern, DQAS succeeds in maintaining a lower delay over the presented load states. So, it is remarkable that DQAS reduces the D_{E2E} up to 48.5% compared to PPM when a 100 UEs are involved in low mobility. This is due to the dynamic delay control threshold, $\overline{D_{HOL}}$, that enforces flows to be processed with low buffer delays. Conversely, a high level of D_{E2E} is seen by EXP–PF and EXP–Rule, hence both emphasize on UEs queues balancing where delay is compromised for keeping an acceptable QoS level.

Flows of NRT application tolerate more marginal delay budgets while seeking a high achievable data rate and less traffic loss. DQAS exhibits a steady and reasonably low delay pattern for NRT video flows as shown in Fig. 9. It benefits from LDI mechanism in which flows can be scheduled if they impose a limited impact on delay. Unlike the deteriorated delay behavior of EXP–PF and EXP–Rule during high UE mobility, DQAS shows a robust trend of delay as low as 54.1% against EXP-PF with a cell load beyond 50 UEs.

The discussion of DQAS performance is also extended over the measured goodput. According to [11], goodput is defined as the amount of successfully transmitted useful bits belong a certain traffic source and can be utilized at the application layer of the UE side. Goodput is then calculated as,

$$goodput = \frac{\sum_{j=0}^{K} \sum_{n=0}^{N} size_{n,j}}{flow_j(t)} \tag{22}$$

Whereby, $size_{n,j}$ is the size of packet n belongs to flow j in bits. $flow_j(t)$ is the time duration that j is active.

Results in Fig. 10 show an outstanding trend of DQAS on RT VBR video goodput. In general, QoS of RT video essentially depends on guaranteeing high data rate for the UE [6]. DQAS succeeded to scale up goodput rate by utilizing LDI mechanism wherein flows are scheduled based on their data rates. Besides, the metric weight for video flows in EDC proportionally increases by the channel data rate. This allows DQAS to dramatically increase goodput up to 33.7% better than PPM when the cell is fully loaded with 100 UEs move in 3km/h. In addition, PPM is seen to achieve a high goodput trend compared with EXP-PF and EXP-Rule since it considers throughput maximization among its procedures. Both EXP-PF and EXP-Rule demonstrate a limited goodput level compared to PPM at high UE mobility. This is because both schemes attempt to provide a balanced service for other types of traffic by the adopted rule of PF.

For NRT video traffic, although EXP-Rule is proposed as a throughput–optimal [7], DQAS outperforms it by 5.8% when 100 UEs with pedestrian mobility is connected to the eNB as shown in Fig. 11. Moreover, in high UE mobility (120km/h), DQAS still harvests the most consistent and outperforming goodput for NRT video traffic. These improvements are tightly related to the inhabited channel–awareness scheduling rule for NRT flows EDC mechanism. Besides, these flows benefit from the LDI mechanism in a way to compensate the lack of data rate for cell–edge UEs[5] in order to ameliorate the overall QoS for NRT flows. Although PPM returns a high level of goodput on RT video, NRT video does not seem a preferable application by this scheme. This is because the scheduling decision in PPM is tightly related to handling the most

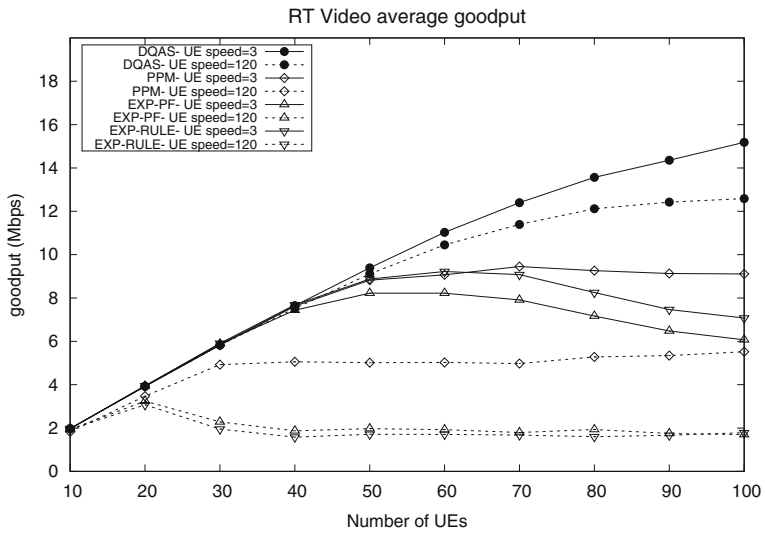

Fig. 10 Average goodput on RT VBR Video traffic

burst traffic, and thus other traffic types may suffer limited data delivery rate.

In general, being a lightweight traffic, RT VoIP flows add no burden to be transmitted by MAC scheduler. This can be obviously seen from the measured goodput results of VoIP traffic in Fig. 12. In DQAS, the priority metric rule for VoIP flows (by Eq. 2) considers UE data rate in a way to provide a linear increase on the achieved throughput. If for instance a flow is transmitted to a high CQI UE, the scheduling decision mostly turns as a channel–aware behavior. Besides, LDI contributes to throughput improvement by scheduling flows that possess a reasonably low buffer delay and decent channel data rate. As stated above, in PPM, traffic is always assumed to come in burst and thus data loss may occur frequently. This makes VoIP flows to suffer from poor delivery rate since they are of small buffers and most of RBs are assigned to burst RT video flows. EXP–Rule and EXP–PF on the other side attempt to keep a balanced service for different flows so that small RT flows can be transmitted by compromising a certain level of throughput on RT burst video flows.

Furthermore, packet dropping ratio is measured on different traffic types. By the time NRT Video flows tolerate a certain level of data drop, VoIP has a crucial limit on this parameter hence excessive dropping severely degrades the call quality. In this work, data dropping is calculated at

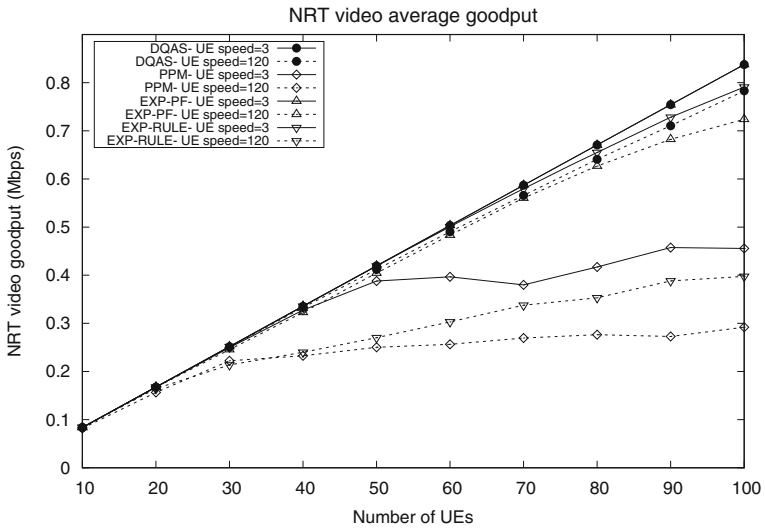

Fig. 11 Average goodput on NRT Video traffic

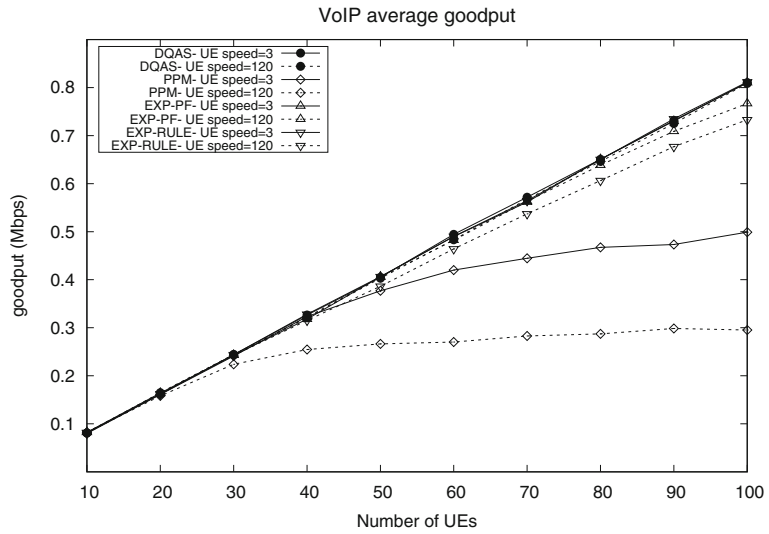

Fig. 12 Average goodput on RT VoIP traffic

both RLC and MAC layers as a ratio of the total transmitted amount of data. Maintaining low data drop ratio reveals the robustness of the scheduler to react toward high overload network states.

Figure 13 demonstrates the results of data drop ratio on VoIP flows. It is obvious that DQAS maintains a very low percentage of data dropping. Basically, for MAC scheduler, keeping a low data dropping ratio is strictly related to the amount of achievable data rate. Therefore, referring to the discussed results of goodput on VoIP flows, DQAS and the reference schemes, except PPM, have a very minor dropping ratio, since the majority of data flows are scheduled for transmission. For PPM, the situation appears very negative on VoIP QoS in general. PPM employs an aggressive drop–based mechanism to punish all flows that violate the defined delay threshold in order to alleviate queuing deadlock. This consequently results in an excessive dropping with the increased load phases.

Drop ratio results for RT Video flows exhibit varied trends by the scheduling schemes as in Fig. 14. Benefiting from the different basis of scheduling decisions in DQAS, majority of the flows are transmitted before violating the delay bound. With that, DQAS succeeds to keep a low data dropping ratio to 54.02% less than EXP–Rule for a range of 40–100 connected UEs and moving in 3 km/h.

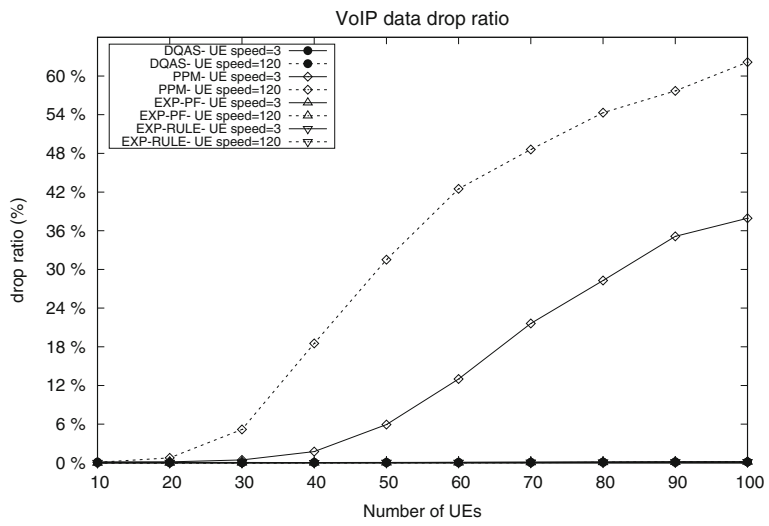

Fig. 13 Data drop ratio on RT VoIP traffic

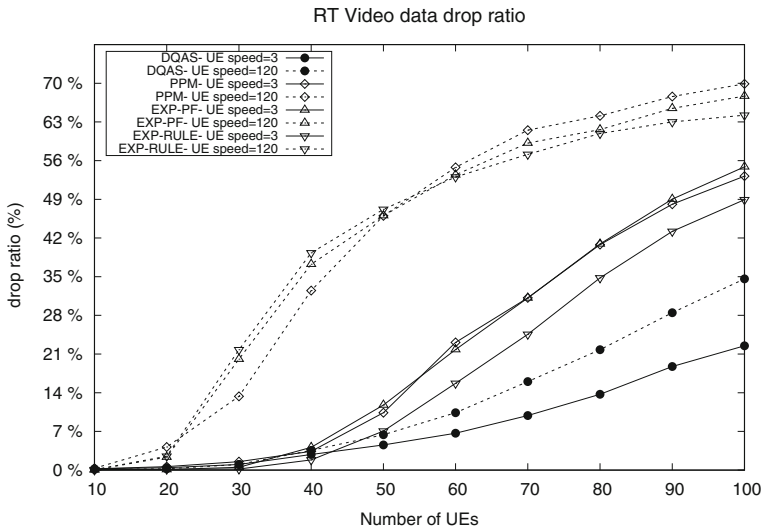

Fig. 14 Data drop ratio on RT VBR Video traffic

Moreover, unlike DQAS, at the high user mobility case, flows with low priority metric are immediately dropped in EXP–PF and EXP–Rule schemes. In PPM, the principle of delay control is mainly based on dropping flows that do not comply with a hard delay threshold.

In DQAS, scheduling NRT flows is greatly based on throughput maximization. Although RT traffic is given more priority since the D_{E2E} is considered as key criteria of scheduling, DQAS still influences the MAC queue to schedule more NRT flows so as to return a low dropping ratio as presented in Fig. 15. In addition, it is interesting to observe that at high UE mobility case, DQAS is able to sustain a minor drop ratio pattern that is independent of the incremented offered load.

From the narrated performance discussion above, the significance of DQAS in figuring out the delay problem is highlighted with respect to the involved reference schemes as follows:

- The performance results for both of PPM and EXP-Rule show QoS improvement in one dimension (throughput) for RT burst traffic. Although delay is considered in both schemes, PPM follows an aggressive drop procedure to flush out long flows that violate its defined delay threshold. This negatively impacts the delay on small flows. EXP-Rule, however, shows a reasonable multi–class QoS level by balancing throughput of different flows. Nonetheless, D_{E2E} for RT traffic is still not desirable as it proportionally increases by the offered load.
- EXP-PF illustrates an acceptable overall performance compared with the other reference schemes. It attempts to balance between delay–sensitivity and throughput properties to serve different QoS needs. This bond is nonetheless compromised during high network load and UE mobility. EXP-PF adopts a dropping principle to overcome the delayed transmissions on burst traffic to favor small and delay-sensitive flows.
- DQAS significantly succeeds in maintaining a tight balance between delay and QoS for multi-traffic types. The numerical results show that DQAS operates as a robust MAC scheduler toward different network loads and under user mobility constraints. The central principle in DQAS that different flows are scheduled by emphasizing on their standardized QoS indices. Therein D_{E2E} on different flows is ensured in low values by using EDC mechanism, while a high level of goodput, as well as low dropping ratio, is obtained by LDI mechanism.

6.2 Impact of EDC QoS-related parameters on the overall performance

In this part, the impact of QoS–related parameters in EDC mechanism, i.e. α is discussed on the obtained results. The results below are aggregated from the three traffic types involved and sampled over a total number of 20 UEs. RT Video, VoIP, and NRT video applications are distributed complying to 2:2:1 relation, respectively. These results are generated based on a simulation scenario involving EDC mechanism.

In Figs. 16 and 17, end–to–end delay and drop ratio, respectively, are measured against different configurations of QoS parameters, i.e. α. In this scenario, given a fixed D_{max}, i.e. 0.1 for all the flows while several values of σ, the pattern of α is obtained as expressed in Eq. 3. From the presented results, it can be seen that higher values of σ, i.e. $\sigma > 0.5$ makes the scheduler decision more

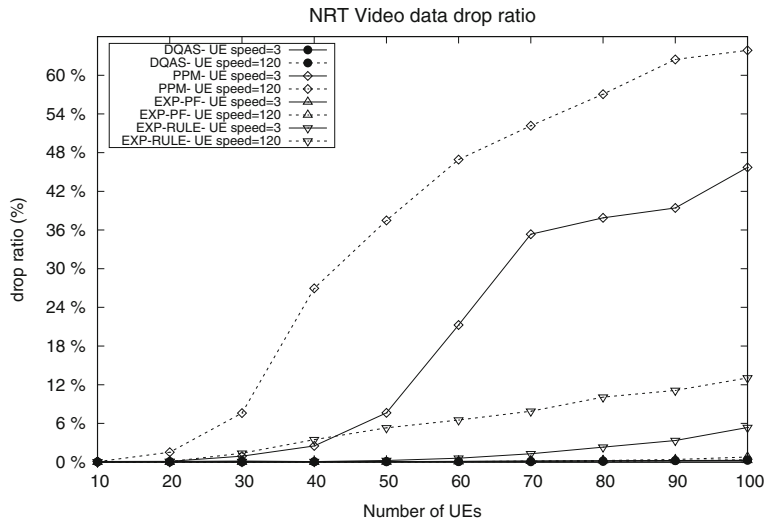

Fig. 15 Data drop ratio on NRT Video traffic

permissive toward ensuring low delay as shown in Fig. 16. Hence, in this case, the endeavor is to enforce an equable delay distribution among the different users' queues. This lenient delay behavior (when choosing high values of σ) precludes further reduction of dropping ratio, particularly in scenarios where multiple traffic of different QoS profiles are transmitted as presented in Fig. 17. Therefore, based on the tailored discussions and results above, we believe that assigning σ to a relatively low value ($\sigma = 0.35$) returns more reasonable and robust performance level of the scheduling decision to handle multi-traffic scenarios, especially for RT flows.

7 Conclusion

In this article, DQAS scheme was introduced as a MAC scheduler for the downlink LTE channel. With the aim of transmitting RT traffic in low latency and high throughput, in DQAS, the problem of QoS guarantee from different dimensions was the main emphasis. In the first part of DQAS, EDC mechanism was developed to determine priority metric weights for different flows to fulfill their delay needs. Besides, the LDI mechanism is adapted to schedule flows with minor impact on delay. These mechanisms enable DQAS to maintain a good balance between delay and improved the application throughput. On the

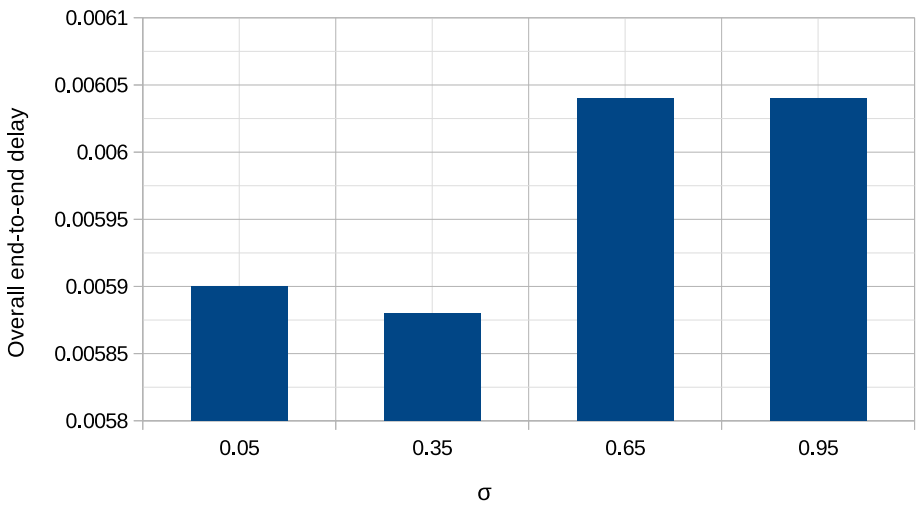

Fig. 16 Overall system D_{E2E} ratio based on different σ values

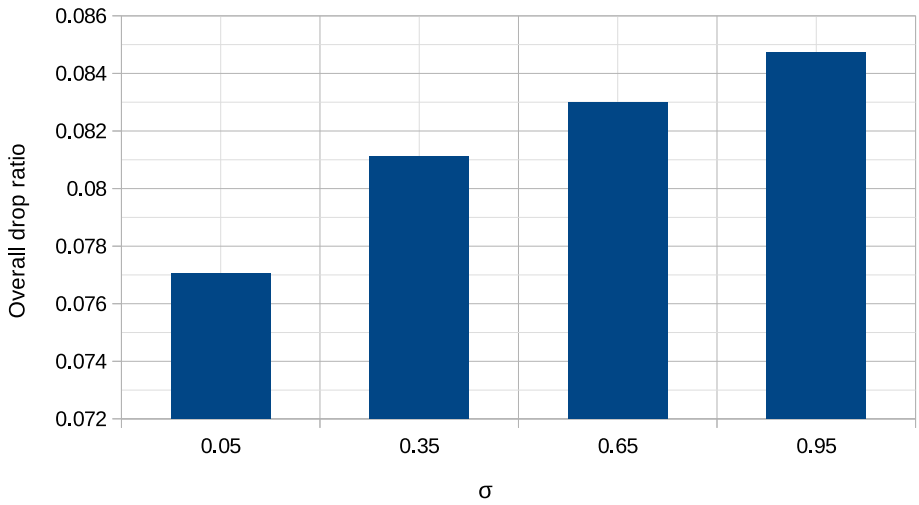

Fig. 17 Overall system drop ratio based on different σ values

other hand, analysis of the overhead complexity revealed that DQAS has a lightweight operation in MAC scheduler which makes it a possible solution in real mobile network scenarios. Performance evaluation of DQAS against other existing schedulers and over different UE mobility scenarios was presented over a 4G LTE network scenario. It is evident that DQAS sustained a robust and low behavior of delay on RT traffic that is independent of the grown network load. Besides, it returns the highest amount of goodput with a low fraction of data drop. These interesting results encourage further endeavors to possibly implement DQAS scheme in 5G multicell environments and considering traffic applications which further serve IoT scenarios.

Endnotes

[1] Features such as channel–awareness, low complexity, and fair service level.

[2] RT traffic with small flows, i.e. VoIP application.

[3] Possible dropping on RLC might be due to detected errors or exceeded retransmissions of certain packets.

[4] With respect to the practicality of multi-cell scenarios, the single–cell scenario allows obtaining a straightforward evaluation of the behavior of the adopted scheduling scheme with a realistic QoS (especially, in types of macrocell base stations) which fulfills the main objective of this work.

[5] UEs with a low channel quality due to multiuser diversity in the wireless environment.

Abbreviations

4G: Fourth Generation; 5G: Fifth Generation; AMC: Adaptive Modulation and Coding; AQM: Active Queue Management; CBR: Constant Bit Rate; CDMA: Code Division Multiple Access; CQI: Channel Quality Indicator; DQAS: Delay-based and QoS-Aware Scheduling; DRX: Discontinuous Reception; EDC: Efficient Delay Control; eNB: evolved NodeB; EXP-PF: EXPonential-PF; EXP-Rule: EXPonential-Rule; FD: Frequency Domain; FDD: Frequency Division Duplex; HoL: Head of Line; ITU: International Telecommunication Union; LDI: Least Delay Increase; Log-Rule: Logarithmic-Rule; LTE: Long Term Evolution; MAC: Medium Access Control; MCS: Modulation and Coding Scheme; MDP: Markov Decision Process; M-LWDF: Modified-Largest Delay First; MT: Maximum Throughput; NRT: Non-Real Time; OFDMA: Orthogonal Frequency-Division Multiple Access; PDCCH: Physical Downlink Control Channel; PDF: Probability Density Function; PDSCH: Physical Downlink Shared Channel; PF: Proportional Fairness; PPM: Packet Prediction Mechanism; PRB: Physical Resource Block; QoS: Quality of Service; RB: Resource Block; RLC: Radio Link Control; RRM: Radio Resource Management; RT: Real Time; SINR: Signal-Interference-plus-Noise-Ratio; TBS: Transport Block Size; TCP: Transmission Control Protocol; TD: Time Domain; TTI: Time Transmission Interval; UE: User Equipment; VBR: Variable Bit Rate; VoIP: Voice-over-Internet Protocol

Acknowledgements

This work has been partially supported by the Malaysian Ministry of Education under the Fundamental Research Grant funding ID UPM-FRGS-08-02-13-1364FR for financial support.

Availability of data and materials

The source code as well as datasets generated and analyses during the current study are not publicly available as they are being considered for further study and deployments; however, inquiring about illustrations of the presented work is possible by contacting the corresponding authors.

Authors' contributions

NKM conceived and designed the study and then performed the experiments and wrote the paper. NKM and ZMH reviewed and edited the manuscript. ZMH, MO, and SS supervised the study and approved the final manuscript.

Competing interests

The authors declare no conflict of interest. The funding sponsors has no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Publisher's Note

Author details

[1] Department of Communication Technology and Networks, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia. [2] Computational Science and Mathematical Physics Lab, Institute of Mathematical Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia.

Received: 10 January 2018 Accepted: 19 June 2018
Published online: 17 July 2018

References

1. D Astély, E Dahlman, A Furuskär, Y Jading, M Lindström, S Parkvall, LTE: the evolution of mobile broadband. IEEE Commun. Mag. **47**(4), 44–51 (2009)
2. P Kela, J Puttonen, N Kolehmainen, T Ristaniemi, T Henttonen, M Moisio, in *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium On*. Dynamic packet scheduling performance in utra long term evolution downlink (IEEE, Santorini, 2008), pp. 308–313
3. S Schwarz, C Mehlführer, M Rupp, in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference On*. Low complexity approximate maximum throughput scheduling for LTE (IEEE, Pacific Grove, 2010), pp. 1563–1569
4. R Kwan, C Leung, J Zhang, Multiuser scheduling on the downlink of an LTE cellular system. Res. Lett. Commun. **2008**, 3 (2008)
5. F Kelly, Charging and rate control for elastic traffic. Eur. Trans. Telecommun. **8**(1), 33–37 (1997)
6. ETSI, Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Quality of Service (QoS) concept and architecture. European Telecommunications Standards Institute. Ts 123 107 v12.0.0 (2014)
7. B Sadiq, R Madan, A Sampath, Downlink scheduling for multiclass traffic in LTE. EURASIP J. Wirel. Commun. Netw. **2009**(1), 1–18 (2009)
8. M Andrews, K Kumaran, K Ramanan, A Stolyar, R Vijayakumar, P Whiting, CDMA data QoS scheduling on the forward link with variable channel conditions. Bell Labs Technical Memorandum Report (2000)
9. KM Elsayed, AK Khattab, Channel-aware earliest deadline due fair scheduling for wireless multimedia networks. Wirel. Pers. Commun. **38**(2), 233–252 (2006)
10. M Brehm, R Prakash, Overload-state downlink resource allocation in LTE MAC layer. Wirel. Netw. **19**(5), 913–931 (2013)
11. WK Lai, C-L Tang, QoS-aware downlink packet scheduling for LTE networks. Comput. Netw. **57**(7), 1689–1698 (2013)
12. S Ali, M Zeeshan, A Naveed, A capacity and minimum guarantee-based service class-oriented scheduler for LTE networks. EURASIP J. Wirel. Commun. Netw. **2013**(1), 67 (2013)
13. G Wunder, P Jung, M Kasparick, T Wild, F Schaich, Y Chen, S Ten Brink, I Gaspar, N Michailow, A Festag, *et al*, 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. IEEE Commun. Mag. **52**(2), 97–105 (2014)
14. C Wengerter, J Ohlhorst, AGE von Elbwart, in *Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st, vol. 3*. Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA (IEEE, Stockholm, 2005), pp. 1903–1907
15. X Li, B Li, B Lan, M Huang, G Yu, in *Information and Communication Technology Convergence (ICTC), 2010 International Conference On*. Adaptive pf scheduling algorithm in LTE cellular system (IEEE, Jeju, 2010), pp. 501–504
16. H Ahmed, K Jagannathan, S Bhashyam, in *Global Communications Conference (GLOBECOM), 2013 IEEE*. Queue-aware optimal resource allocation for the LTE downlink (IEEE, Atlanta, 2013), pp. 4122–4128
17. MM Nasralla, MG Martini, in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium On*. A downlink scheduling approach for balancing QoS in LTE wireless networks (IEEE, London, 2013), pp. 1571–1575
18. J Xu, C Guo, H Zhang, J Yang, Resource allocation for real-time traffic in unreliable wireless cellular networks. Wirel. Netw. **24**(5), 1–14 (2016)
19. HR Chayon, K Dimyati, H Ramiah, AW Reza, An Improved Radio Resource Management with Carrier Aggregation in LTE Advanced. Appl. Sci. **7**(4), 394 (2017)
20. S Liu, C Zhang, Y Zhou, Y Zhang, Delay-Based Weighted Proportional Fair Algorithm for LTE Downlink Packet Scheduling. Wirel. Pers. Commun. **82**(3), 1955–1965 (2015)
21. K Kaewmongkol, A Jansang, A Phonphoem, in *Computer Science and Engineering Conference (ICSEC), 2015 International*. Delay-aware with resource block management scheduling algorithm in LTE (IEEE, Chiang Mai, 2015), pp. 1–6
22. C Wang, Y-C Huang, Delay-scheduler coupled throughput-fairness resource allocation algorithm in the long-term evolution wireless networks. IET Commun. **8**(17), 3105–3112 (2014)
23. S Shakkottai, AL Stolyar, Scheduling algorithms for a mixture of real-time and non-real-time data in HDR. Teletraffic Sci. Eng. **4**, 793–804 (2001)
24. SJ Bae, B-G Choi, MY Chung, in *Communications (APCC), 2011 17th Asia-Pacific Conference On*. Delay-aware packet scheduling algorithm for multiple traffic classes in 3GPP LTE system (IEEE, Sabah, 2011), pp. 33–37
25. HR Chayon, KB Dimyati, H Ramiah, AW Reza, Enhanced Quality of Service of Cell-Edge User by Extending Modified Largest Weighted Delay First Algorithm in LTE Networks. Symmetry. **9**(6), 81 (2017)
26. B Sadiq, SJ Baek, G De Veciana, Delay-optimal opportunistic scheduling and approximations: the log rule. IEEE/ACM Trans. Networking (TON). **19**(2), 405–418 (2011)
27. AK Paul, H Kawakami, A Tachibana, T Hasegawa, Effect of AQM-Based RLC Buffer Management on the eNB Scheduling Algorithm in LTE Network. Technologies. **5**(3), 59 (2017)
28. H Bo, T Hui, C Lan, Z Jianchi, DRX-aware scheduling method for delay-sensitive traffic. IEEE Commun. Lett. **14**(12), 1113–1115 (2010)
29. J-M Liang, J-J Chen, H-H Cheng, Y-C Tseng, An energy-efficient sleep scheduling with QoS consideration in 3GPP LTE-advanced networks for internet of things. IEEE J. Emerg. Sel. Top. Circ. Syst. **3**(1), 13–22 (2013)
30. Y Li, KK Chai, Y Chen, J Loo, Duty cycle control with joint optimisation of delay and energy efficiency for capillary machine-to-machine networks in 5G communication system. Trans. Emerg. Telecommun. Technol. **26**(1), 56–69 (2015)
31. NK Madi, Z Mohd Hanapi, M Othman, S Subramaniam, On multi-cell packet scheduling of LTE-a cellular networks: a survey of concepts related challenges and solutions. J. Appl. Sci. **14**(20), 2422–2438 (2014)
32. ETSI, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis. Tr136 921 v10.0.0 (2011)
33. M Masum, M Babu, End-to-End Delay Performance Evaluation for VoIP in the LTE network. Master of science thesis, Blekinge Institute of Technology, School of Engineering (2011). http://urn.kb.se/resolve?urn=urn:nbn:se:bth-4264
34. G Piro, LA Grieco, G Boggia, F Capozzi, P Camarda, Simulating LTE cellular systems: an open-source framework. IEEE Trans. Veh. Technol. **60**(2), 498–513 (2011)
35. A Mecozzi, R-J Essiambre, Nonlinear Shannon limit in pseudolinear coherent systems. J. Light. Technol. **30**(12), 2011–2024 (2012)
36. 3GPP, 3rd generation partnership project; technical specification group radio access network; fdd base station (bs) classification (release 10). Tech. Specification. **10.0.0**(3GPP TR 25.951), 1–63 (2011)
37. WC Jakes, DC Cox, *Microwave Mobile Communications*. (Wiley-IEEE Press, New York, 1994)
38. P Seeling, M Reisslein, Video transport evaluation with H. 264 video traces. IEEE Commun. Surv. Tutorials. **14**(4), 1142–1165 (2012)
39. R Salami, C Laflamme, J-P Adoul, A Kataoka, S Hayashi, T Moriya, C Lamblin, D Massaloux, S Proust, P Kroon, *et al*, Design and description of CS-ACELP: a toll quality 8 kb/s speech coder. IEEE Trans. Speech Audio Process. **6**(2), 116–130 (1998)
40. C-N Chuah, RH Katz, in *Communications, 2002. ICC 2002. IEEE International Conference On, vol. 2*. Characterizing packet audio streams from internet multimedia applications (IEEE, New York, 2002), pp. 1199–1203