Scheduling Algorithm with Delay-limited for VoIP in LTE

Juan Chen [#], Wenguo Yang [#], Suixiang Gao [#], Lei Zhou [*]

[#] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China
Chenjuan12@mails.ucas.ac.cn; yangwg@ucas.ac.cn; sxgao@ucas.ac.cn
[*] Huawei Technologies Co. LTD. Beijing, China
kevin.zhoulei@huawei.com

Abstract — **The long term evolution (LTE) is a breakthrough technology with respect to the previous generation of cellular networks, as it is based on an all-IP architecture that aims at supporting several high quality services, such as video streaming, voice-over-internet-protocol (VoIP) and everything related to the wideband Internet access. In the recent years, VOIP service is becoming more and more popular and important, while the tight delay requirements and the scarcity of radio resources problem remain the main challenges for it. To solve these problems, we construct a mathematical model and design a novel algorithm for the scheduling problem of VOIP. Limiting the delay, our algorithm solve the scarcity of radio resources problem by minimizing the total number of radio resources scheduled during a certain period of time, which can ensure the quality-of-experience (QOE) of user equipment (UE) and high resource utilization rate. Moreover, the maximum delay of a UE with a certain average transmitting rate is determined theoretically. Finally, we give simulation results of our scheduling algorithm and the comparison with the modified largest weighted delay first (M-LWDF) and the exponential/proportional fair (EXP/PF) algorithms, by which we demonstrate the effectiveness of our algorithm on delay and resource utilization.**

Index Terms — **delay-limited, radio resource, scheduling, VoIP, LTE**

I. INTRODUCTION

The LTE system, which is designed based on the orthogonal frequency division multiplexing (OFDM) to provide services including real-time multimedia services in the IP-based PS domain, represents a very promising answer to the ever rising bandwidth demand of mobile applications and enables diverse mobile multimedia service provision.

The continuous rise of real-time multimedia services (such as voice and video service) over the Internet, and the need for ubiquitous access to them, are driving the evolution of cellular networks. Such services usually exhibit a large variety of quality-of-service (QoS) requirements, such as transmission rate, delay, and packet-dropping ratio, which are difficult to satisfy in a wireless environment due to the limit of radio resources, the variability of the channel condition, and resource contention among multiple UEs. Although the latest mobile broadband technologies, including 3GPP technologies based on LTE and LTE-advanced standards, offer much higher capacities with peak rates up to 340 Mb/s over a 20 MHz channel [1, 2], the successful implementation of multimedia services in the PS domain is still a challenge when loss sensitive real-time multimedia service characteristics are taken into account, especially in the wireless environment with fading and delay.

Despite the increasing demand for packet-data based communication, the voice communication will always play an important role in wireless networks. In the existing CDMA and WCDMA based cellular systems, the voice service is provided separately in the circuit switching (CS) domain. On the other hand, LTE is designed to provide all services including the voice service in the IP-based PS domain. QoS provision for VoIP traffic in the PS domain is difficult because the VoIP service is sensitive to packet delay and loss. Nevertheless, UEs expect higher voice service quality in LTE than in CS domain based voice services in the existing cellular systems. Therefore, it is pivotal to satisfy the QoS requirements of VoIP-based voice services in LTE. In general, the most important objective of a multimedia service is the satisfaction of the end UEs, i.e., their QoE. This is strictly related to the system's ability to provide for the application flows a suitable QoS [3], generally defined in terms of network delivery capacity and resource availability (i.e. there should be limited packet loss ratios and delays).

The delay requirements are tight in the multimedia services. As an example, in real-time multimedia services, end-to-end delay constraints in content delivery have to match the requirements related to the human perception of interactivity. In wireless communication the problem of lost packets attracted much attention. Fortunately, with the introduction of H-ARQ in LTE, the packet loss rate over the air interface has been greatly reduced (this has already been proved using 3GPP system simulations). Besides, the problem of radio-resource scarcity has been an important issue. So, supporting VoIP in LTE encounters the following problems [4]:

(1) The tight delay requirement combined with the frequent arrival of small VoIP packets;

(2) The scarcity of radio resources along with control channel restrictions in LTE.

Resource scheduling plays a key role in both controlling delay and saving resources in LTE. Thus, designing effective scheduling methods to meet the stringent delay requirements of VoIP and solve the radio-resource scarcity problem in LTE is very urgent.

We consider the downlink of VoIP in this paper as there is far more downlink traffic in the system than uplink traffic, which causes the need for downlink optimization to be more urgent.

In this paper, our main contributions are as follows:

(1) A stochastic mathematical model is established with hard packet-delivery deadline constraints. The model is proposed by limiting the delay within a certain deadline to meet the stringent delay requirements of VoIP. It also minimizes the total number of radio resources used in a given period to solve the problem of radio-resource scarcity.

(2) We design an effective delay-limited (DL) algorithm by forecasting and estimating the future wireless channel conditions based on the known actual channel conditions. Moreover, given a certain average transmitting rate, we determine the maximum delay of a UE theoretically.

(3) We simulate the DL scheduling algorithms, in comparison with M-LWDF and EXP/PF algorithms. On this basis, we analyse the effects of DL algorithm. The results demonstrate that the DL algorithm has a beneficial effect on delay and resource utilization.

The remainder of this paper is organized as follows. Section II outlines other work related to scheduling in the VoIP service. Section III describes the downlink resource allocation in LTE, stated as a specific optimization problem, and gives the mathematical model for the problem. Section IV presents the DL algorithm for VoIP service in LTE, and determines theoretically the maximum delay of the UE with a certain average transmission rate. Section V gives the simulation results obtained using the DL algorithm and compares DL algorithm with M-LWDF and EXP/PF algorithms. The final part gives a summarization.

II. RELATED WORK

In the LTE system, resource scheduling for VOIP services is very important and challenging because of the continuous increase in the demand for voice services and the difficulty in satisfying delay and packet-dropping ratio requirements. Many algorithms for VoIP in LTE systems have been designed [5–12]. To date, the scheduling methods used for VoIP in LTE systems can be divided into three types: persistent scheduling, semi-persistent scheduling, and dynamic scheduling.

A. Persistent Scheduling

Persistent scheduling [5, 6] has been proposed to reduce signalling overheads in VoIP services. The idea behind persistent scheduling is to pre-allocate a sequence of frequency-time resources with a fixed MCS to a VoIP UE at the beginning of a specified period. This allocation remains valid until the UE receives another allocation due to a change in channel quality or the expiration of a timer. A big disadvantage of such a scheme is the lack of flexibility in the time domain which may result in a problematic difference between allocated and required resources. To reduce wastage when a VoIP call does not make full use of its allocated resources, two ‘paired’ VoIP UEs can share resources dynamically. Persistent scheduling always divides and allocates the resources to specific groups of UEs, omitting the fading in the communication channel. This blind strategy is neither effective nor fair.

B. Semi-persistent Scheduling

Semi-persistent scheduling [7, 8] is an algorithm which can reduce signalling overheads and increase system capacity. In a semi-persistent scheduling algorithm, only when the UE equipment creates a new voice, can it apply for resources and it always occupies the same resource block (RB) in each 20 *ms*. As VoIP is characterized by small size and frequent and regular arrival of data packets, the semi-persistent scheduling algorithm reduces the radio resources occupied by a large number of scheduling signals. However, as the UE requests in the semi-persistent scheduling method are on a first-come first-served basis, this relatively fair scheduling scheme can lead to taking up too many available radio resources caused by retransmitting data associated with UEs with poor channel quality, thus reducing the overall radio resource utilization.

C. Dynamic scheduling

Dynamic scheduling is a general scheduling method. In radio resource scheduling for each basic RB unit, the scheduler allocates resources based on the resource request by the UE equipment. Application of this scheduling method to VOIP services in LTE causes an increased consumption of radio resources. Typical dynamic scheduling algorithms include the polling scheduling algorithm, the maximum carrier to interference ratio (MAX C/I) scheduling algorithm [9, 10], and the proportional fair (PF) scheduling algorithm [11]. In addition, scholars have proposed improved algorithms for maximum weight delay priority and other enhanced algorithms, which, to some extent, enhance the performance of the dynamic scheduling. Recently, a very thorough discussion of the related work in this field has been reported in [12], and EXP and LOG rules have been presented as the most promising approaches for downlink scheduling in LTE systems with delay-sensitive applications. However, for VOIP in LTE, these improved dynamic scheduling algorithms have a limited scope to optimization business. In particular, the dynamic scheduling algorithms cannot meet the stringent delay requirements of VoIP. Both the time and frequency domains can be optimized using dynamic algorithms, leading to probably ideal schemes for the UE.

In fact, the design of an efficient packet scheduler for a wireless system is a difficult task that typically involves a large number of conflicting requirements. These must be analysed and weighed-up before a balanced solution can be implemented [13]. On the one hand, the scheduler must be efficient in utilizing the available radio resources as the wireless spectrum is the most precious resource in the wireless communication system. On the other hand, services should be fairly scheduled so as to guarantee a certain level of service for UEs with bad channel conditions. Among the various fairness criteria, proportional fairness (PF) scheduling is widely considered as a good solution because it provides an attractive trade-off between the maximum average throughput and UE fairness. It does this by exploiting the temporal diversity and game-theoretic equilibrium in a multi UE environment [14]. A typical PF scheduler only considers the performance of an average system throughput and fairness. Such a PF scheduler usually does not consider other QoS performance metrics, e.g. packet-dropping probability and packet delay. The average delay performance in multi UE OFDM was studied in [15]. However, the delay problem has not been effectively solved. Some scheduling strategies, such as M-LWDF [16] and EXP/PF [17] algorithms which are not originally designed for OFDMA systems, guarantee the delay requirements. Then M-LWDF and EXP/P algorithms have been applied to LTE based on the assumption that a strictly positive probability of discarding packets [18] is acceptable. And they have been demonstrated that they guarantee good throughput, little delay and an acceptable level of fairness. However, unlike in other systems, in LTE, good throughput is not equal to saving resources.

As the tight delay requirement and saving resource are not guaranteed at the same time in most of the current algorithms, we construct a mathematical model and design an algorithm to meet the tight delay requirement and solve the problem of radio-resource scarcity at the same time. This is the main contribution of this paper.

III. PROBLEM STATEMENT AND MATHEMATICAL MODEL

A. Downlink Resource Allocation in LTE

In LTE [19–21], access to the radio spectrum is based on OFDM. The air interface has been designed to use OFDM for the downlink, i.e., from the evolved node B (eNB) to UEs. Radio resources are apportioned into the time/frequency domain. In the time domain, they are distributed every transmission time interval (TTI), each one lasting 1 *ms*. Furthermore, each TTI is composed of two time slots of 0.5 *ms*, corresponding to 7 OFDM symbols in the default configuration with a short cyclic prefix. Ten consecutive TTIs form an LTE frame which lasts 10 *ms*.

In the frequency domain, the total bandwidth is instead divided into180 kHz sub-channels, each with 12 consecutive and equally-spaced OFDM sub-carriers. A time/frequency radio resource, spanning over one time slot of 0.5 *ms* in the time domain and over one sub-channel in the frequency domain, is called a resource block (RB) and corresponds to the smallest radio resource unit that can be assigned to a UE for data transmission. Note that as the sub-channel dimension is fixed, the number of sub-channels varies according to the different system bandwidth configuration. Portions of the spectrum should be distributed every TTI among the UEs. Packet schedulers work in the time and frequency domain with a coarseness of one TTI or one RB, respectively. The fastest scheduling is required to be done within 1 *ms* according to the symbol length of RB.

At the base station, i.e. the eNB, the packet scheduler distributes radio resources among UEs. The scheduling decisions are strictly related to the channel quality experienced by UEs. The downlink resource scheduling, a resource allocation process, refers to the system determining when and what resources are available for the UE to transfer data. The whole process can be divided into the following sequence of operations that are repeated:

(1) The eNB prepares the list of RBs which can be scheduled in the current TTI and then sends it to the UEs.

(2) Each UE decodes the reference signals and reports the channel quality to the eNB which helps to estimate the quality of the downlink channel.

(3) The eNB allocates resources using a scheduling strategy depending on whether the downlink channel quality is good or bad.

(4) The eNB transmits data in the downlink channel according to the allocation results. The UEs receive the data and determine whether to send a retransmission request indication.

(5) The eNB transmits the new or retransmission data.

Now assume that when a new scheduling time tis reached, the first and second operations can be completed. In other words, the channel quality is known at t. The main problem is the third operation. That is, designing an effective scheduling strategy to allocate resources (which is the main purpose of this paper). As we can calculate channel transmission rate according to the channel condition, we use transmission rate instead of channel condition in the following.

B. Problem Restatement

To make the problem easier to solve, the following assumptions are set out:

(1)Each data packet can be transmitted after being separated into many parts.

(2)Each data packet can be transmitted successfully one time. In other words, there are no transmission errors in the system. The lost packet is inserted into the front of the queue and retransmitted when a transmission error happens. This will cause no problem for our algorithm.

VoIP is characterized by the small size, frequent transmission, and regular arrival of data packets. Based on these characteristics, we give another assumption

(3) The time interval of producing packet is already identified and known, denoted by Δt (*ms*).

For Internet telephony, a delay of 100 *ms* is considered as the limit for a good perceived quality, while the delay has to be less than 300 *ms* for satisfactory quality [22]. In Ref. [23], for example, a delay of 200 *ms* is considered for video interactive applications. So, to meet the tight delay requirements of VoIP in the LTE system, we limit the delay of each packet with a certain deadline which is less 200ms according to the delay requirements of each UE.

The problem of radio-resource scarcity has been an important issue the wireless communication facing. In LTE, the radio resource is divided into some RBs in frequency domain and time domain. Moreover, each RB can be only used by one UE. Therefore, to save radio resource for other service, we minimize the total number of radio resources used in a certain period of time.

Based on the assumptions, the problem can be described strictly. There are N UEs which continue to produce data packets, and R available RBs at each TTI to be allocated to UEs to transmit packets. The time interval of producing data packets for each UE is Δt. The rate that one RB transmits packets and the size of the data packets are all random. Each RB can only transmit data packets for one UE. The delay limit of UE n is DL_n *ms*. Now, we need to design a method to schedule R

available RBs for N UEs in time T, and the resources will be scheduled once each TTI. Moreover, the method needs to meet the following three conditions:

(1) The total number of RBs used in T is as little as possible.

(2) Each data packet of UE n must be transmitted successfully within DL_n *ms*.

C. Mathematical Model

We formulate the mathematical model of the above optimization problem as follows:

Symbol explanation:

N: Total number of UEs;

R: Total number of available RBs at each time;

T: Total time;

DL_n : Delay limit of UE n.

$x_{n,i,t,r}$: The proportion of the ith packet accounting at RB when the ith data packet produced by UE n is transmitted at time t by the rth RB;

$y_{n,t,r}$: If the rth RB transmits UE n's data packets at time tit equals to 1, else it equals to 0;

$v_{n,t,r}$: The rate of the rth RB transmitting UE n's data packet at time t;

$c_{n,i}$: The size of the ith data packet produced by UE n;

Mathematical modelI

$$\textbf{Min} \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{R} y_{n,t,r}(\mathbf{v})$$

Such that

$$\sum_{n=1}^{N} y_{n,t,r} \le 1, \quad \forall t,r \quad (1)$$

$$\sum_{t=(i-1)\Delta t}^{(i-1)\Delta t+DL_n} \sum_{r=1}^{R} x_{n,i,t,r} \cdot v_{n,t,r} \ge c_{n,i}, \ \forall n,i \qquad (2)$$

$$\sum_{i} x_{n,i,t,r} \le y_{n,t,r}, \ \forall n,t,r \quad (3)$$

$$x_{n,i,t,r} \in [0,1] \ ; y_{n,t,r} \in \{0,1\} \qquad (4)$$

Where $\boldsymbol{v} = \{v_{n,t,r}\}$ and $\boldsymbol{c} = \{c_{n,i}\}$ are random variables and $n = 1,2,...,N;\ i = 1,2,...;\ t = 0,1,2,...,T;\ r = 1,2,....,R$.

The goal is to minimize the total number of RGBs used in time T using an objective function depending on the random variable $\boldsymbol{v}$. Eq. (1) ensures that each RB can only transmit data packets for one UE. Eq. (2) indicates that the size of UE n's data transmitted in the time interval $[(i-1)\Delta t, (i-1)\Delta t + DL_n]$ $(i = 1,2,...)$ must not be less than the size of the ith data packet of UE n. This ensures that each data packet for UE n must be transmitted within DL_n *ms*. As $\boldsymbol{v}$ and $\boldsymbol{c}$ are random variables, Eq. (2) is, in essence, a random constraint. Eq. (3) indicates that the data transmitted by one RB must be less than the largest amount of data that the RB can transmit. The model has $RNT(2+T/\Delta t)$ variables and $RT + N(1+T/\Delta t) + RNT$ constraints.

Obviously, this is a 0–1 large-scale integer linear random programming problem. Generally speaking, integer linear programming problems are NP-hard. Thus, we cannot solve our problem in polynomial time. Consequently, we propose to use a heuristic algorithm to tackle this problem.

IV. ALGORITHM AND ANALYSIS

A. Delay-limited (DL) Scheduling Algorithm

In part A, we transform the random problem to the deterministic problem by some assumptions. However those assumptions do not always comply with the actual situation. For example, in actual situations, the rate after the scheduled time, and the sizes of data packets are random and not known precisely. So to solve this problem, we have to design an algorithm which can be used in actual settings. In this part, we will design an online algorithm which ensures that every packet of user n can be transmitted successfully within DL_n *ms* at a high resource utilization rate.

From Section III.C, we know that the proposed optimization problem is a large-scale 0–1 integer linear random programming problem, which are NP-hard. Thus, we cannot precisely solve the problem in polynomial time. Consequently, we propose to use a heuristic algorithm to tackle this problem.

In this section, we design an algorithm which ensures that every packet of UE n can be transmitted successfully within DL_n at a high resource utilization rate.

The main idea of the algorithm is to first forecast UE's average transmitting rate in the next TTI using the known transmitting rates and then choose some data packets of UEs whose channel conditions are very good to be transmitted together.

In order to introduce our algorithm more clearly, we explain the mean of some symbols appear in the algorithm in TABLE I following.

TABLE I

EXPLANATION OF THE SYMBOLS USED IN ALGORITHM 1

Symbols	Meaning
$v(n,r,t)$	The transmission rate of the rth RB transmitting the data packet of UE n at time t(known only after the scheduling time $t-1$)
$v_{n,t}$	The average rate of UE n with respect to RB at time t, calculated using $v_{n,t} = \sum_r v(n,r,t) \Big/ R$ only when the scheduling time is t
$\overline{v_n}$	The average rate of $v_{n,t}$ with respect to time t. It will be updated using the formula $\overline{v_n} = a\overline{v_n} + (1-a)v_{n,t}$,where a is an adjustable constant
U	The set of UEs that need RBs
b_n	The size of data of UE n
Δs_n	The size of a new data packet produced by UE n, it can be updated when the scheduling time reaches the time when a new packet is produced
k_n	The current maximum delay of data packets not transmitted to UE n
m_n	The number of RBs that UE n needs in the next Δt
m	The total number of RBs that will be scheduled in the next Δt

m_e	The number of RBs that will be scheduled at each time in the next Δt
tv	The time interval for scheduling resources in the next Δt
tt	The next scheduling time

Note: Except $v(n,r,t)$, other symbols in Table I are related with the scheduling time and they are updated as the scheduling.

The algorithm includes the following steps:

Step1. Forecast each UE's average rate in the next Δt at the time when a new data packet is produced.

```
for UE n =1 to N do
  Calculate v_{n,t} = Σ_r v(n,r,t) / R ;
  Update v̄_n = a v̄_n + (1 − a) v_{n,t} ;
end for
```

Where a is an adjustable constant, and its value is in the interval [0, 1]. What a's specific value is does not affect the algorithm. The optimal value of a can be got with statistics on a large amount of data.

Step2. Judge and choose the UEs that need some RBs in the next Δt.

```
for UE n = 1 to N do
  Update b_n = b_n + Δs_n ;
  if v̄_n ≤ b_n or k_n = (DL_n/Δt − 1) * Δt
  then
    Put n into U;
    Update k_n = 0 ;
  else
    Update k_n = k_n + Δt ;
  end if
end for
```

Where the UEs whose average transmission rate is less than the amount of data or delay time is close to delay limit are chosen to get RBs in the next Δt.

Step3. Estimate the number of RBs that needs to be scheduled in the next Δt according to the average rate, and find the number m_e of RBs that will be scheduled at each TTI of the next Δt.

```
for UE n ∈ U do
  Calculate m_n = ⌈b_n / v̄_n⌉ ;
end for
  Calculate m = Σ_{n∈U} m_n ;
  if m/Δt ≥ 1, then m_e = ⌈m/Δt⌉ , tv = 1 ;
  else m_e = 1 , tv = ⌊Δt/m⌋ ;
  end if
```

Where the number m of RBs that needs to be scheduled in the next Δt is the sum of number of RBs needed by UEs that are chosen in step 2. Then let it be assigned to each TTI of the next Δt averagely to get the number m_e.

Step4. With the largest rate method, assign m_e RBs to those UEs chosen in Step 2 at each TTI of the next Δt.

```
for i = 1 to m_e do
  if U ≠ ∅ then
    Find out the biggest rate v_max from
    V = {v(n,r,t) | n ∈ U, r = 1,...,R} ,
    denote the UE corresponding to v_max by n_max and the RB b
    r_max ;
    Schedule the r_max th RB to the UE n_max ;
    Update b_{n_max} = b_{n_max} − v_max , v(n_max, r_max, t) = 0,
    m_{n_max} = m_{n_max} − 1 ;
   if v̄_{n_max,t} ≥ (3/2) b_{n_max} and k_n < DL_n/Δt − 1 then
    Update m_{n_max} = 0 ;
   end if
   if m_{n_max} = 0 then get out UE n_max from U;
   end if
  end if
end for
```

Where the largest rate method is to assign one RB to the UE who needs RBs and has biggest transmit rate with the RB.

Based on the four steps above, we propose **Delay-limited (DL) scheduling algorithm** as following:

```
Algorithm1: Delay-limited (DL) scheduling algorithm:
 Initialization: Let U = ∅ . Δs_n = size of the first data packet
  produced by UE n. Let v̄_n = v_{n,0} , b_n = Δs_n , k_n = 0,
  n ∈ {1,...,N} ;
 for time t = 0 to T do
   Call Step 1;
   if t is the time of a new data packet being produced, then
       Call Step 2 ;
       Call Step 3;
   end if
   if t is the time of new data packet being produced or t = tt , then
       Call Step 4 ;
   end if
   tt = t + tv ;
 end for
```

B. Algorithm Analysis

The core idea of the DL algorithm above is delay limiting. Actually, we can prove that limiting is implemented in the DL algorithm using the following theorem.

Theorem: Let the size of all the data packets be Δs .Then, if the DL algorithm is feasible, the relation between the biggest delay of UE n's data packet and UE n's average transmission rate $\overline{v_n}$ is as follows:

(1) If $0 \le \overline{v_n} \le \Delta s$, the maximum delay of the UE is less than or equal to $2\Delta t$;

(2) If $\overline{v_n} > \Delta s$ and $\lceil \overline{v_n}/\Delta s \rceil \Delta t < DL_n$, the maximum delay of the UE is less than or equal to $\lceil \overline{v_n}/\Delta s \rceil \Delta t$;

(3) If $\overline{v_n} \geq \Delta s$ and $\lceil \overline{v_n}/\Delta s \rceil \Delta t \geq DL_n$, the maximum delay of the UE is less than or equal to DL_n.

Proof: For an arbitrary UE *n*, let its arbitrary data packet be produced at time t. Let the average transmission rate of the UE be $\overline{v_n}$ and the size of each data packets be Δs. And let the data not transmitted of the previous data packet be b_s, thus $\overline{v_n} \geq \frac{3}{2} b_s$. So

(1)If $0 \leq \overline{v_n} \leq \Delta s$, according to the DL algorithm, each data packet can be scheduled in an RB within Δt after it is produced. Moreover, despite b_s increase after each Δt *ms*, in all time $\overline{v_n} \geq \frac{3}{2} b_s$, so $b_s \leq 2/3\overline{v_n} < \overline{v_n} < \Delta s$. Thus, each data packet can be transmitted successfully within $2\Delta t$ after it is produced, that is, the maximum delay is $2\Delta t$.

(2) If $\overline{v_n} > \Delta s$ and $\lceil \overline{v_n}/\Delta s \rceil \Delta t < DL_n$ for a particular data packet whose production time is t_0, then according to the DL algorithm, the latest interval it can be scheduled in is $\left[t_0 + \left(\lceil \overline{v_n}/\Delta s \rceil - 1\right)\Delta t, t_0 + \lceil \overline{v_n}/\Delta s \rceil \Delta t\right]$. For $\overline{v_n} \geq \Delta s$, it can be scheduled completely, so the data packet can be scheduled successfully within $\lceil \overline{v_n}/\Delta s \rceil \Delta t$ after it was produced, and the maximum delay is $\lceil \overline{v_n}/\Delta s \rceil \Delta t$.

(3) If $\overline{v_n} \geq \Delta s$ and $\lceil \overline{v_n}/\Delta s \rceil \Delta t \geq DL_n$, according to the DL algorithm, the data packet produced at time t_0 will be scheduled at the latest in $[t_0 + DL_n - \Delta t,\ t_0 + DL_n]$, so the maximum delay is DL_n.

V. Simulation And Result

The key performance indicators of the algorithm given in this paper are the delay of each UE and the extent of saving radio resource. So, in this section, we analyse and evaluate the performance of the DL scheduling algorithm by simulation results.

We simulate the DL scheduling algorithm using real data from the Huawei Company. The data consist of the transmission rates of some UEs with 16 RBs available in 1000 *ms*. According to the actual situation, we set $\Delta t = 20$ ms, and the same size $\Delta s = 400$ bit for all packets. We also set the same delay limit for all UEs $DL_n = 140$ ms. In order to successfully simulate the situation, we let the data packets be produced in the first 860 *ms*, and no data packets are allowed to be produced in the final 140 *ms*. As a result, the total number of data packets being produced by each UE is 44. Based on the above assumptions, we make the following simulations:

D. Scheduling Results of DL Algorithm

For comparison purposes, we solve mathematical model I in Section III.C using CPLEX software whose result is close to the optimal solution and simulate the DL algorithms with the following three different cases:

Case1.1: T = 1000; N = 19; R = 16;
Case1.2: T = 1000; N = 19; R = 3;
Case1.3: T = 1000; N = 10; R = 16.

The results are shown in Table II.

TABLE II

TOTAL NUMBER OF RBs USED OF TWO METHODS

	Total number of RBs scheduled	
	DL algorithm	**CPLEX method**
Case1.1	450	339
Case1.2	530	393
Case1.3	220	175

As shown in Table II, the total number of RBs scheduled by DL algorithm is only about 30% larger than using solving the mathematic model directly. This indicates that the forecasting and estimation of the channel conditions are feasible and reasonable. Also, the DL algorithm takes the smallest amount of time to run, which shows that the DL algorithm is both feasible and effective. So, combining together the feasibility, effectiveness, and actual situation considered, the DL algorithm is reasonable.

In order to better understand the scheduling results of the DL algorithm, we simulate the DL algorithm using the data of Case 1.2. In Table III, we show explicitly the number of RBs used by each UE, the average delay and maximum delay of each UE with a certain average transmission rate.

TABLEIII

SCHEDULING RESULT FOR 19 UEs USING THE DL ALDORITHM

UE	1	2	3	4	5	6	7	8	9	10
Average transmission rate (bit)	277	284	317	356	502	570	523	579	666	687
Number of RBs used	58	66	57	48	36	30	31	29	25	25
Average delay (*ms*)	23	23	20	26	23	28	26	20	25	24
Max. delay (*ms*)	38	50	48	52	47	64	65	43	47	64
UE	11	12	13	14	15	16	17	18	19	

Average transmission rate (bit)	914	1019	1086	1209	1278	1370	1446	1858	2098	
Number of RBs used	19	18	16	14	14	13	12	10	9	
Average delay (*ms*)	34	34	34	32	33	43	45	44	54	
Max. delay (*ms*)	63	84	65	82	81	83	100	102	120	

As can be seen from Table III, the number of RBs scheduled for the UEs decreases with increasing average transmission rate. This is in line with the actual situation in which a UE with good channel conditions needs fewer resources to transmit the same amount of data. Besides, Table III also shows that the maximum delay for each UE is far less than 140 *ms*, which proves that the DL algorithm can guarantee the delay limit. Moreover, Table III shows that the average delay of each UE is less than 140/2 *ms*, which demonstrates that the DL algorithm has a good effect on delay.

E. Effect of the DL Algorithm on Delay

As the delay is one of the key issues of the VOIP services in LTE, we analyse the effect of delay of the DL algorithm in the next subsection.

From Table III, we know that the average delay for each UE is less than 140/2 *ms*, which means that the DL algorithm has a good effect on delay. Next, we investigate the stability of the delay resulting from the DL algorithm.

First, we consider the distribution of the data packets with different delays for each UE (see Fig. 1). For the sake of convenience, the delay is divided into the following three ranges:

Range 1: the delay is less than the average delay minus 20 *ms*;

Range 2: the delay is within±20 *ms* of the average delay;

Range 3: the delay is larger than the average delay plus 20 *ms*.

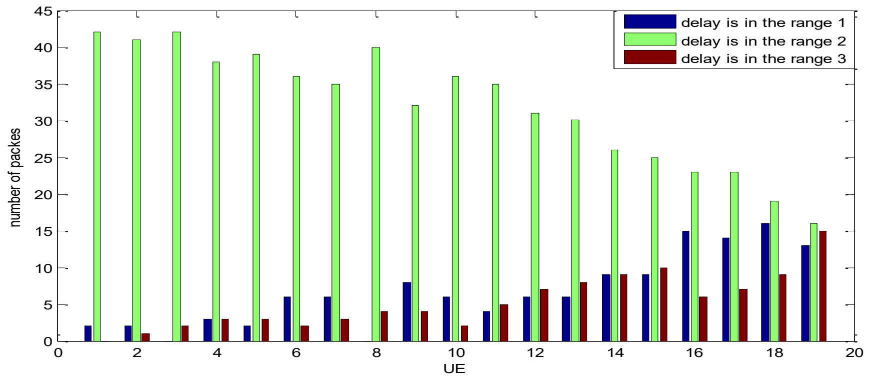

Fig.1. The distribution of packets for each UE according to their delay with respect to the average delay.

As can be seen from Fig. 1, the delays of most data packets fall in Range 2. In other words, the delays of most data packets are close to the average delay, which indicates that data packets with larger delays are relatively rare. In fact, further calculations show that the ratio of the number of 'late' data packets to the total number is less than 0.2 for all UEs. Besides, Fig. 1 also shows that the delays of the data packets for each UE are stable.

Next we consider the distribution of the maximum delay in the entire system, as shown in Fig. 2.

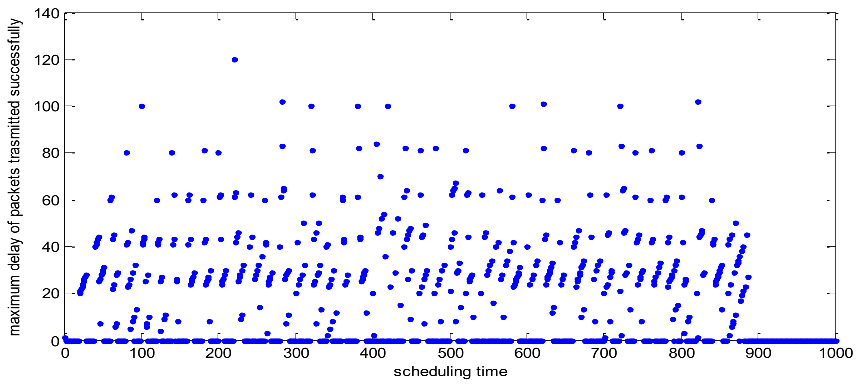

Fig. 2. The maximum delay of the packets transmitted completely at each time slot.

Fig. 2 shows the maximum delay of the packets that were transmitted completely at scheduling time t in the entire time domain. From the figure, we see that the maximum delay is mainly concentrated in the range 20–80 *ms* in the time domain. This is a good indication that the delay in the system as a whole is small and stable.

F. Effect on Resource Utilization of the DL Algorithm

As stated in the introduction, the scarcity of radio resources is one of the major problems the VOIP services encounter in scheduling resources. So, using a few resources and having a high resource utilization rate must be achieved for a good scheduling algorithm. Next, we will show the effect of the DL algorithm on resource utilization.

From Table II, we know that the total number of RBs scheduled using the DL algorithm is only about 30% more than the optimal solution, which indicates that the DL algorithm has a good effect on saving resources. Next we consider the resource utilization rate. Let TD_n be the total size of all data packets of UE *n*, and let CA_n be the total size of data that the RBs assigned to UE *n* can transmit. We define resource utilization rate RU_n of UE *n* with the formula

$$RU_n = TD_n / CA_n$$

In the simulations above, $TD_n = \left[\left(T - DL_n \right) / \Delta t \right] * \Delta s$ = $\left[\left(1000 - 140 \right) / 20 \right] * 400$ = 17,600. Then, according to the data in Table III, we can get the RU rate for each UE. The results are shown in Table IV.

TABLE IV

RESOURCE UTILIZATION RATE OF EACH UE

UE	1	2	3	4	5	6	7	8	9	10
RU	0.929	0.917	0.923	0.953	0.939	0.950	0.969	0.930	0.955	0.895
UE	11	12	13	14	15	16	17	18	19	
RU	0.939	0.925	0.932	0.888	0.893	0.943	0.975	0.884	0.918	

As Table IV shows, the RUs of all the UEs are almost larger than 0.9, meaning that when an RB is scheduled to one UE, it is fully utilized. Thus, the waste of resources is very little. The results therefore show that the DL algorithm has a high resource utilization.

Next, we demonstrate that the DL algorithm not only ensures a high resource utilization, but also has good uniformity across the entire time domain and stability at any particular time.

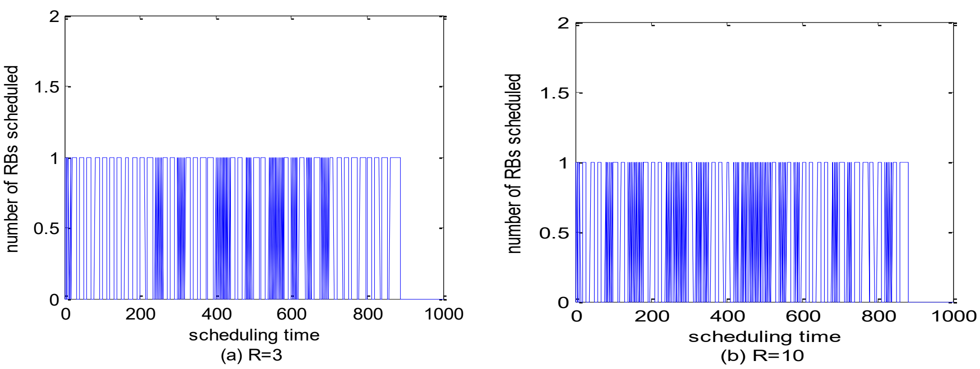

Fig.3. The number of RBs scheduled at each TTI.

Fig. 3 shows the number of RBs scheduled at time t when the total number (R) of RBs available at each TTI is equal to 3 and 10. From the figure, we can see that the number of RBs scheduled is uniformly distributed over the entire time domain. In addition, comparing Fig. 3(a) with 3(b), we can see that the numbers of RBs scheduled at each TTI in these two figures are almost the same. In other words, the number of RBs scheduled almost remains constant with the addition of available RBs. This indicates that the scheduling of the DL algorithm is stable in the entire system, and this also increases the likelihood of more UEs joining the system.

G. Compare the DL Algrithm with M-LWDF and EXP/PF Algrithms

In the last three parts of this section, we analyse and evaluate the performance of the DL scheduling algorithm. In order to show the performance of the DL scheduling algorithm, we compare it with M-LWDF and EXP/PF algorithms have been proved to be better scheduling algorithms for VOIP in LTE and have a good effect on delay. We simulate the DL, M-LWDF and EXP/PF algorithms in the following three different cases:

Case2.1: $T = 1000$; $N = 19$; $R = 3$;

Case2.2: $T = 1000$; $N = 80$; $R = 3$;

Case2.3: $T = 1000$; $N = 80$; $R = 5$.

The results are shown in Table V.

TABLE V

SCHEDULING RESULT OF THREE ALGORITHMS

	The number of RB scheduled in *T*		
	DL algorithm	M-LWDF algorithm	EXP/PF algorithm
Case 2.1	530	991	1155
Case 2.2	2415	2764	2811
Case 2.3	2218	4260	4058

As can be seen from Table V, the number of RBs scheduled using DL algorithm is far less than using M-LWDF and EXP/PF algorithms. Moreover, when the total time (T) and total number (N) of UEs don't vary while the number(R) of available RBs increases, the number of RBs scheduled by DL algorithm has no significant change, but the numbers of RBs scheduled by M-LWDF and EXP/PF algorithms substantially increase. This demonstrates that the DL algorithm does effectively ensure that the number of RBs used is as little as possible.

As is known, saving resources not only requires the number of RBs used to be as little as possible, but also needs full use of each RB. In order to analyse the extent of using each RB, we show the RU of each UE using DL, M-LWDF and EXP/PF algorithms with the result of Case 2.2 in Figure 4.

As shown in Fig. 4, the RU of each UE using DL algorithm is larger than 0.9, which is almost equal to the maximum RU of all UEs using LWDF and EXP/PF algorithms. This demonstrates that our DL algorithm makes full use of each RB, which is better than the other two algorithms.

As the above analysis shows, the DL algorithm can not only make the number of scheduled RBs to be very little, but also ensure each RB to be used fully.

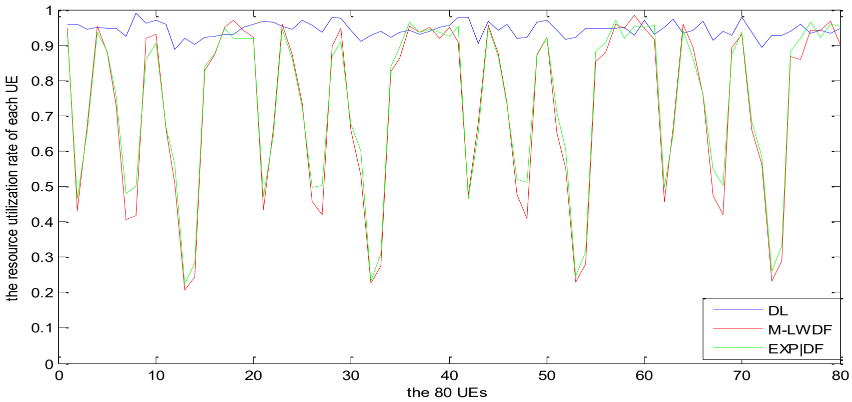

Fig. 4. The RU of each UE using DL, M-LWDF and EXP/PF algorithms

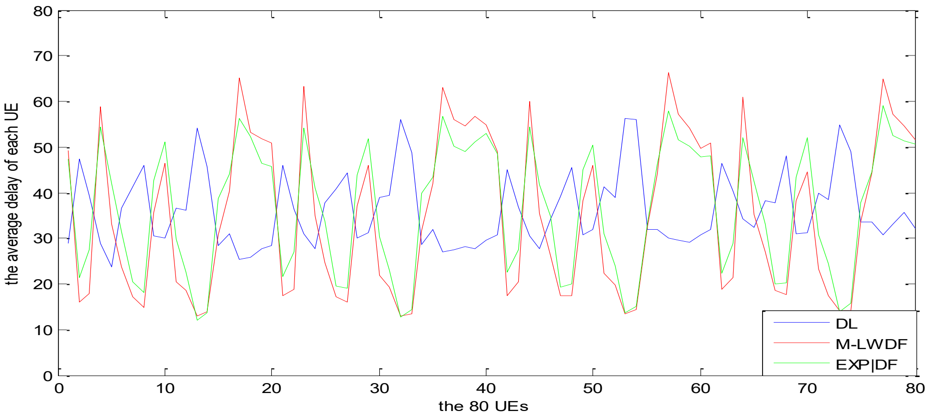

Fig. 5. The average delay of each UE using DL, M-LWDF and EXP/PF algorithms

In Part A, we analyse the effect on delay of DL algorithm through comparison with M-LWDF and EXP/PF algorithms. We analyse the fairness by showing average delay of each UE using DL, M-LWDF and EXP/PF algorithms in Fig. 5.

As can be seen from Fig. 5, when considered as a whole, the average delay of each UE using DL algorithm is not bigger than using M-LWDF and EXP/PF algorithms. Moreover, we can see that the gap of the average delay in different UEs using DL algorithm is less than using M-LWDF and EXP/PF algorithms, showing that the DL algorithm has a better fairness for all UEs on delay. By these facts, we can see that the DL algorithm has a good effect on fairness.

VI. CONCLUSIONS

In this paper, we construct a mathematical model and propose an efficient VoIP packet scheduling algorithm for VoIP in LTE. The cores of our model and algorithm are to meet the stringent delay requirements of VOIP by limiting the delay with a certain deadline and to solve the problem of scarcity of radio resources by minimizing the total number of radio resources scheduled in a certain period of time. As a result of applying our algorithm, we are able to ensure a high QOE for every UE and a high resource utilization rate. By conducting the simulations for our algorithm and comparing it with M-LWDF and EXP/PF algorithms, we found that our algorithm has a good effect on both delay and resource utilization.

ACKNOWLEDGMENT

This work is supported by the National 973 Plan project under Grant No. 2011CB706900 , the National 863 Plan project under Grant No.2011AA01A102 , the NSF of China (11331012 , 71171189), the "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06010302), and Huawei Technology Co. Ltd.

REFERENCES

[1] 3G Americas, "HSPA to LTE Advanced: 3GPP Broad band Evolution to IMT Advanced," *RYSAVY Research Rep*. Sept. 2009.

[2] S. Ahmadi,"An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Commun. Mag.* Vol. 47, no. 6, June 2009, pp. 84–98.

[3] S. Khirman and P. Henriksen, "Relationship between quality-of-Service and quality-of-Experience for public internet service," in Proc. *Passive and Active Measurement PAM, Fort Collins*, CO, Mar. 2002.

[4] Bang Wang et al, "Performance of VoIP on HSDPA", *IEEE VTC*, Vol.4:2335 – 2339, May 2005.

[5] 3GPP TSG RAN WG1 Meeting #47, R1-063275, "Discussion on control signalling for persistent scheduling of VoIP", Riga, Latvia, November 6 – 10, 2006.

[6] 3GPP TSG RAN WG1 Meeting #47bis, R1-070098, "Persistent Scheduling in E-UTRA", Sorrento, Italy, January 15 – 19, 2007.

[7] Dajie Jiang, Haiming Wang, EsaMalkamaki."Principle and Performance of Semi-persistent Scheduling for VoIP in LTE System", *in International Conference on Wireless Communications, Networking and Mobile Computing*, 2007.

[8] Jinhua Liu, Chunjing Hu, Zhangchao Ma, "Semi-Persistent Scheduling for VoIP Service in the LTE-Advanced Relaying Networks" *IEEE ICCCAS*, 2010.

[9] Ericsson, N.S, "Adaptive modulation and scheduling of IP traffic over fading channels", *VTC 1999-Fall. IEEE VTS 50th,* 1999(2): 849-853.

[10] Lang T J, Williams E A, Crossley W A. "Average and Maximum Revisit Time Trade Studies For Satellite Constellation Using a Multiobjective Genetic Algorithm" [J]. JOURNAL of Astronautical Sciences, 2001, 49<3>: 385-400.

[11] Kian Chung Beh, Armour, S. and Doufexi, A. "Joint Time-Frequency Domain Proportional Fair Scheduler with HARQ for 3GPP LTE systems". *Vehicular Technology Conference*, 2008.VTC 2008-Fall. IEEE 68th, 2008: 1-5.

[12] Bilal Sadiq, Seung Jun Baek, Member. "Delay-Optimal Opportunistic Scheduling and Approximations: The Log Rule. "IEEE: 2013.

[13] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 146–158, Apr. 1999.

[14] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[15] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett*, vol. 9, no. 3, pp. 210–212 Mar. 2005.

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

[17] H. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well-known packet scheduling algorithms in the downlink 3GPP LTE system," in Proc. o*f IEEE Malaysia International Conf. on Comm., MICC, Kuala Lumpur,Malaysia*, 2009, pp. 815 –820.

[18] F Capozzi, G Piro, La Grieco, G boggia, P Camarda, "Downlink packet scheduling in lte cellular networks: Key Design Issues and A survey," *IEEE Commun. Surv. Tutorials*, 2012.

[19] 3GPP, "Technical Specifications Group Radio Access Network -Physical channel and Modulation (release 8)," 3GPP TS 36.211.

[20] Giuseppe Piro, Luigi Grieco, Gennaro Boggia, etc. "Two-level downlink scheduling for real-time multimedia services in LTE networks." *IEEE Trans. on Multimedia*, vol.13, no.5, Oct. 2011.

[21] Davinder Singh, Preeti Singh, "Radio Resource Scheduling in 3GPP LTE: A Review" IJETT, 2013.

[22] S. Na and S. Yoo, "Allowable propagation delay for VoIP calls of acceptable quality," in Proc. 1st Int. *Workshop Advanced Internet Services and Applications, Springer-Vorlage*, London, U.K., 2002, pp.47–56.

[23] G.-M. Su, Z. Han, M. Wu, and K. Liu, "Joint uplink and downlink optimization for real-time multiuser video streaming over WLANs," *IEEE J. Sel. Topics Signal Process*, vol. 1, no. 2, pp. 280–294, Aug.2007.