Downlink Scheduling in LTE: Challenges, Improvement, and Analysis

Mohamad Omar Kayali, Zeinab Shmeiss, Haidar Safa, and Wassim El-Hajj
Department of Computer Science
American University of Beirut
Beirut, Lebanon
Email: {mmk77, zhs07}@mail.aub.edu, {hs33, we07}@aub.edu.lb

***Abstract*—Long Term Evolution (LTE) was developed by 3GPP to cope with the increasing demand for better Quality of service (QoS) and the emergence of bandwidth-consuming multimedia applications. Today's data transmission networks face extreme challenges in providing high data rate and low latency. Scheduling paradigms such as Round Robin, Best Channel Quality Condition and Proportional Fair are commonly adopted in current LTE downlink scheduling algorithms, but they are far from optimal for satisfying latency requirements. In this paper, we first survey the state of the art downlink scheduling algorithms in LTE and identify their main challenges. We then formulate the LTE downlink scheduling problem as an optimization problem in order to meet the flow deadlines, then incorporate the formulation within the surveyed scheduling algorithms, to produce better performance. We consider strict deadlines for different types of packets with the goal of maximizing resource distribution. Additionally, in our formulation the buffer state for each user is taken into consideration in order to minimize the packet loss. We evaluate the proposed formulation using LTE-Sim and study its positive impact on the existing LTE downlink scheduling algorithms; the performance in terms of QoS, packet loss and fairness is improved throughout all evaluations.**

***Index Terms*—LTE, QoS, Downlink Scheduling, Integer Linear Programming, Buffer State.**

I. Introduction

Long term evolution (LTE) offers revolutionary performance when compared to its predecessors, GSM/EDGE and UMTS/HSPA network technologies. LTE is part of the evolved packet system (EPS) which was developed by the third generation partnership project (3GPP). EPS incorporated several technologies such as orthogonal frequency division multiplex (OFDM) and multiple inputs multiple outputs (MIMO). It is an all-IP network, consisting of the evolved UTRAN (eUTRAN) and the evolved packet core (EPC), as shown in Figure 1. The eUTRAN contains only one simplified entity, the evolved NodeB (eNB) that is responsible for radio resource management, making the core of the network flatter and less complex, thus allowing for reduced latencies. The eUTRAN is the access point of the user equipment (UE) to the network. The EPC is comprised of several components such as the packet data network gateway (P-GW), the serving gateway (S-GW), the mobility management entity (MME), and the home subscriber server (HSS). LTE specification was published and first deployed in 2009 [1]. It supports high data transmission rate of more than 100 Mb/s and operates on the downlink and the uplink air interfaces. Downlink is the transmission of data from the base station to the mobile station. As opposed to downlink, uplink is the transmission of data from the mobile station to the base station.

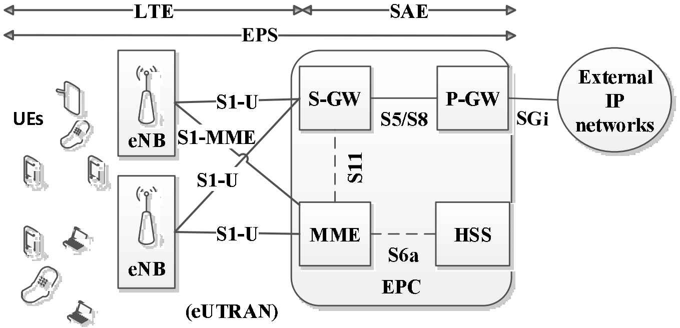


Fig. 1: Evolved Packet System Overview

OFDM has been adopted as the downlink transmission scheme for the 3GPP LTE. It involves frequency division multiplexing (FDM) and multi-carrier communication. It divides the available bandwidth into many sub-carriers and allows multiple users to access the system simultaneously. Guard bands are used between sub-carriers to avoid interference. In multi-carrier FDM, the data of the user can be split into multiple sub-streams and transmit them in parallel to make the data rate higher. Orthogonality allows sub-carriers to overlap and save bandwidth. Thus, achieving higher data rate.

A scheduler is a key element in the eNB which assigns the shared physical resources to different users. There are several downlink scheduling algorithms such as the Round Robin (RR) Scheduling, Best Channel Quality Indication (Best CQI) Scheduling, and Proportional Fair (PF) Scheduling. In RR, terminals are assigned one after another without taking any factor into consideration. Although this method results in poor performance, fairness is guaranteed since all terminals are equally scheduled. Best CQI scheduling assigns resource blocks to the user with the best radio link conditions. On the other hand, Proportional Fair (PF) scheduling tries to maximize total throughput while providing all users at least a minimal level of service. Thus, it balances between throughput and fairness among all the UEs. Because the scheduler aims to maximize the system performance, the design of the scheduling algorithms has became a major issue. However, the sharp growth of QoS applications makes the aim much more challenging. It is well known that best-effort applications that require non-real time traffic do not call for strict requirements on packet delay, whereas real time services are delay sensitive and should be transmitted as soon as possible. The 3GPP specifications did not define scheduling algorithms that support real-time QoS applications [2].

In this paper, we formulate the LTE downlink scheduling as an optimization problem where the objective is to optimize parameters such as the resource distribution, data rate, packet delay, and even buffer overflow. Our formulation is then incorporated in the existing state of the art LTE downlink scheduling algorithms leading to enhanced performance on all fronts. The remaining of this paper is organized as follows. In Section 2, we present some basic concepts and related work. In Section 3, we mathematically formulate the scheduling problem in LTE using integer linear programming and incorporate the formulation in the surveyed algorithms. In Section 4, we evaluate the performance of the proposed solution. Conclusion is drawn in Section 5.

II. Basic concepts and related work

In a base station (BS), the scheduler is a key element that assigns the shared physical resources to different users in the cell. LTE downlink physical resource is represented as a time-frequency resource grid consisting of multiple Resource Blocks (RB). A resource block (RB) is the smallest unit of resources that can be allocated to a user. A resource block is 180 kHz wide in frequency and has a duration of 0.5 msec (one slot). Figure 2 shows the resource blocks which are divided into multiple Resource Elements (RE).

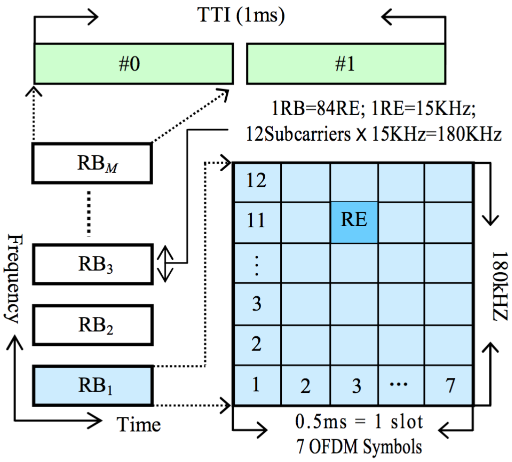


Fig. 2: Resource Blocks

Many state-of-the-art downlink scheduling algorithms were proposed in the literature. The most relevant ones for our work are:

1) **Proportional Fair (PF)**: It aims to balance between maximizing the bit rate and fairness. The PF scheduling effectively reduces variations in user bit rates. It was shown that as long as the user average signal-to-interference-plus-noise ratio (SINR) are fairly uniform the PF scheduler provides an ultimate fairness performance with a moderate loss in throughput [3].
2) **Maximum-Largest Weighted Delay First (M-LWDF)**: It uses LWDF for bounded delay and PF for fairness. M-LWDF scheduler outperforms other downlink scheduling algorithms by granting higher system throughput and guaranteeing fairness [4]. Thus, M-LWDF is considered one of the best candidates for real-time packet scheduling.
3) **EXP/PF**: It uses exponential rule for bounded delay and PF for fairness. It was demonstrated that M-LWDF scheduler outperforms EXP/PF algorithm for lower loads while the computationally complex EXP/PF scheduler performs better for higher loads in downlink LTE system supporting multimedia services [5].
4) **EXP-rule**: It uses exponential rule for bounded delay and PF for channel awareness. [6] proposes the EXP rule to provide QoS guarantees for real-time packet scheduling.
5) **LOG-rule**: It uses logarithm rule for bounded delay and PF for fairness. The LOG rule scheduler was designed to balance in QoS metrics taking the mean delay and robustness into consideration [7]. Moreover, it allocates resources to users in the same way as EXP rule to maximize throughput.
6) **DP-VT-MLWDF**: It is a scheduler that maximize the QoS performance of real-time traffic while sacrificing an acceptable performance of nonreal-time traffic by effectively using the delay priority function [8].

Every scheduler uses different strategy for assigning resources. Table I shows the parameters used in the process of resource allocation.

TABLE I: Parameters Used by Schedulers

	PF	M-LWDF	EXP/PF	[EXP/LOG]-rule
SINR	x	x	x	x
Throughput	x	x	x	x
HoL PD		x	x	x
Target Delay		x	x	x
Target PLR		x	x	
Queue Length				
Buffer State				

The LTE downlink scheduling problem has been addressed by many researchers in the field. [9] addresses the state-of-the-art algorithms that tackle the problem of scheduling QoS applications. However, these algorithms fail to meet all QoS requirements mentioned in the literature [10]. Some work such as [8] proposes a variation to some state-of-the-art algorithms in order to increase QoS performance by taking into consideration the delay sensitivity. Alternatively, some proposes solutions that tackle the problem of scheduling by monitoring the buffer state of end users [11]. However, they didn't take into account the QoS of real time and non-real time applications.

The downlink scheduling problem was formulated as an optimization problem in approaches [12], [13], [14], [15].

However, [13], [15] didn't take into consideration the buffer state or major QoS parameters. Moreover, [12], [14] took into consideration the QoS but only for specific class of real time applications such as VoIP or Video Stream.

III. Proposed Work

Our proposed approach takes into consideration all QoS parameters including the buffer state of end users and formulates the scheduling problem as a linear optimization problem. Our main goal is to optimize the usage of available network resources in order to provide better QoS in LTE networks. This is done by selecting the optimal parameters and the appropriate scheduling algorithm to obtain the best QoS. Indeed, our solution aims to maximize QoS performance while maintaining fairness among all packets. The value of the parameters used in the scheduling process can be determined based on the following factors [12]:

1) Quality of Service (QoS): data that has the lowest Quality Class Identifier (QCI) value will have the highest metric.
2) Channel Quality Indicator (CQI): based on the CQI value, the highest throughput will have the highest metric.
3) The transmission queues state: the longest queue will have the highest metric.
4) Resource Allocation History: based on the performance history, the lowest previous throughput will have the highest metric.
5) Buffer State: based on the buffer state at the UE, the buffer with the highest available space will have the highest metric.

It is worth mentioning that overwhelming the user with packets will result in buffer overflow which leads to packet loss. In our scheduling process the buffer status is considered while assigning resource blocks to different users which is done to avoid packet loss. UEs provide the eNodeB scheduler with its buffer state information such as the terminal buffer length L_{buff} and the amount of current received packets N_{curr}. The variable N_{curr} is initialized to 0 when the data traffic process is established. The increment step of variable N_{curr} is 1. The N_{curr} is flushed to 0 when it reaches buffer length L_{buff} [11]. The buffer of user *i* is $buff_i$ which is described as follows:

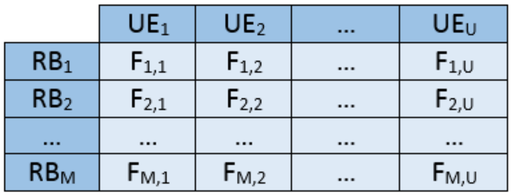
$$Buff_i = \frac{L_{buff} - N_{curr}}{L_{buff}} \tag{1}$$

In order to decrease the packet dropping probability due to buffer overflow, our approach schedules more resources to users with higher spare buffer space.

We next discuss the proposed problem formalization. Suppose we have M available RBs and U users. We define the variable matrix x[M, U] such that:

$$x_{i,j}(t) = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ RB is allocated to the } j^{\text{th}} \text{ user at time t} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the sum of all $x_{i,j}$ is less than or equal to 1 for any particular RB implying that only one user can employ the i^{th} resource block. We also define the matrix $F[M, U](t)$ where $F_{i,j}$ represents the behavior of the j^{th} user on the i^{th} RB at time *t* according to each of the selected algorithms (Figure 3).



	UE$_1$	UE$_2$	...	UE$_U$
RB$_1$	F$_{1,1}$	F$_{1,2}$	...	F$_{1,U}$
RB$_2$	F$_{2,1}$	F$_{2,2}$	...	F$_{2,U}$
...	...	...	...	...
RB$_M$	F$_{M,1}$	F$_{M,2}$	...	F$_{M,U}$

Fig. 3: $F_{i,j}$ Matrix

We will formulate many state-of-the-art scheduling algorithms using our optimization solution by adding to them all the QoS parameters described in Table II. In the proportional fairness scheduling algorithm [3], $F_{i,j}(t)$ is given as:

$$F_{i,j}(t) = r_{i,j}(t) = \frac{\log\left(1 + SINR_{i,j}(t)\right)}{R^i(t-1)} \tag{3}$$

where SINR is the signal to noise ratio reported by the UE and $R^i(t-1)$ is the past average throughput achieved by user *j* at time t - 1.

In the M-LWDF scheduling algorithm [12], $F_{i,j}(t)$ is given as:

$$F_{i,j}(t) = \begin{cases} \alpha_j * DHOL_j * r_{i,j}(t) & \text{for real time applications} \\ r_{i,j}(t) \quad \text{otherwise} \end{cases} \tag{4}$$

where

$$\alpha_j = \frac{\log \delta_j}{\tau_j}$$

with δ_j is the acceptable packet loss rate, τ_j is the delay threshold, and $DHOL_j$ is the head of line delay for the j^{th} user.

In EXP/PF scheduling algorithm [12], $F_{i,j}(t)$ is given as:

$$F_{i,j}(t) = \begin{cases} exp(\frac{\alpha_j * DHOL_j - \chi}{1 + \sqrt{\chi}}) * r_{i,j}(t) & \text{for real time applications} \\ r_{i,j}(t) \quad \text{otherwise} \end{cases} \tag{5}$$

where

$$\chi = \frac{1}{N_{rt}} \sum_{k=1}^{N_{rt}} \alpha_k * DHOL_k$$

and N_{rt} is the number of active downlink real-time flows.

In delay priority scheduler (DP-VT-MLWDF)[8], $F_{i,j}(t)$ is given as:

$$F_{i,j}(t) = \begin{cases} \alpha_j * \frac{Q_j(t)}{\tau_j - DHOL_j} * r_{i,j}(t) & \text{for real time applications} \\ r_{i,j}(t) \quad \text{otherwise} \end{cases} \tag{6}$$

with $Q_j(t)$ being the length of token queue for the j^{th} user.

Furthermore, the end-user buffer ($Buff_j$) status which is reported by the UE can be easily expressed through an optimization constraint where we force the scheduler to assign RB to the user such that the eNodeB does not transfer data more than the UE can handle. When we do so we are preventing the UE buffer from overflow. Thus, minimizing the ratio of packet loss making the system perform better on the long term.

The formalization of the optimization approach is expressed as follows:

$$\begin{aligned} \underset{X}{\text{maximize}} \quad & \sum_{i=1}^{M}\sum_{j=1}^{U} x_{i,j}F_{i,j}(t) \\ \text{subject to} \quad & C1: \sum_{j=1}^{U} x_{i,j} \leq 1 \quad \forall i = 1, ..., M \\ & C2: x_{i,j} \in 0, 1 \\ & C3: \sum_{i=1}^{M} x_{i,j}r_{i,j}(t) \leq r_j \quad \forall j = 1, ..., U \\ & C4: \sum_{i=1}^{M} x_{i,j}r_{i,j}(t) \leq Buff_j \quad \forall j = 1, ..., U \end{aligned} \tag{7}$$

Constraint C1 ensures that every RB is allocated to at most one user. C2 forces the value of the variable $x_{i,j}$ to be 0 or 1. C3 implies that the total data rate assigned to the users should not exceed their demands, and also should not lead to an overflow on the user side buffer which is expressed in constraint C4. The optimization problem presented is a 0-1 integer linear programming in which the variables are restricted to be integers. This type of problems is considered NP-hard where the unknowns are binary. Thus, relaxation must be done on constraint C2 by replacing it with the following constraints: $0 \leq x_{i,j}$ and $x_{i,j} \leq 1$. After this relaxation, the problem is transformed to linear programming optimization which can be easily solved by the two-phase simplex algorithm. However, in order to preserve the binary property of the unknowns the outcome of the simplex algorithm is rounded to 0 or 1.

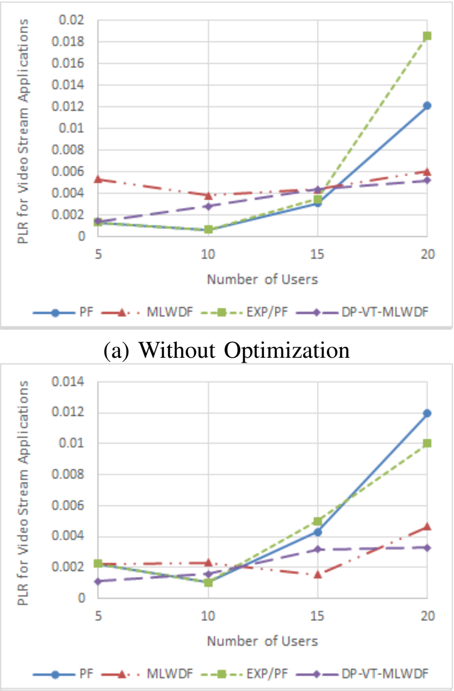
IV. Evaluation and Preliminary Results

In this section, we present the performance of some of the state-of-the-art algorithms (PF, MLWDF, EXP/PF, DP-VT-MLWDF) with and without the optimization while taking the buffer state into consideration. LTE-Sim simulator was used for simulations. The experiments were performed on a uniformly distributed real-time and nonreal-time applications on the available users. To ensure accuracy, each run was conducted 5 times, with different number of users and the average was taken. Table III summarizes the system parameters used in the simulations.

TABLE III: Simulation Parameters

Parameters	**Values**
Bandwidth	10MHz
Number of RBs	50
Frame Structure	FDD
Cell Radius	1 kmn
No. of Users	5:5:20
UE speed	3 km/h

In our evaluation, we measured the average packet loss ratio and the average total throughput for the video stream applications and nonreal-time applications. Figure 4 shows the packet loss ratio for video stream applications. It can be noticed that as the number of users increase the packet loss ratio for PF, EXP/PF, and DP-VT-MLWDF increases in both (A) and (B). However, the PLR in MLWDF scheduler is not affected by the number of users. Moreover, Figure 5 presents a comparison of the PLR between both optimization and non-optimization approach for 20 users. We noticed that the PLR for MLWDF, EXP/PF, and DP-VT-MLWDF is less when using the optimization approach. On the other hand, the PLR for PF scheduler is the same for both approaches. This is because the PF scheduler doesn't take the packet priority into consideration while assigning resources. Consequently, adding the formulation we proposed to the existing scheduling algorithms, either improved on the ”average packet loss ratio” metric or did not worsen it.



(a) Without Optimization

(b) With Optimization

Fig. 4: PLR of Video Stream Applications.

TABLE II: Standardized QoS Class Identifiers for LTE

QCI	Resource Type	Priority	Packet Delay Budget [ms]	Packet Loss Rate	Example services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-Conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real time gaming
5	non-GBR	1	100	10^{-6}	IMS signaling
6	non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	non-GBR	8	300	10^{-6}	TCP based (e.g., WWW, e-mail)
9	non-GBR	9	300	10^{-6}	TCP based (e.g., WWW, e-mail)

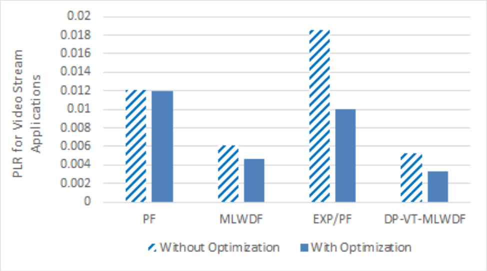
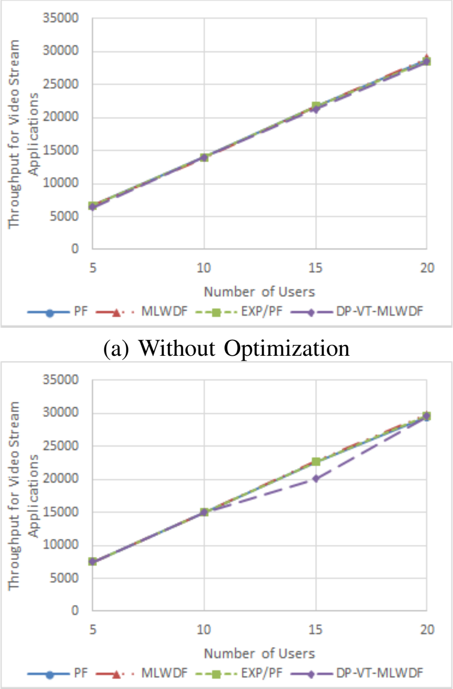


Fig. 5: PLR of Video Stream Applications for 20 Users

Figure 6 presents the throughput for video stream applications with and without optimization. As presented in the figure 6-(a), in all cases, the throughput tends to increase as the number of users increases. However, and as shown in figure 6-(b) and figure 7, when the suggested optimization is incorporated, the optimized approach outperforms the traditional approaches, in terms of throughput, for all selected algorithms. Hence, the optimization techniques force the scheduler to select the best choice with the best throughput.



(a) Without Optimization

(b) With Optimization

Fig. 6: Throughput of Video Stream Applications.

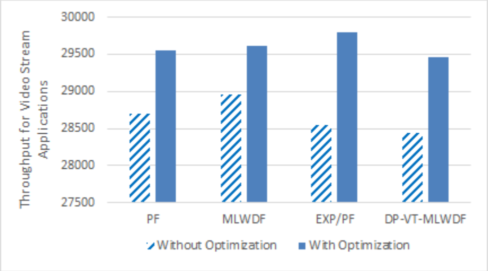
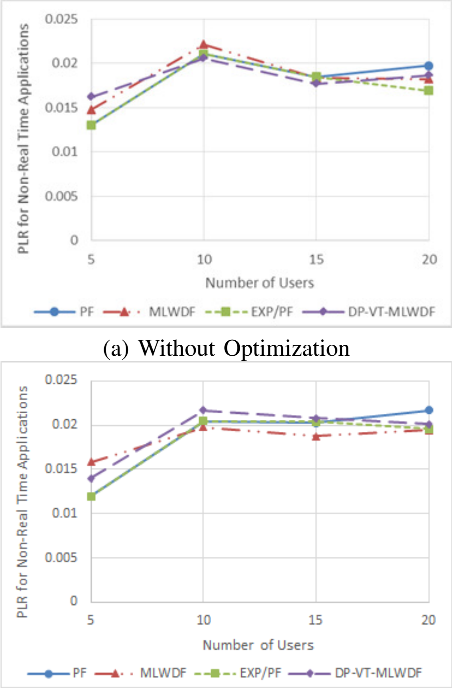


Fig. 7: Throughput of Video Stream Applications for 20 Users

Figure 8 shows the packet loss ratio of non-real time applications. The PLR tends to remain within the same range when the number of users exceeds 10 users. Moreover, the results tend to be slightly less when using the non-optimized approach. This is because the optimized approach gives priority for QoS packets while sacrificing an acceptable performance for non-real time traffic.

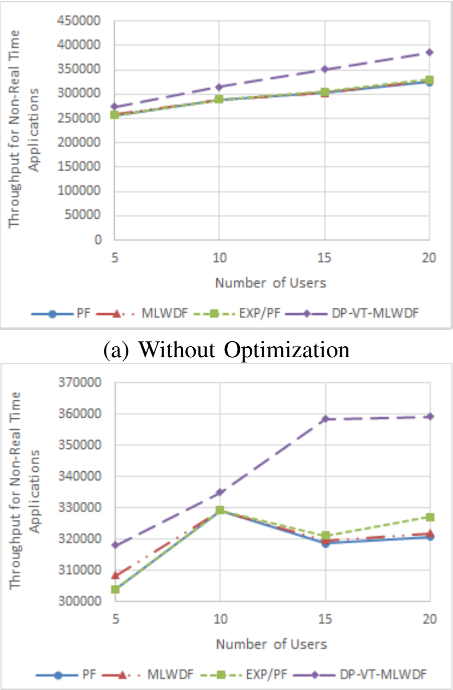


(a) Without Optimization

(b) With Optimization

Fig. 8: PLR of Non-Real Time Applications.

Figure 9 shows the throughput when using non-real time applications. The results show that when the optimized formulation is incorporated in the scheduling algorithms (PF, MLWDF, and EXP/PF), the algorithms perform better than the regular non-optimized approach.



(a) Without Optimization

(b) With Optimization

Fig. 9: Throughput of Non-Real Time Applications

V. Conclusion

In this paper, we have addressed the problem of downlink scheduling for QoS packet flow in LTE networks. The state-of-the-art scheduling algorithms have been formulated using integer linear programming, and solved. The formulation considers in addition to the regular QoS parameters, the UE buffer state parameter as an enhancement. The effects of the proposed approach have been studied and evaluated to demonstrate that it is suitable to provide better services for QoS and best-effort applications. Obtained simulation results confirm the effectiveness of the proposed approach. They show the performance of the original state-of-the-art scheduling algorithms and how incorporating our optimization formulation have impacted positively their performance while measuring parameters such as throughput and the packet loss rate.

References

[1] C. Cox. *An introduction to LTE: LTE, LTE-advanced, SAE and 4G mobile communications*, 2012.

[2] Bin Liu, Hui Tian, and Lingling Xu. An efficient downlink packet scheduling algorithm for real time traffics in lte systems. In *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, pages 364–369, Jan 2013.

[3] R. Kwan, C. Leung, and J. Zhang. Proportional fair multiuser scheduling in lte. *IEEE Signal Processing Letters*, 16(6):461–464, June 2009.

[4] H. A. M. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand. Performance of well known packet scheduling algorithms in the downlink 3gpp lte system. In *2009 IEEE 9th Malaysia International Conference on Communications (MICC)*, pages 815–820, Dec 2009.

[5] R. Basukala, H. A. M. Ramli, and K. Sandrasegaran. Performance analysis of exp/pf and m-lwdf in downlink 3gpp lte system. In *2009 First Asian Himalayas International Conference on Internet*, pages 1–5, Nov 2009.

[6] Sanjay Shakkottai and Alexander L Stolyar. Scheduling for multiple flows sharing a time-varying channel: The exponential rule. *Translations of the American Mathematical Society-Series 2*, 207:185–202, 2002.

[7] Bilal Sadiq, Seung Jun Baek, and Gustavo De Veciana. Delay-optimal opportunistic scheduling and approximations: The log rule. *IEEE/ACM Transactions on Networking (TON)*, 19(2):405–418, 2011.

[8] Yuan-Ping Li, Bin-Jie Hu, Hui Zhu, Zong-Heng Wei, and Wei Gao. A delay priority scheduling algorithm for downlink real-time traffic in lte networks. In *Information Technology, Networking, Electronic and Automation Control Conference, IEEE*, pages 706–709. IEEE, 2016.

[9] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda. Downlink packet scheduling in lte cellular networks: Key design issues and a survey. *IEEE Communications Surveys Tutorials*, 15(2):678–700, Second 2013.

[10] Giuseppe Piro, Luigi Alfredo Grieco, Gennaro Boggia, Rossella Fortuna, and Pietro Camarda. Two-level downlink scheduling for real-time multimedia services in lte networks. *Trans. Multi.*, 13(5):1052–1065, October 2011.

[11] Yan Lin and Guangxin Yue. Channel-adapted and buffer-aware packet scheduling in lte wireless communication system. *Proc. Int. Conf. Wireless Communications, Networking and Mobile Computing, WiCOM, Dalian, China*, 2008.

[12] Tarik Ghalut, Hadi Larijani, and Ali Shahrabi. Qoe-aware optimization of video stream downlink scheduling over {LTE} networks using {RNNs} and genetic algorithm. *Procedia Computer Science*, 94:232 – 239, 2016. The 11th International Conference on Future Networks and Communications (FNC 2016) / The 13th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2016) / Affiliated Workshops.

[13] Raymond Kwan, Cyril Leung, and Zhang Jie. Proportional fair multiuser scheduling in lte. *IEEE Signal Processing Letters*, 16(6):461–464, 2009.

[14] Duy-Huy Nguyen, Hang Nguyen, and Eric Renault. We-mqs-voip priority: An enhanced lte downlink scheduler for voice services with the integration of voip priority mode. *International Journal of Advanced Computer Science and Applications IJACSA*, 7(7):560–567, 2016.

[15] Yunzhi Qian, Canjun Ren, Suwen Tang, and Ming Chen. Multi-service qos guaranteed based downlink cross-layer resource block allocation algorithm in lte systems. In *Wireless Communications & Signal Processing, 2009. WCSP 2009. International Conference on*, pages 1–4. IEEE, 2009.