Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey

F. Capozzi, *Student Member, IEEE,* G. Piro, *Student Member, IEEE,*
L.A. Grieco, *Member, IEEE,* G. Boggia, *Senior Member, IEEE,* and P. Camarda

Abstract

Future generation cellular networks are expected to provide ubiquitous broadband access to a continuously growing number of mobile users. In this context, LTE systems represent an important milestone towards the so called 4G cellular networks. A key feature of LTE is the adoption of advanced Radio Resource Management procedures in order to increase the system performance up to the Shannon limit. Packet scheduling mechanisms, in particular, play a fundamental role, because they are responsible for choosing, with fine time and frequency resolutions, how to distribute radio resources among different stations, taking into account channel condition and QoS requirements. This goal should be accomplished by providing, at the same time, an optimal trade-off between spectral efficiency and fairness. In this context, this paper provides an overview on the key issues that arise in the design of a resource allocation algorithm for LTE networks. It is intended for a wide range of readers as it covers the topic from basics to advanced aspects. The downlink channel under frequency division duplex configuration is considered as object of our study, but most of the considerations are valid for other configurations as well. Moreover, a survey on the most recent techniques is reported, including a classification of the different approaches presented in literature. Performance comparisons of the most well-known schemes, with particular focus on QoS provisioning capabilities, are also provided for complementing the described concepts. Thus, this survey would be useful for readers interested in learning the basic concepts before going into the details of a particular scheduling strategy, as well as for researchers aiming at deepening more specific aspects.

Index Terms

LTE, scheduling, Radio Resource Management, QoS, System-Level Simulation.

F. Capozzi is with ITIA-CNR, v. P Lembo, 38F - 70124, Bari, Italy. e-mail: francesco.capozzi@itia.cnr.it. This work was done when he was at "DEE - Dip. di Elettrotecnica ed Elettronica", Politecnico di Bari.

Other authors are with the "DEE - Dip. di Elettrotecnica ed Elettronica", Politecnico di Bari, v. Orabona, 4 - 70125, Bari, Italy. e-mail: {g.piro,a.grieco,g.boggia,camarda}@poliba.it.

Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey

List of Acronyms

AMC	Adaptive Modulation and Coding
BLER	Block Error Rate
BET	Blind Equal Throughput
CQI	Channel Quality Indicator
DCI	Downlink Control Information
eNB	evolved NodeB
FDPS	Frequency Domain Packet Scheduler
FLS	Frame Level Scheduler
GBR	Guaranteed bit-rate
HARQ	Hybrid Automatic Retransmission Request
LTE	Long Term Evolution
LWDF	Largest Weighted Delay First
MCS	Modulation and Coding Scheme
M-LWDF	Modified LWDF
MT	Maximum Throughput
OFDM	Orthogonal Freq. Division Multiplexing
OFDMA	Orthogonal Freq. Division Multiple Access
PDCCH	Physical Downlink Control Channel
PDSCH	Physical Downlink Shared Channel
PUSCH	Physical Uplink Shared Channel
PF	Proportional Fair
PLR	Packet Loss Rate
PSS	Priority Set Scheduler
QCI	QoS Class Identifier
QoS	Quality of Service
RB	Resource Block

RLC	Radio Link Control
RR	Round Robin
RRM	Radio Resource Management
SC-FDMA	Single Carrier Freq. Division Multiple Access
SGW	Serving Gateway
TDPS	Time Domain Packet Scheduler
TTA	Throughput To Average
TTI	Transmission Time Interval
UE	User Equipment
VPM	VoIP priority mode

I. Introduction

The growing demand for network services, such as VoIP, web browsing, video telephony, and video streaming, with constraints on delays and bandwidth requirements, poses new challenges in the design of the future generation cellular networks. 3GPP [1] introduced the Long Term Evolution (LTE) specifications [1] as an answer to this need, aiming at ambitious performance goals and defining new packet-optimized and all-IP architectures for the radio access and the core networks. According to [2], there are already more than 20 LTE cellular operators worldwide, and more than 32 million LTE subscribers are foreseen by 2013. For this reason, both research and industrial communities are making a considerable effort on the study of LTE systems, proposing new and innovative solutions in order to analyze and improve their performance.

LTE access network, based on Orthogonal Freq. Division Multiple Access (OFDMA), is expected to support a wide range of multimedia and Internet services even in high mobility scenarios. Therefore, it has been designed to provide high data rates, low latency, and an improved spectral efficiency with respect to previous 3G networks. To achieve these goals, the Radio Resource Management (RRM) block exploits a mix of advanced MAC and Physical functions, like resource sharing, Channel Quality Indicator (CQI) reporting, link adaptation through Adaptive Modulation and Coding (AMC), and Hybrid Automatic Retransmission Request (HARQ).

In this context, the design of effective resource allocation strategies becomes crucial. In fact, the efficient use of radio resources is essential to meet the system performance targets and to satisfy user

[1] http://www.3gpp.org.

needs according to specific Quality of Service (QoS) requirements [3].

The packet scheduler works at the radio base station, namely the evolved NodeB (eNB), and it is in charge of assigning portions of spectrum shared among users, by following specific policies. With respect to well-known schemes conceived for wired networks (and summarized in section IV-A), in a wireless scenario the packet scheduler plays an additional fundamental role: it aims to maximize the spectral efficiency through an effective resource allocation policy that reduces or makes negligible the impact of channel quality drops. In fact, on wireless links the channel quality is subject to high variability in time and frequency domains due to several causes, such as fading effects, multipath propagation, Doppler effect, and so on.

For these reasons, channel-aware solutions are usually adopted in OFDMA systems because they are able to exploit channel quality variations by assigning higher priority to users experiencing better channel conditions.

Nevertheless, many issues arise in the design of such solutions in LTE systems, spanning from the provisioning of high cell capacity to the satisfaction of fairness and QoS requirements. Considering the relevance of the topic, we believe that a detailed classification and description of design approaches would be highly beneficial for the audience of this journal.

To this end, in this paper we overview the key facets of LTE scheduling. The downlink shared channel is taken as a reference, but most of the considerations pointed out in this work remain valid also for the uplink direction. We discussed and classified a wide range of techniques according to their working rationales. In addition, a survey on the current research status in the field is presented along with a performance comparison of the most well-known techniques. For each of them, we detail the main properties and summarize learned lessons so far from its adoption.

The rest of the paper is organized as follows. In Sec. II an overview of LTE system is provided, focusing on the architectural and physical aspects that are relevant from the point of view of the resource sharing problem. In Sec. III, the key issues regarding the design of a packet scheduler for LTE are presented and some possible approaches to solve the problem are shown. A comprehensive survey as well as a detailed classification of the most known strategies is reported in Sec. IV. Moreover, in this section, performance comparison of the most representative solutions is carried out through system-level simulations and, when it is possible, by comparing results with the ones already present in literature. Sec. V depicts new research directions and open design challenges. Finally, in Sec. VI we draw the conclusions, with particular attention to the lessons learned with this work.

II. Overview on LTE Networks

A deep understanding of radio propagation issues in cellular networks has driven the LTE standardization process towards the use of advanced and high performance techniques [4].

In fact, in order to efficiently support the current high variety of applications, LTE networks have been conceived with very ambitious requirements that strongly overtake features of 3G networks, mainly designed for classic voice services [1]. LTE aims, as minimum requirement, at doubling the spectral efficiency of previous generation systems and at increasing the network coverage in terms of bitrate for cell-edge users. Moreover, several new performance targets, with respect to other technologies, have been addressed during the standardization phase, spanning from increased data rates (i.e., the allowed peak data rates for the downlink and uplink are equal to 100 Mbps and 50 Mbps, respectively) to the support of very high user mobility. To make LTE networks highly flexible for a worldwide market, a variable bandwidth feature, that gives to network operators the possibility to throttle the bandwidth occupation between 1.4 and 20 MHz, is also included. Among all these performance targets, summarized in Tab. I, the most important novelty introduced by LTE specifications is the enhanced QoS support by means of new sophisticated RRM techniques.

TABLE I

Main LTE Performance Targets

Peak Data Rate	- Downlink: 100 Mbps - Uplink: 50 Mbps
Spectral Efficiency	2 - 4 times better than 3G systems
Cell-Edge Bit-Rate	Increased whilst maintaining same site locations as deployed today
User Plane Latency	Below 5 ms for 5 MHz bandwidth or higher
Mobility	- Optimized for low mobility up to 15 km/h - High performance for speed up to 120 km/h - Maintaining connection up to 350 km/h
Scalable Bandwidth	From 1.4 to 20 MHz
RRM	- Enhanced support for end-to-end QoS - Efficient transmission and operation of higher layer protocols
Service Support	- Efficient support of several services (e.g., web-browsing, FTP, video-streaming, VoIP) - VoIP should be supported with at least a good quality as voice traffic over the UMTS network

This section gives an overview of the main LTE features. First of all the system architecture is described, including the main aspects of the protocol stack. The OFDMA air interface is also illustrated, with particular focus on issues related to scheduling. Finally, a comprehensive description of the fundamental RRM procedures is provided.

A. System Architecture and Radio Access Network

The LTE system is based on a flat architecture, known as the “Service Architecture Evolution”, with respect to the 3G systems [5]. This guarantees a seamless mobility support and a high speed delivery for data and signaling. As depicted in Fig. 1, it is made by a core network, namely the “Evolved Packet Core”, and a radio access network, namely the Evolved-Universal Terrestrial Radio Access Network (E-UTRAN).

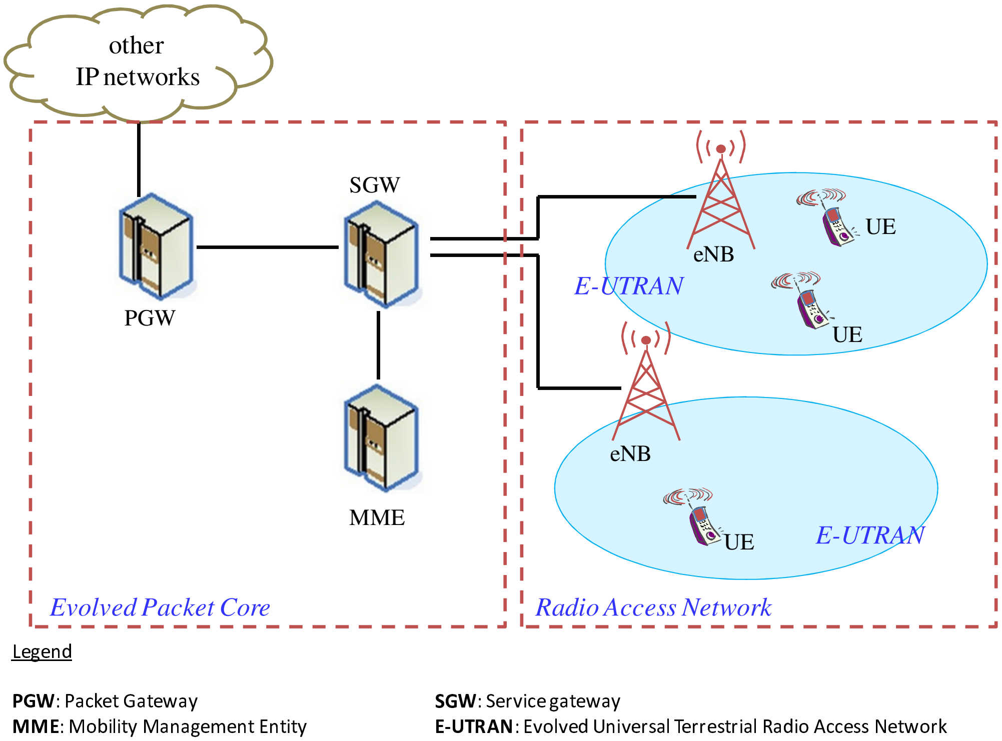


Fig. 1. The Service Architecture Evolution in LTE network.

The Evolved Packet Core comprises the Mobility Management Entity (MME), the Serving Gateway (SGW), and the Packet Data Network Gateway (PGW). The MME is responsible for user mobility, intra-LTE handover, and tracking and paging procedures of User Equipments (UEs) upon connection establishment. The main purpose of the SGW is, instead, to route and forward user data packets among LTE nodes, and to manage handover among LTE and other 3GPP technologies. The PGW interconnects LTE network with the rest of the world, providing connectivity among UEs and external packet data networks.

The LTE access network can host only two kinds of node: the UE (that is the end-user) and the eNB. Note that eNB nodes are directly connected to each other (this obviously speeds up signaling procedures) and to the MME gateway. Differently from other cellular network architectures, the eNB is the only device in charge of performing both radio resource management and control procedures on the radio interface.

To the purpose of our dissertation, in what follows we will focus only on those aspect of the LTE architecture that are strictly related to the radio resource management, such as the radio access network, the radio bearer management, and the physical layer design. We suggest the reader to refer to [4] and [5] for more details on entities and physical interfaces of the core network.

B. Radio Bearer Management and Protocol Stack

A radio bearer is a logical channel established between UE and eNB. It is in charge of managing QoS provision on the E-UTRAN interface. When an UE joins the network, a *default bearer* is created for basic connectivity and exchange of control messages. It remains established during the entire lifetime of the connection. *Dedicated bearers*, instead, are set up every time a new specific service is issued. Depending on QoS requirements, they can be further classified as Guaranteed bit-rate (GBR) or non-guaranteed bit rate (non-GBR) bearers [6].

In this context, the general definition of QoS requirements is translated in variables that characterize performance experienced by users [3]. A set of QoS parameters is therefore associated to each bearer depending on the the application data it carries, thus enabling differentiation among flows. To this aim, during LTE standardization phase, several classes of QoS services have been identified through QoS Class Identifiers (QCIs) [7], i.e., scalar values used as a reference for driving specific packet forwarding behaviors. As comprehensively pictured in Tab. II, each QoS class is characterized by its resource type (GBR or not-GBR), a priority level, the maximum admitted delivery delay, and the acceptable packet loss rate. The RRM module translates QoS parameters into scheduling parameters, admission policies, queue management thresholds, link layer protocol configurations, and so on. LTE specifications introduced also specific protocol entities:

- the Radio Resource Control, which handles the establishment and management of connections, the broadcast of system information, the mobility, the paging procedures, and the establishment,

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), chat, FTP, P2P file sharing
9	non-GBR	9	300	10^{-6}	

reconfiguration and management of radio bearers [8];

- the Packet Data Control Protocol, which operates header compression of upper layers before the MAC enqueueing [9];
- the Radio Link Control (RLC), which provides interaction between the radio bearer and the MAC entity [10];
- the MAC, which provides all the most important procedures for the LTE radio interface, such as multiplexing/demultiplexing, random access, radio resource allocation and scheduling requests [11].

C. Physical layer

LTE has been designed as a highly flexible radio access technology in order to support several system bandwidth configurations (from 1.4 MHz up to 20 MHz). Radio spectrum access is based on the Orthogonal Freq. Division Multiplexing (OFDM) scheme. In particular, Single Carrier Freq. Division Multiple Access (SC-FDMA) and OFDMA are used in uplink and downlink directions, respectively. Differently from basic OFDM, they allow multiple access by assigning sets of sub-carriers to each individual user. OFDMA can exploit sub-carriers distributed inside the entire spectrum, whereas SC-

FDMA can use only adjacent sub-carriers. OFDMA is able to provide high scalability, simple equalization, and high robustness against the time-frequency selective nature of radio channel fading. On the other hand, SC-FDMA is used in the LTE uplink to increase the power efficiency of UEs, given that they are battery supplied.

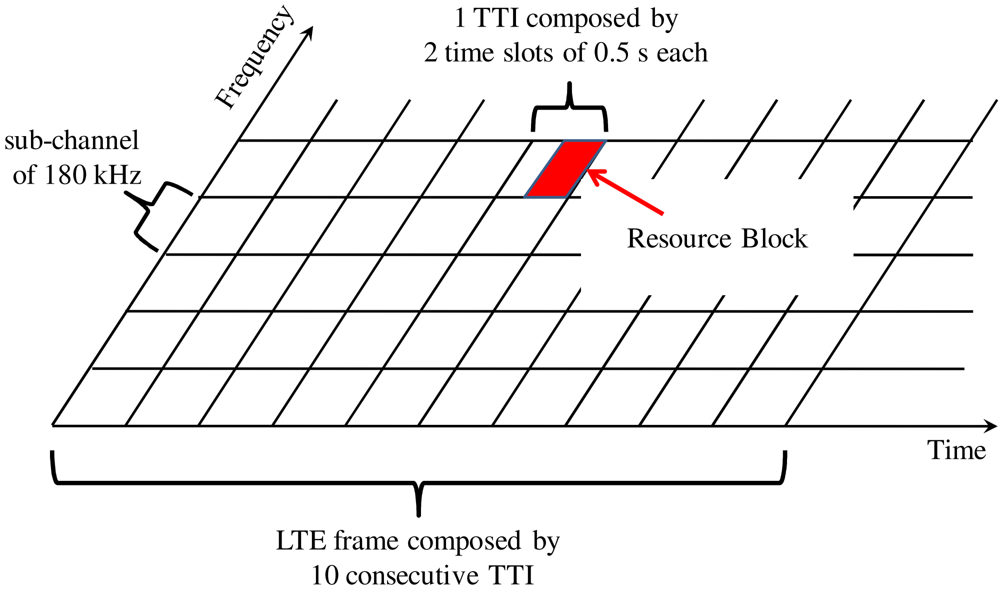


Fig. 2. Time-Frequency radio resources grid.

Radio resources are allocated into the time/frequency domain [12] (see Fig. 2). In particular, in the time domain they are distributed every Transmission Time Interval (TTI), each one lasting 1 ms. The time is split in frames, each one composed of 10 consecutive TTIs. Furthermore, each TTI is made of two time slots with length 0.5 ms, corresponding to 7 OFDM symbols in the default configuration with short cyclic prefix. In the frequency domain, instead, the total bandwidth is divided in sub-channels of 180 kHz, each one with 12 consecutive and equally spaced OFDM sub-carriers. A time/frequency radio resource spanning over two time slots in the time domain and over one sub-channel in the frequency domain is called Resource Block (RB) and corresponds to the smallest radio resource unit that can be assigned to an UE for data transmission. As the sub-channel size is fixed, the number of RBs varies according to the system bandwidth configuration (e.g., 25 and 50 RBs for system bandwidths of 5 and 10 MHz, respectively).

As described in [12], the LTE radio interface supports two types of frame structure, related to the different duplexing schemes. Under Frequency Division Duplex, the bandwidth is divided in two parts, allowing simultaneous downlink and uplink data transmissions, and the LTE frame is composed of 10 consecutive identical sub-frames. For Time Division Duplex (TDD), instead, the LTE frame is divided into two consecutive half-frames, each one lasting 5 ms. In this case, several frame configurations allow different balance of resources dedicated for downlink or uplink transmission. As schematized in Tab. III, in fact, a frame could be characterized by a very high presence of downlink (e.g., configuration 5) or uplink sub-frames (e.g., configuration 0). Note that there is in all the configurations a special downlink sub-frame that handles synchronization information. The selection of the TDD frame configuration is performed by the RRM module, and, as expected, it is based on the proportion between downlink and uplink traffic loading the network.

TABLE III
TDD FRAME CONFIGURATIONS

	sub-frame number									
	1^{st} half frame					2^{nd} half frame				
configuration number	0	1	2	3	4	5	6	7	8	9
0	D	S	U	U	U	D	S	U	U	U
1	D	S	U	U	D	D	S	U	U	D
2	D	S	U	D	D	D	S	U	D	D
3	D	S	U	U	U	D	D	D	D	D
4	D	S	U	U	D	D	D	D	D	D
5	D	S	U	D	D	D	D	D	D	D
6	D	S	U	U	U	D	S	U	U	D

D = downlink sub-frame; U = uplink sub-frame; S = Special sub-frame.

D. Radio Resource Management

Besides resource distribution, LTE makes massive use of RRM procedures such as link adaptation, HARQ, Power Control, and CQI reporting. They are placed at physical and MAC layers, and strongly interact with each other to improve the usage of available radio resources. Fig. 3 shows an overall description of the main features and their interaction in terms of both data exchange and signaling.

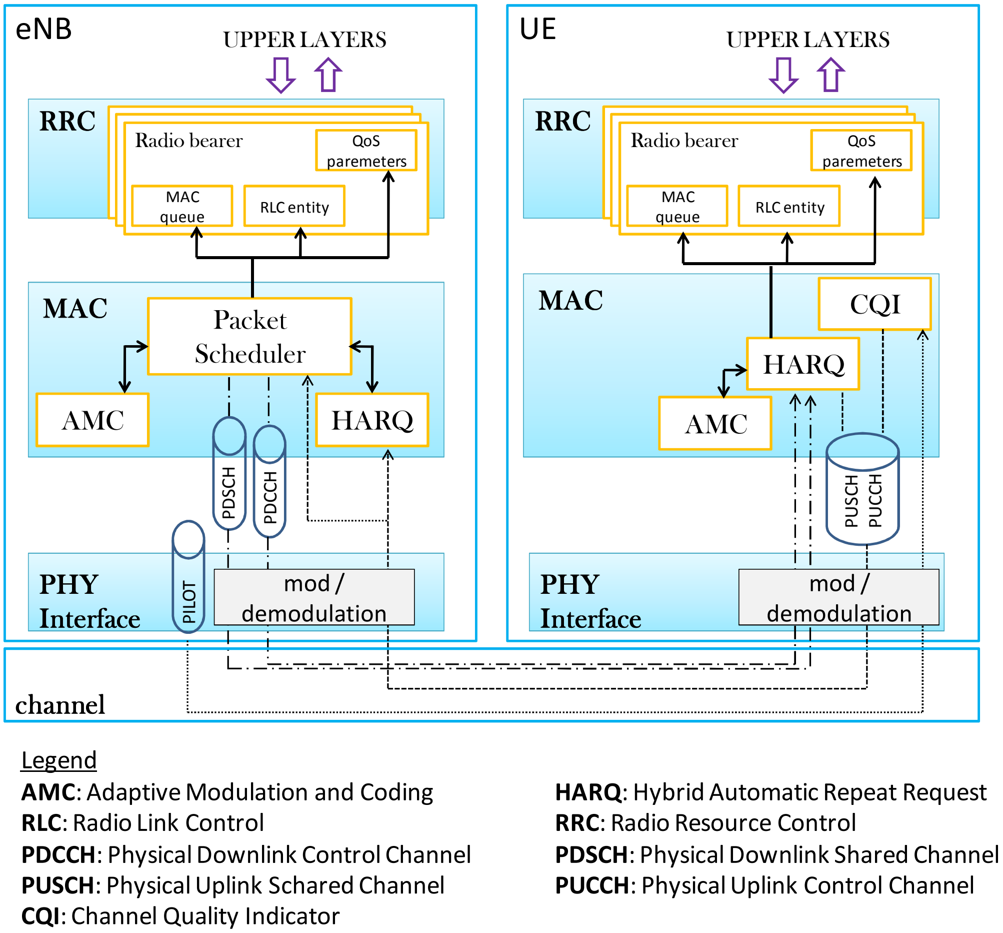


Fig. 3. Interaction of the main RRM features.

1) CQI reporting: The procedure of the CQI reporting is a fundamental feature of LTE networks since it enables the estimation of the quality of the downlink channel at the eNB. Each CQI is calculated as a quantized and scaled measure of the experienced Signal to Interference plus Noise Ratio (SINR). The main issue related to CQI reporting methods is to find a good tradeoff between a precise channel quality estimation and a reduced signaling overhead. CQI reporting, however, is out of the scope of this paper; we suggest to refer to [13] for further details on the topic.

2) AMC and Power Control: The CQI reporting procedure is strictly related to the AMC module, which selects the proper Modulation and Coding Scheme (MCS) trying to maximize the supported throughput with a given target Block Error Rate (BLER) [5]. In this way, a user experiencing a higher SINR will be served with higher bitrates, whereas a cell-edge user, or in general a user experiencing bad channel conditions, will maintain active connections, but at the cost of a lower throughput. It is important to note that the number of allowed modulation and coding schemes is limited. Hence, the system throughput is upper-bounded: over a certain threshold an increase in the SINR does not bring to any throughput gain. This is why the AMC often works together with the power control module. In fact, as well-known, power control is a dynamic procedure that adjusts transmission power on the radio-link in order to compensate for variations of the instantaneous channel conditions [5]. The aim of these adjustments is twofold: to save energy while maintaining a constant bitrate (i.e., power reduction) or to increase the bitrate selecting a higher MCS (i.e., power boosting); in both cases, the goal is obtained keeping the expected BLER below a target threshold.

3) Physical Channels: Downlink data are transmitted by the eNB over the Physical Downlink Shared Channel (PDSCH). As its name states, it is shared among all the users in the cell as, in general, no resource reservation is performed in LTE. Transmission of PDSCH payloads is allowed only in given portion of the spectrum and in certain time interval according to a scheme. An example of the typical structure of the LTE downlink subframe is shown in Fig. 4.

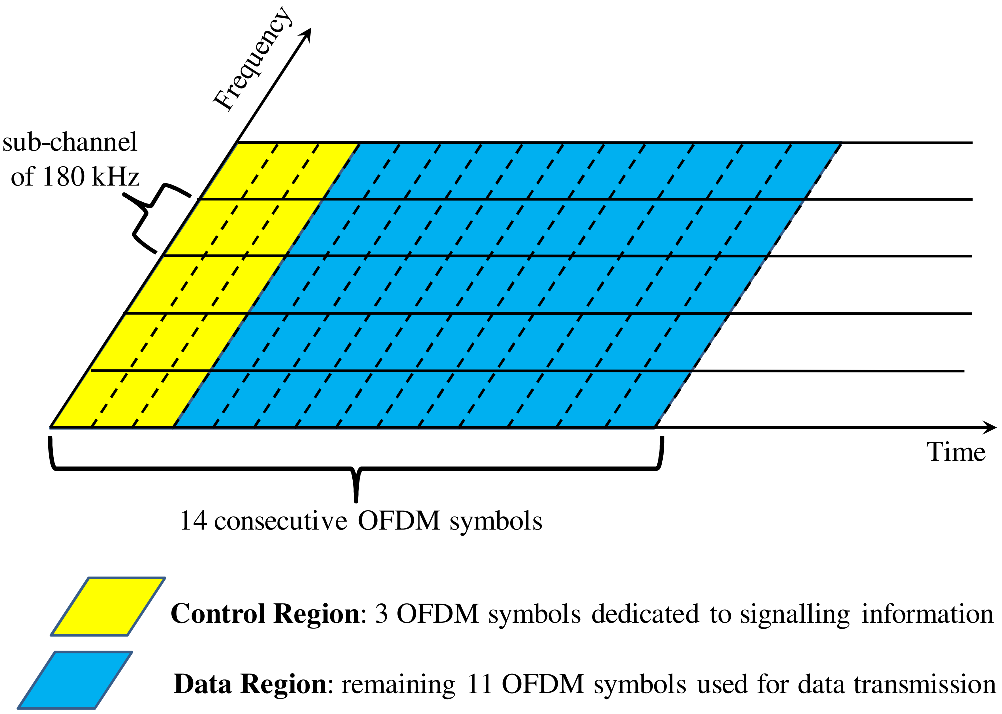

Frequency
sub-channel
of 180 kHz
Time
14 consecutive OFDM symbols
Control Region: 3 OFDM symbols dedicated to signalling information
Data Region: remaining 11 OFDM symbols used for data transmission

Fig. 4. Time/Frequency structure of the LTE downlink subframe in the 3 MHz bandwidth case (example with 3 OFDM symbols dedicated to control channels).

It is based on the time/frequency grid previously explained and depicted in Fig. 2. It defines how radio resources are utilized by upper layers. Downlink data and signaling information are time multiplexed within the subframe. In details, control channels occupy, in each TTI, the first 1 to 3 OFDM symbols over the 14 available. Consequently, data transmission is allowed during the remaining time. In Fig. 4, it is possible to observe the case of a LTE downlink subframe when 3 OFDM symbols are dedicated to control messages.

Downlink control signaling is carried by three physical channels [14]; the most important from a scheduling perspective is the Physical Downlink Control Channel (PDCCH), which carries assignments for downlink resources and uplink grants, including the used MCS.

It is worth to observe the influence that the control overhead has on the downlink performance [15], because every TTI a significant amount of radio resources is used for sending such information. PDCCH, in particular, carries a message known as Downlink Control Information (DCI), that conveys various pieces of information depending on the specific system configuration. Several DCI formats are defined depending on the type of information they carry and on their packet size [16]. The impact of the control overhead on the scheduler design is object of study in section III-B.

In the uplink direction, two physical channels are defined: the Physical Uplink Control Channel (PUCCH) and the Physical Uplink Shared Channel (PUSCH) [12]. Due to single carrier limitations, simultaneous transmission on both channels is not allowed. When no uplink data transmission is foreseen in a given TTI, PUCCH is used to transmit signaling (e.g., ACK/NACK related to downlink transmissions, downlink CQI, and requests for uplink transmission). On the other hand, PUSCH carries the uplink control signals when the UE has been scheduled for data transmission; in this case data and different control fields, such as ACK/NACK and CQI, are time multiplexed in the PUSCH payload.

4) HARQ: It is the retransmission procedure at MAC layer, based on the use of the well-known stop-and-wait algorithm [4]. This procedure is simply performed by eNB and UE through the exchange of ACK/NACK messages. A NACK is sent over the PUCCH when a packet transmitted by the eNB is unsuccessfully decoded at the UE. In this case the eNB will perform a retransmission, sending the same copy of the lost packet. Then, the UE will try to decode the packet combining the retransmission with the original version, and will send an ACK message to the eNB upon a successfully decoding.

III. Scheduling in LTE systems

Multi-user scheduling is one of the main feature in LTE systems because it is in charge of distributing available resources among active users in order to satisfy their QoS needs.

As explained in the previous section, in fact, the data channel (i.e., the PDSCH) is shared among the users, meaning that portions of the spectrum should be distributed every TTI among them. Packet schedulers (for both the downlink and the uplink) are deployed at the eNB, and, since OFDMA ideally provides no inter-channel interference. They work with a granularity of one TTI and one RB in the time and frequency domain, respectively.

Resource allocation for each UE is usually based on the comparison of per-RB metrics: the k-th RB is allocated to the j-th user if its metric $m_{j,k}$ is the biggest one, i.e., if it satisfies the equation:

$$m_{j,k} = \max_{i}\{m_{i,k}\} \ . \tag{1}$$

These metrics can be somehow interpreted as the transmission priority of each user on a specific RB. Based on the desired performance requirement, their computation is usually evaluated starting from information related to each flow and useful to drive the allocation decision:

- *Status of transmission queues*: the status of transmission queues at UEs could be used for minimizing packet delivery delays (e.g., the longer the queue, the higher the metric).
- *Channel Quality*: reported CQI values could be used to allocate resources to users experiencing better channel conditions (e.g., the higher the expected throughput, the higher the metric).
- *Resource Allocation History*: information about the past achieved performance can be used to improve fairness (e.g., the lower the past achieved throughput, the higher the metric).
- *Buffer State*: receiver-side buffer conditions might be used to avoid buffer overflows (e.g., the higher the available space in the receiving buffer, the higher the metric).
- *Quality of Service Requirements*: the QCI value associated to each flow might be used to drive specific policies with the aim of meeting QoS requirements.

Every TTI the scheduler performs the allocation decision valid for the next TTI and sends such information to UEs using the PDCCH (see previous section). DCI messages in the PDCCH payload inform UEs about RBs allocated for data transmission on the PDSCH in the downlink direction. Moreover, DCI messages are used to inform users about the dedicated radio resources for their data transmission on the PUSCH in the uplink direction.

In this work we mainly focus on the downlink, but most of the considerations hold for the uplink as well. In section III-B differences between the two directions, mainly concerning the characteristics of OFDMA and SC-FDMA, are pointed out.

A relevant importance in LTE schedulers is assigned to the “channel sensitivity” concept. The basic idea is to schedule transmission for UEs that, at the current time and on a given frequency, are experiencing

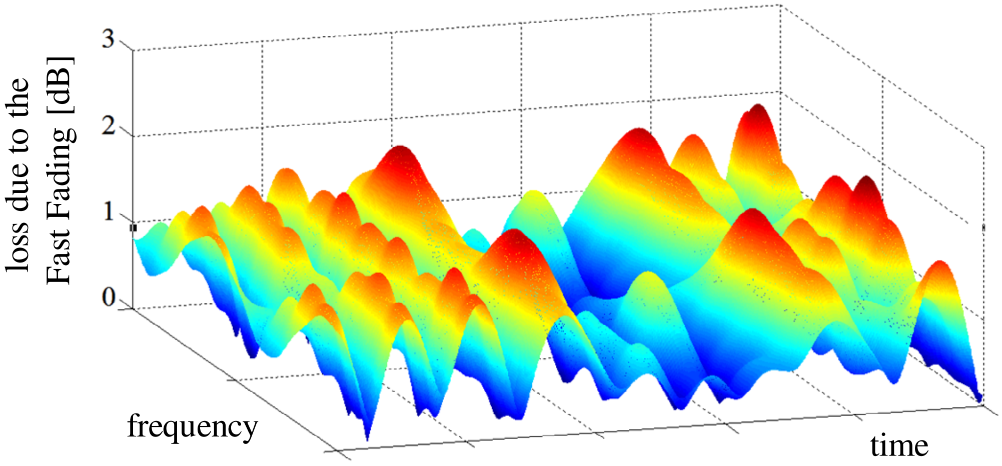


Fig. 5. Loss due to the fast fading in the time-frequency domain.

"good" channel conditions based on the selected metric. This approach, also known as Frequency Domain Packet Scheduler (FDPS), counteracts the time-varying and frequency-selective nature of the wireless channel (as an example, Fig. 5 shows the loss due to the fast fading effect).

Furthermore, the characteristic of the fast fading to be independent on users can be exploited by allocation procedures, obtaining what is usually addressed as "multi-user diversity" gain. In [17], authors show that the overall system capacity grows with the number of users; therefore, we can define the multi-user diversity gain as the advantage, in terms of system capacity, of serving more than one user. In fact, in a scenario with many users experiencing independent fading effects, the probability to find a user with good channel conditions at a given time is very high. The advantage of this behavior is twofold: it enables the transmission when high data rates are achievable (i.e., under good channel conditions the AMC module will select a more effective MCS) and, at the same time, it is naturally immune to frequency-selective fading effects (i.e., a user experiencing very bad channel condition will not be served). Nevertheless, as we will describe in this section, multi-user diversity gain appears to be upper bounded, and this should be taken into account during the design phase. As matter of fact, increasing the number of users in the system, also the control overhead increases.

Fig. 6 represents the main RRM modules that interact with the downlink packet scheduler. The whole process can be divided in a sequence of operations that are repeated, in general, every TTI:

1) each UE decodes the reference signals, computes the CQI, and sends it back to the eNB.
2) The eNB uses the CQI information for the allocation decisions and fills up a RB "allocation mask".

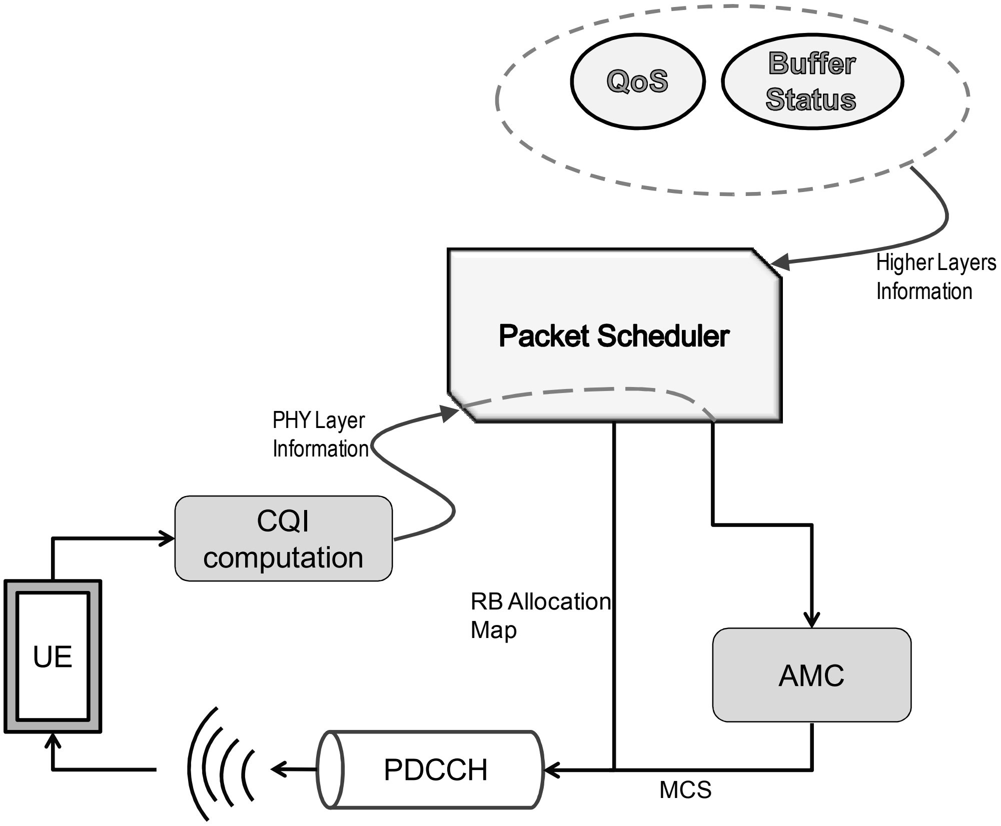


Fig. 6. Simplified model of a packet scheduler.

3) The AMC module selects the best MCS that should be used for the data transmission by scheduled users.
4) The information about this users, the allocated RBs, and the selected MCS are sent to the UEs on the PDCCH.
5) Each UE reads the PDCCH payload and, in case it has been scheduled, accesses to the proper PDSCH payload.

The outlined working flow slightly differs in the uplink direction as the eNB does not need any additional information on the uplink channel quality .

In the rest of this section we focus on the key aspects that should be considered when designing a dynamic resource sharing scheme for LTE, and on the main issues that arise in realistic scenarios. Finally, we discuss basics of some persistent techniques, as an alternative to dynamic ones.

A. Key Design Aspects

Differences among resource allocation strategies are mainly based on the trade-off between decision optimality and computational complexity. Hereafter, we present a list of the main design factors that always should be taken into account before defining an allocation policy for LTE.

1) Complexity and Scalability: A LTE packet scheduler works with a time granularity of 1 ms: it has to take allocation decisions every TTI. Low complexity and scalability are therefore fundamental requirements for limiting processing time and memory usage. Finding the best allocation decision through complex and non-linear optimization problems or through an exhaustive research over all the possible combinations would be too expensive in terms of computational cost and time [17]. For this reason, FDPS decisions are usually based on the computation of *per-RB* metrics for each user. In this manner, complexity reduction is achieved because each RB is allocated to the user with the highest metric independently of other RBs. Let N and R be the number of active users in the current TTI and the number of available RBs, respectively; the scheduler has to calculate $M = N \cdot R$ metrics every TTI. This assures the scalability requirement thanks to the linear dependence on the number of resource blocks and users.

2) Spectral Efficiency: Effective utilization of radio resources is one of the main goals to be achieved. To this aim, several types of performance indicators can be considered: for instance, a specific policy could aim at maximizing the number of users served in a given time interval or, more commonly, the spectral efficiency (expressed in bit/s/Hz) by always serving users that are experiencing the best channel conditions. One of the most used efficiency indicators is the *user goodput*, that is a measure of the actual transmission data rate without including layer two overheads and packet retransmissions due to physical errors.

3) Fairness: A blind maximization of the overall cell throughput surely enables effective channel utilization in terms of spectral efficiency, but also brings to very unfair resource sharing among users. Fairness is therefore a major requirement that should be taken into account to guarantee minimum performance also to the cell-edge users (or in general to users experiencing bad channel conditions). As we will explain in section IV, this issue can be overcome by considering, inside the selected metric, a measure of the past service level felt by each user; making this, it is possible to counteract the unfair sharing of the channel capacity.

4) QoS Provisioning: As well-known, QoS provisioning is very important in next generation mobile networks. It is a major feature in all-IP architectures. As previously mentioned (see section II-D), LTE maps QoS constrained flows to dedicated radio bearers that, depending on their QCIs, enable special RRM procedures. QoS constraints may vary depending on the application and they are usually mapped into some parameters: minimum guaranteed bitrate, maximum delivering delay, and packet loss rate. Thus, it is important to define QoS-aware schedulers.

B. Practical limitations in real LTE systems

Several aspects of LTE deployment in real environment may impact on the choice of the best allocation schemes to be adopted.

1) Uplink Limitations: We consider, at first, the impact of the different assumptions on downlink and uplink radio access methods. In downlink, due to the use of OFDMA, the scheduler can fill out the RB allocation mask without ordering constraints. Instead, SC-FDMA method, used for uplink transmissions, allows the UEs to transmit only in a single carrier mode. Therefore, the scheduler for the uplink has limited degrees of freedom: it has to allocate contiguous RBs to each user without the possibility of choice among the best available ones. Further details related to the allocation in the uplink can be found in [18][19] and in the references therein.

2) Control Overhead: As already mentioned in section II-C, the PDCCH, which carries DCI messages, is time-multiplexed with the data-channel occupying a variable number of OFDM symbols (up to 3). As a consequence, the amount of resources dedicated to the PDCCH is limited, thus decreasing the degrees of freedom for the downlink scheduler. PDCCH overhead can be reduced using specific RRM procedures [15], either by lowering a priori the number of users to be served (and hence the number of DCI messages) or by utilizing low bitrate DCI formats [16]. The latter solution poses some constraints on the construction of the RB allocation mask (for example allowing the assignment of only contiguous RBs to the same UE). Detailed information on characteristics of DCI messages can be found in [12].

3) Limitations on the Multi-User Diversity Gain: An important limitation for LTE resource sharing algorithms can derive from the availability, at the eNB, of accurate channel quality measurements. For the downlink, UEs are in charge of sending CQIs to the serving eNB as described in section II-D. For the uplink, instead, the eNB may use reference signals transmitted by the UEs to estimate uplink channel quality. The chosen CQI reporting scheme has great impact on the multi-user diversity gain, as it defines the time and frequency resolution of the channel quality information available at the scheduler [20]. This consideration can be easily confirmed observing, as representative example, that in the case of wideband CQI reporting scheme (i.e., a single CQI value is transmitted for the entire spectrum) the sensitivity to channel conditions would become totally useless and the multi-user diversity gain would go to zero. Thus, it appears clear how, in such situations, the adoption of sophisticated techniques would not bring to any gain.

Furthermore, efficiency is limited because multi-user diversity gain results to be upper bounded; that is, increasing indefinitely the number of users competing for allocation does not bring to any gain over a certain threshold [17]. According to [21], moreover, it actually exists an optimal number of candidate users depending on the system bandwidth. We will show in next sections that this characteristics can be fruitfully exploited for reducing the computational complexity of the selected strategy.

4) Energy Consumption: Energy saving is a required feature for mobile terminals, and in LTE it is achieved through Discontinuous Reception (DRX) methods. The basic idea of DRX is to allow an UE to save energy turning off its radio equipment when there are no data transmissions [22], as in typical alternation of on/off periods (DRX cycles). In LTE, moreover, DRX can be enabled in presence of active flows as well, allowing off periods between packet burst transmissions. Such operations require the knowledge of the activation of DRX cycles. In principle, if the network could predict when a certain user needs to transmit/receive data, an UE would only need to be awake at such specific time instants. On the other hand, in absence of a precise coupling between scheduling and DRX procedures, it may occur that resources are allocated to an UE during its off periods with consequent loss of data and waste of bandwidth. In next sections we will demonstrate how persistent solutions proposed in literature face this problem. More details on DRX procedures and related performances can be found in [22] and [23].

C. Persistent and semi-persistent scheduling

As already pointed out, dynamic frequency domain strategies have the main benefit of exploiting multi-user diversity gain, but this comes at the cost of increased control overhead, due to the need of forwarding DCI messages to scheduled users every TTI. For this reason, especially in scenarios with high traffic load, the limited amount of radio resources dedicated to control information transmission can become a bottleneck, with consequent degradation of QoS provisioning capabilities [24].

To face this problem, persistent solutions have been investigated [25]. According to [26], in these approaches, the control overhead is used to preallocate to the same UE certain RBs over a time sequence, distinguishing between active and inactive periods. Under this scheme, once the eNB has informed a user on the persistent allocation, the interested UE will know in advance the specific TTI/RB couples where it should decode PDSCH payloads (or, for the uplink direction, it should transmit PUSCH payloads) with no need of any additional PDCCH overhead.

In Fig. 7, a graphical example of a persistent resource allocation is given. Here, for instance, user 1 has resources allocated every three TTIs and user 2 every five TTIs, always on the same sub-carriers.

Besides the impossibility to exploit channel sensitivity, a major drawback of persistent approach is that

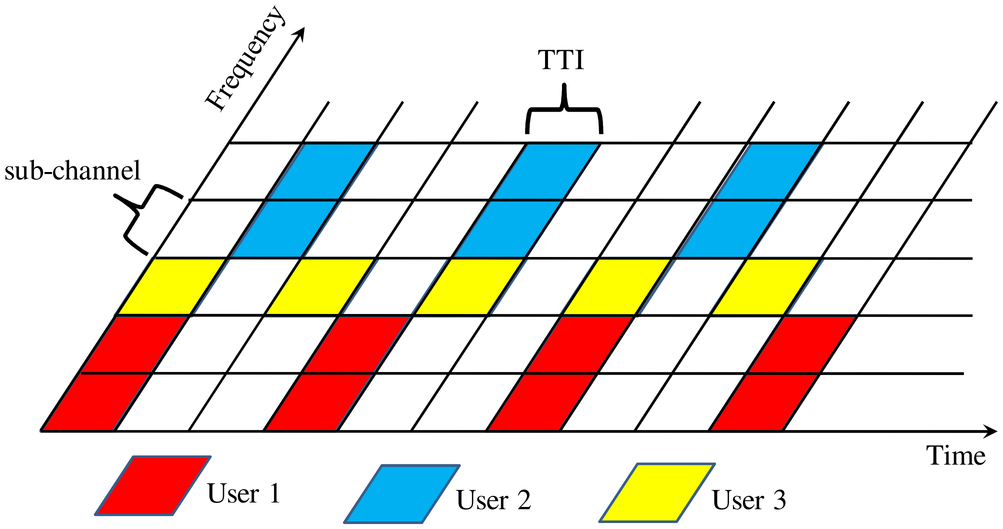


Fig. 7. Example of persistent scheduling allocation.

it does not appear suitable for supporting HARQ. In practice, a single packet needs on average more than a single transmission to be correctly decoded at the receiver, and the number of retransmissions varies for each user, depending on many factors such as the channel quality, the user mobility, and the perceived inter-cell interference.

For these reasons, semi-persistent algorithms are being deeply investigated as a trade-off among static persistent approaches and fully dynamic ones, especially if conceived for VoIP traffic (see Sec. IV-D).

IV. Scheduling Strategies for LTE Downlink

In this section, we will illustrate different allocation strategies introduced for LTE systems, highlighting pros and cons related to each solution. Moreover, we will compare performance of some of them through system-level simulations. The simulated algorithms are well-known schemes widely used in literature and represent, from our point of view, the most interesting approaches that can be used for easing the understanding of several aspects connected to the scheduling design. They differ in terms of input parameters, objectives, and service targets. To simplify the reading of the survey, we have classified them in five groups of strategies: (i) channel-unaware; (ii) channel-aware/QoS-unaware; (iii) channel-aware/QoS-aware; (iv) semi-persistent for VoIP support; and (v) energy-aware.

We will first illustrate channel unaware approaches, which have been historically adopted to face fairness, flow priorities, and deadline expiration in all packet switching networks. They will help us to introduce some basic parameters such as instantaneous data rate, past average throughput, head of line packet delay, target delay, target packet loss rate, and more, whose definitions are given in Tab. IV. These parameters are widely used for the definition of metrics. Channel aware schedulers (with and without QoS support) are then introduced and deeper analyzed, because more suitable for wireless environments. Finally, we will illustrate some semi-persistent solutions, conceived in literature for VoIP flow support, and energy-aware schemes. Let us point out that such classification can also be seen as a historical evolution of resource allocation techniques. In fact, while the channel unaware schemes were firstly designed to work in operating systems and cabled networks, the channel awareness requirement became a must in wireless communications. A further step was done with the introduction of all-IP architectures (as in LTE network) to provide strong QoS differentiation and thus requiring the development of more complex QoS-aware strategies.

A. Channel-unaware Strategies

Firstly introduced in wired networks [27], channel unaware strategies are based on the assumption of time-invariant and error-free transmission media. While their direct application in LTE is not realistic, they are typically used jointly with channel-aware approaches to improve system performance.

1) First In First Out: The simplest case of channel unaware allocation policy serves users according to the order of resource requests, exactly like a First In First Out (FIFO) queue. We can translate this behavior in LTE expressing the metric of the i-th user on the k-th RB as

$$m_{i,k}^{FIFO} = t - T_i \ , \tag{2}$$

where t is the current time and T_i is the time instant when the request was issued by the i-th user.

This technique is very simple, but both inefficient and unfair.

2) Round Robin: It performs fair sharing of time resources among users. Round Robin (RR) metric is similar to the one defined for FIFO with the difference that, in this case, T_i refers to the last time when the user was served. In this context, the concept of fairness is related to the amount of time in which the channel is occupied by users. Of course, this approach is not fair in terms of user throughput, that, in wireless systems, does not depend only on the amount of occupied resources, but also on the experienced channel conditions. Furthermore, the allocation of the same amount of time to users with very different application-layer bitrates is not efficient.

3) Blind Equal Throughput: Throughput Fairness can be achieved with Blind Equal Throughput (BET) which stores the past average throughput achieved by each user and uses it as metric [28]. In this case

TABLE IV

NOTATIONS USED FOR SCHEDULING METRICS

Expression	**Meaning**
$m_{i,k}$	Generic metric of the i-th user on the k-th RB
$r^i(t)$	Data-rate achieved by the i-th user at time t
$\overline{R^i}(t)$	Past average throughput achieved by the i-th user until time t
$\overline{R^i_{sch}}(t)$	Average throughput achieved by data flow of the i-th user when scheduled
$D_{HOL,i}$	Head of Line Delay, i.e., delay of the first packet to be transmitted by the i-th user
τ_i	Delay Threshold for the i-th user
δ_i	Acceptable packet loss rate for the i-th user
$d^i(t)$	Wideband Expected data-rate for the i-th user at time t
$d^i_k(t)$	Expected data-rate for the i-th user at time t on he k-th RB
Γ^i_k	Spectral efficiency for the i-th user over the k-th RB

the metric (for the i-th user) is calculated as (see table IV for notation meanings):

$$m_{i,k}^{BET} = 1/\overline{R^i}(t-1) \tag{3}$$

with

$$\overline{R^i}(t) = \beta\overline{R^i}(t-1) + (1-\beta)r^i(t) \tag{4}$$

where $0 \leq \beta \leq 1$.

Thanks to its interesting properties, this metric is widely used in most of the state of the art schedulers. First of all, it is easy to note that every TTI, BET allocates resources to flows that have been served with lower average throughput in the past. Under this allocation policy, the user experiencing the lowest throughput, performs, in practice, resource preemption: he will be served as long as he does not reach the same throughput of other users in the cell. In this way, users with bad channel conditions are allocated more often that others, with a consequent fairness improvement. The factor $\overline{R^i}(t)$, that represents the past average throughput experienced by the i-th user at time t, is calculated as a moving average and it is updated every TTI for each user. Its role will be better explained below, where we will also highlight how BET metric assumes a great importance for guaranteeing fairness in channel-aware schemes.

4) Resource Preemption: Fairness is not always required, or at least not for all users. Therefore, several type of priority schemes can be defined. The simplest approach is resource preemption of some high priority users (or classes of users). The idea is that transmission queues are grouped in several priority classes, and a queue belonging to a given class cannot be served until all queues having higher priorities are empty. This approach can be fruitfully exploited to handle the differentiation among QoS (high priority) and non-QoS flows, provided that the scheduler implements some techniques to avoid the starvation of low-priority applications.

5) Weighted Fair Queuing: An alternative way to introduce priorities avoiding the possibility of starvation is through the usage of an approximation of the well-known Weighted Fair Queuing (WFQ) approach. In this case, a specific weight (w_i) is associated to the i-th user (or class of users) and then it is used to correct user-specific RR metric as

$$m_{i,k}^{WFQ} = w_i \cdot m_{i,k}^{RR} \tag{5}$$

where $m_{i,k}^{RR}$ is the RR specific metric for the i-th user.

In this way, resources are shared accordingly to the proportion among the weights (the higher the weights, the higher the allocated resources), but no starvation is possible because the RR metric would control that the waiting time of a given user does not indefinitely grow. In general, this approach holds if also the BET metric is taken into account.

6) Guaranteed Delay: Guaranteed delay services, in particular, require that each packet has to be received within a certain deadline to avoid packet drops. This goal can be accomplished by including into the metric information about the specific packet timing, that is both the time instant when the packet was created and its deadline. Earliest Deadline First (EDF) and Largest Weighted Delay First (LWDF) are two policies, defined mostly for real-time operating system and wired networks [27], [29], [30], that aim at avoiding deadline expiration. EDF policy, as its name itself clearly states, schedules the packet with the closest deadline expiration. Its metric can be easily expressed as:

$$m_{i,k}^{EDF} = \frac{1}{(\tau_i - D_{HOL,i})} \; . \tag{6}$$

Intuitively, the more the head of line delay approaches the expiration time, the more the user metric increases.

On the other hand, LWDF metric is based on the system parameter δ_i, representing the acceptable probability for the i-th user that a packet is dropped due to deadline expiration; the metric is then calculated as:

$$m_{i,k}^{LWDF} = \alpha_i \cdot D_{HOL,i} \tag{7}$$

where α_i is given by

$$\alpha_i = -\frac{\log \delta_i}{\tau_i}. \tag{8}$$

The role that α_i plays in this metrics is quite interesting. Given two flows with equal head of line delay, in fact, α_i weights the metric so that the user with strongest requirements in terms of acceptable loss rate and deadline expiration will be preferred for allocation. In section IV-C a modified channel-aware version of the LWDF for QoS provisioning will be described.

7) General considerations on channel-unaware strategies: In this subsection we summarized allocation approaches widely used in the field of operating systems and cabled networks. We referred to them as channel-unaware schedulers, because their working rationales do not account for channel quality variations, making them unsuitable in cellular networks. Nevertheless, they are fundamental when combined with channel-aware schemes (especially the BET algorithm).

B. Channel-aware/QoS-unaware Strategies

Thanks to CQI feedbacks, which are periodically sent (from UEs to the eNB) using ad hoc control messages, the scheduler can estimate the channel quality perceived by each UE; hence, it can predict the maximum achievable throughput.

Let $d^i(t)$ and $d^i_k(t)$ be the achievable throughput expected for the i-th user at the t-th TTI over all the bandwidth and over the k-th RB, respectively. The mentioned values can be calculated using the AMC module or simply estimated, considering the well-known Shannon expression for the channel capacity, as

$$d^i_k(t) = \log[1 + SINR^i_k(t)]. \tag{9}$$

This definition gives a numerical explanation of the relevance of channel-awareness in wireless contexts. Moreover, we believe that it is of fundamental importance because it synthesizes many concepts expressed in previous sections.

1) Maximum Throughput: The strategy known as Maximum Throughput (MT) aims at maximize the overall throughput by assigning each RB to the user that can achieve the maximum throughput (indeed) in the current TTI. Its metric can be simply expressed as

$$m^{MT}_{i,k} = d^i_k(t). \tag{10}$$

MT is obviously able to maximize cell throughput, but, on the other hand, it performs unfair resource sharing since users with poor channel conditions(e.g., cell-edge users) will only get a low percentage of the available resources (or in extreme case they may suffer of starvation).

A practical scheduler should be intermediate between MT, that maximizes the cell throughput, and BET, that guarantees fair throughput distribution among users, in order to exploit fast variations in channel conditions as much as possible while still satisfying some degrees of fairness.

2) Proportional Fair Scheduler: A typical way to find a trade-off between requirements on fairness and spectral efficiency is the use of Proportional Fair (PF) scheme. Its metric is obtained merging the ones of MT and BET; it can be expressed as

$$m^{PF}_{i,k} = m^{MT}_{i,k} \cdot m^{BET}_{i,k} = d^i_k(t)/\overline{R^i}(t-1) \ . \tag{11}$$

The idea is that the past average throughput can act as a weighting factor of the expected data rate, so that users in bad conditions will be surely served within a certain amount of time. The parameter β in eq. (4) is very important, because it is related to the the time window T_f, over which fairness wants to be imposed, according to the relation [4]

$$T_f = 1/(1-\beta). \tag{12}$$

Intuitively, for $\beta = 0$ the past average throughput results to be equal to the last instantaneous rate and the fairness window T_f would be equal to 1 TTI. On the other hand, for β approaching to 1, the last achieved rate would never be included into the past throughput calculation and the fairness window would theoretically become infinite.

Several algorithms have been proposed in literature to extend PF. In [31], the approach of PF was formulated as an optimization problem, with the objective of maximizing the achieved throughput under the typical constraints of a LTE system. Results showed that performance obtained by using different PF implementations increase with the complexity of the optimization problem. As we recalled in section III-A, in fact, performance improves with the number of considered variables and constraints in the optimization problem, but at the cost of an increased computational complexity, that often makes very hard to realize a given algorithm in real system. We will provide, in the rest of the section, some results on the effectiveness of PF downlink scheduler. Further results on uplink case can be found in [32].

The Generalized Proportional Fair (GPF) approach is developed in [33]. The PF metric is slightly modified by means of two novel parameters, ξ and ψ:

$$m_{i,k}^{GPF} = \frac{\left[d_k^i(t)\right]^{\xi}}{\left[\overline{R^i}(t-1)\right]^{\psi}} \ . \tag{13}$$

The role of ξ and ψ is to modify the impact on the allocation policy of the instantaneous data rate and of the past achieved throughput, respectively. Intuitively, setting $\xi = 0$, the GPF metric would become equal to the BET metric, meaning that fairness can be achieved by the system regardless of the channel conditions. On the other hand, setting $\psi = 0$ would bring to a MT policy with no fairness. Note that the basic PF metric defined in eq. (11) results as a particular case of the GPF with $\xi = \psi = 1$. Similarly to the GPF approach, in [34] and in [35] authors use adaptive schemes capable of tune the achievable fairness level, depending on the system conditions.

3) Throughput to Average: The scheme Throughput To Average (TTA) can be considered as intermediate between MT and PF [28]. Its metric is expressed as:

$$m_{i,k}^{TTA} = \frac{d_k^i(t)}{d^i(t)}. \tag{14}$$

Here, the achievable throughput in the current TTI is used as normalization factor of the achievable throughput on the considered RB. Its meaning is clear, as it quantifies the advantage of allocating a specific RB, guaranteeing that the best RBs are allocated to each user. We note that TTA enables a strong level of fairness on a temporal window of a single TTI. In fact, from its metric it is easy to see that the higher the overall expected throughput of a user is, the lower will be its metric on a single resource block. This means that such a scheduler, as matter of fact, exploits channel awareness for guaranteeing a minimum level of service to every user.

4) Joint Time and Frequency domain schedulers: In [36], a two-step technique for distributing radio resources is presented:

1) at first, a Time Domain Packet Scheduler (TDPS) selects a subset of active users in the current TTI among those connected to the eNB;
2) then, RBs are physically allocated to each user by a FDPS.

The final allocation decision is the outcome of the decisions of two schedulers (one in the time domain and the other one in the frequency domain) that work in series. The main advantage of such a partitioning is that the computational complexity at the FDPS is reduced, due to the number of candidate users for resource allocation decreases. Moreover, authors show that for each phase a different policy can be selected. The latter means that, for instance, RR or PF metrics could be used by the TDPS to obtain fair sharing of time resources among users, and a PF metric could be used by the FDPS to achieve a good trade-off between spectral efficiency and fairness.

In Sec. IV-B7, we will evaluate the performance of a PF scheme applied to both time and frequency domains, which will be referred to as PF-PF, as suggested also in [37].

5) Delay sensitivity: Important works have also been done in order to design delay-sensitive schedulers. The idea is that, even if we neglect the problem of packet deadline expiration (typical of real-time flows), the average data delivering delay can be taken as the overall key performance indicator. In [38], a cross-layer algorithm is presented as an optimization problem to minimize the average delay under several constraints on the assigned MCS, the transmission power, and the BLER requirements. Again, to reduce the complexity of the algorithm, the actual scheduling decisions are taken through multiple stages, each one in charge of optimizing, in an independent way, a given parameter. In their work, authors show the effectiveness of the proposed approach by demonstrating how overall delivering delay can be kept constant despite the increase of cell load.

Delay-sensitive schemes can also be implemented defining metrics as mathematical functions of the experienced delay. As example, in [39], authors associate an utility function to each data packet and use it to weight a MT metric (but it can be applied also considering the PFmetric). The idea is that the utility should be a decreasing function of the experienced delay so that the longer the delay of the packet, the lower the utility, and the higher the probability of the packet to be allocated. Moreover, changing the shape of the utility function, different allocations can be implemented. More details on the utilization of utility functions for dynamic packet scheduling can be found in [40]. Even if in [39] authors do not apply this method directly in a LTE network, this approach can be used such a system for QoS provisioning to real time flows, as we will see in the next sections.

6) Buffer-aware schedulers: Some form of buffer management could be required to avoid packet losses. As example, we can consider the Buffer-Aware Traffic-Dependent (BATD) scheme presented in [41]. Authors deal with the packet dropping probability due to a receiver buffer overflow; they aim at keeping this probability as low as possible while guaranteeing high total system throughput and a certain level of fairness. BATD makes use of buffer status information reported by the user to the eNB and of traffic statistics for setting dynamic priorities associated to each MAC queue. A similar approach for facing the buffer overflow problem is used in [42].

7) Performance evaluation of most relevant channel-aware/QoS-unaware strategies: In order to enrich the analysis of the relevant MT, PF, TTA, and PF-PF schemes, an accurate quantitative performance assessment has been carried out using a system level simulator[2] [43].

The considered scenario is composed by 19 cells and a number of users, chosen in the range [20,100], that move along random paths at 30 km/h within the central cell. A downlink flow, modeled with an infinite buffer source, is assumed to be active for each user. The scenario, moreover, takes into account realistic channel models and several RRM features (the main simulation parameters are summarized in Tab. V).

Figs. 8-10 represent, for each of the considered algorithms, the aggregate cell throughput, the average user throughput, and the well-known Jain fairness index [45], respectively.

As a first consideration, Fig. 8 confirms the effect of multi-user diversity gain, revealing that cell capacity slightly increases with the number of users in the cell. This depends on the multi-user diversity; in fact, growing the number of users in a cell, the probability to find one of such users experiencing good channel conditions at a given time and on a given frequency increases. Obviously, this effect is more evident when the MT metric is used, because PF has to take into account also fairness. Moreover, as expected, TTA does not exploit multi-user diversity gain, since it does not uses channel quality information to increase spectral efficiency, but rather to guarantee some minimum performance to each flow. Therefore, for TTA, an increase of the number of users has the opposite effect, because it will try anyway to allocate, in every TTI, resources to all users.

Fig. 9 shows that the average user throughput for all strategies decreases as the number of users increases. This result is obvious because the same amount of resources has to be share among a higher number of candidates. Presented results demonstrate how, as expected, MT performs always better than the other stragies in terms of the overall achieved throughput, but significantly worse when we consider

[2] Simulation have been carried out with the LTE-Sim simulator [43], available on-line at http://telematics.poliba.it/LTE-Sim.

TABLE V

SIMULATION PARAMETERS

Parameter	Value
Simulation duration	100 s

Physical Details	Carrier Frequency: 2GHz; Bandwidth for the Downlink: 10 MHz;
	# Symbols for TTI: 14; SubFrame length: 1 ms;
	# SubCarries per RB: 12; SubCarrier spacing: 15 kHz;
	Frequency Reuse Scheme: clusters of 4 cells;
	eNB: Power transmission = 43 dBm uniformly distributed among sub-channels;
	Propagation Model: Macro-Cell Urban Model

Link Adaptation	Modulation Schemes: QPSK, 16QAM, and 64QAM with all available coding rate Target BLER: 10^{-1}
Control Overhead	RTP/UDP/IP with ROCH compression: 3 bytes MAC and RLC: 5 bytes; PDCP: 2 bytes; CRC: 3 bytes PHY: 3 symbols
Cell layout	radius: 0.5 km
CQI	Method: Full bandwidth and periodic reporting Measurement period: 1 ms;
UE mobility	Mobility model: Random way-point (see [44]); UE speed: 30 km/h
Traffic Models	real time traffic type: H264, VoIP best effort flows: infinite buffer

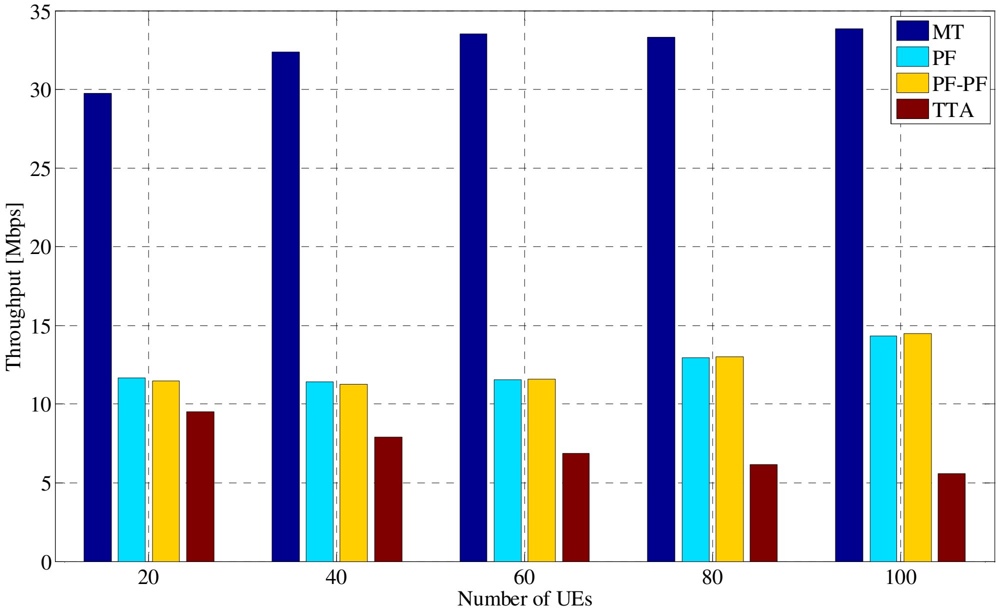


Fig. 8. Aggregate cell throughput for different number of users.

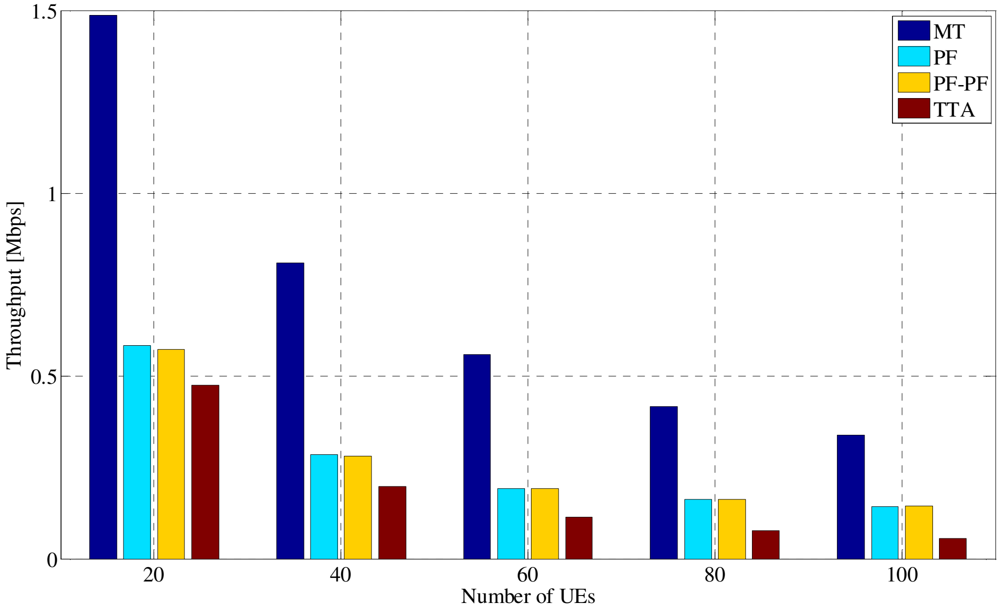


Fig. 9. Average user throughput for different number of users.

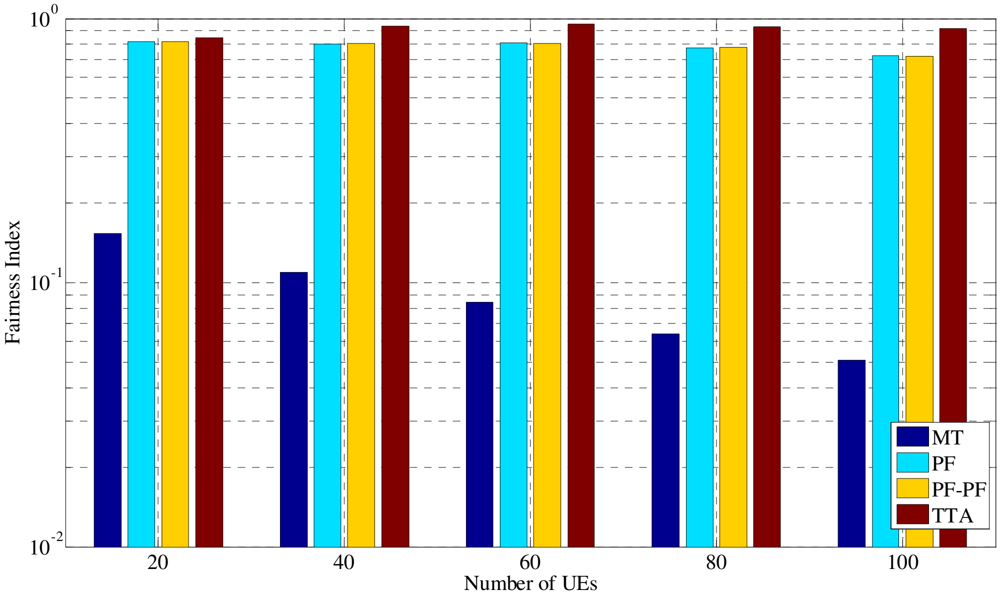
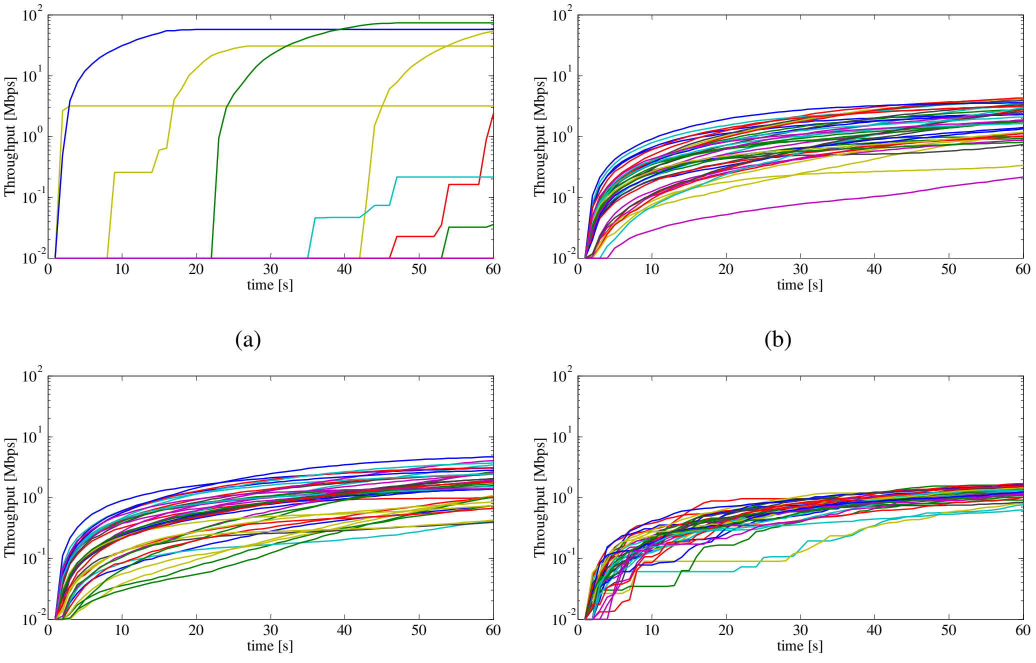


Fig. 10. Fairness index for different number of users.

the achieved fairness level (see Fig. 10). The reason is that MT is able to guarantee a high throughput to a limited number of users, whereas the rest of the users experience very low throughputs. TTA and PF, on the other hand, are able to guarantee high fairness[3] regardless the number of users in the cell.

An overall analysis on the evolution of user throughput in the time domain can provide a further insight on the behavior of different algorithms.



(a)

(b)

(c)

(d)

Fig. 11. Time traces of users' throughput for (a) MT, (b) PF, (c) PF-PF, and (d) TTA schedulers, obtained in a scenario with 40 UEs in the cell.

Fig. 11 shows the time traces of users' throughput in the scenario with 40 users in the cell. It is possible to see the fundamental difference among MT and other schemes. Maximum throughput policy, in fact, tends to allocate resources always to users experiencing best channel conditions. That is, only users very close to the eNB are served while the other ones are starved, as we can see from Fig. 11(a). Other approaches, instead, are more regular in the distribution of resources. As expected, PF and PF-PF perform in a quite similar way, and, after a transient stage, adopting them user throughputs grow with almost the same slope. The high overall fairness achieved by TTA (as shown in Fig. 10) is again observed in Fig. 11(d) thanks to the very compact sheaf of the throughput curves.

[3] Note that the maximum value of the Jain fairness index is 1; that is, the more it is close to 1 the more the allocation is fair.

Finally, it is interesting to compare PF with PF-PF. They perform mostly in the same way. The main difference among this two strategies is in term of complexity. It is easy to demonstrate that the joint time-frequency domain structure allows a significant reduction of the complexity, as it lowers the number of needed metric computations. To this aim, let N and R be the user in the cell and the number of resource blocks, respectively. The conventional frequency domain PF has to calculate $M_{PF} = N \cdot R$ metrics every TTI. On the other hand, considering a TDPS that selects a subset of N_{mux} users to be scheduled (with $N_{mux} < N$) before the FDPS, the joint time-frequency structure should calculate $M_{PF-PF} = N + N_{mux} \cdot R$ metrics every TTI. In our simulations, we have assumed $N_{mux} = 10$ for the PF-PF case (meaning that the TDPS selects only 10 candidate users which resources into the frequency domain are assigned to). Considering, as example, the case with $N = 60$ and $R = 25$, it will result that $M_{PF} = 1500$ and $M_{PF-PF} = 310$, that is, a significant reduction of computational cost, with no performance degradation.

8) General considerations on channel-aware/QoS-unaware strategies: Channel-awareness is a fundamental concept for achieving high performance in a wireless environment, and it can be used by exploiting RRM features such as CQI reporting and link adaptation. If one can estimate the channel quality perceived by a user on a given RB, in fact, it is possible to allocate radio resources obtaining very high data rate. To this aim, the most significant parameter is the expected user throughput, i.e., $d_k^i(t)$, as defined in eq.(9). Nevertheless, spectral efficiency is not the unique objective for a cellular network operator, as it should be able to guarantee a specific level of service also to cell-edge user. In this sense, PF is widely considered the most representative approach for facing the fairness problem in a spectral efficient way.

C. Channel-aware/QoS-aware Strategies

As explained in section II-B, QoS differentiation is handled by associating a set of QoS parameters to each flow. Knowing the values of such parameters, the scheduler can treat data to guarantee some minimum required performances, either in terms of guaranteed data rates or of delivery delays.

In this subsection, we give a comprehensive overview on QoS-aware solutions presented in literature for LTE systems. Once more, it is important to note that QoS-awareness does not necessarily mean QoS provisioning, since it consists in taking allocation decision depending on the requirements of each flow, without necessarily guaranteeing the meeting of such requirements, because it could be unfeasible if procedures for admission control are not implemented.

1) Schedulers for Guaranteed Data-Rate: A generic QoS-aware solution for flows requiring guaranteed data-rate is proposed in [46]. It works in both time and frequency domains. For the time domain, the

Priority Set Scheduler (PSS) has been devised to select users with the highest priority. In particular, users with flows below their target bitrate (i.e., those needing to be urgently allocated to meet QoS requirements) form a high priority set. The rest of the users, instead, forms a lower priority set. Users belonging to first and second sets are managed by using BET and PF algorithms, respectively. Once a number of candidate users has been selected by the TDPS, the FDPS allocates available resources through the *PF scheduled* (PFsch) metric based on the common PF scheme:

$$m_{i,k}^{PF_{sch}} = d_k^i(t)/\overline{R_{sch}^i}(t-1) \tag{15}$$

where $\overline{R_{sch}^i}(t-1)$ is similar to the past average throughput defined in eq. (4), with the difference that it is updated only when the i-th user is actually served.

Intuitively, in this way the value assumed by $\overline{R_{sch}^i}(t-1)$ represents an estimate of the throughput that the i-th user can achieved when he has allocated resources.

A similar approach is followed in [47]. The priority sets are populated depending on the QCI of each data bearer (see section II-B) and classified as GBR and non-GBR set. After this step, the FDPS orderly assigns the best RB to each user in the GBR set, updating the achieved bitrate. When all users in the list have reached their target bitrate, if spare RB are still available, the scheduler assigns them to users in the non-GBR list using PF metric.

Priorities are also used in [48] and calculated, for the i-th user, as:

$$P_i = D_{HOL,i}/\tau_i \ . \tag{16}$$

It is worthwhile to note that raising the value of P_i, the transmission of the head of line packet becomes more urgent. Resource block allocation is then performed, allocating all the resources needed to reach the guaranteed bitrate to the user with the highest priority . If some resources are left free after this allocation, the same operations are repeated considering the user with the second highest priority, and so on. At the end of the process, any free resource is again allocated to the users following the priority order, as far as they have data to transmit.

From an overall point of view, all these approaches use ordered lists to prioritize the most delayed flows and to meet their required bitrate. Nevertheless, the first approach aims at high spectral efficiency, given that the decision process at the FDPS is QoS-unaware. The other two cases appear to be more robust in terms of QoS provisioning, due to the strong prioritization applied also in the assignment of radio resources in the frequency domain.

2) Schedulers for Guaranteed Delay Requirements: Strategies that aim to guarantee bounded delay are the most representative within the category of the QoS-aware schemes since one of the main requirement for a QoS constrained packet is to be delivered within a certain deadline. The same, of course, applies for real-time applications as well as for video streaming and VoIP flows.

Herein, we will first describe algorithms that make use of per-RB metrics, including a performance comparison. Then, for the sake of completeness, more complex procedures present in literature will be surveyed.

The Modified LWDF (M-LWDF) [49] is a channel-aware extension of LWDF and provides bounded packet delivering delay. Even if not originally designed for OFDMA systems, M-LWDF metric can be easily expressed as in Eq. (1). Moreover, non-real and real-time flows are treated differently, handling the first ones with PF and weighting metrics for the second ones as follows [50]:

$$m_{i,k}^{M-LDWF} = \alpha_i D_{HOL,i} \cdot m_{i,k}^{PF} = \alpha_i D_{HOL,i} \cdot \frac{d_k^i(t)}{\overline{R^i}(t-1)} \tag{17}$$

where $D_{HOL,i}$ is the delay of the head of line packet and α_i is calculated as in eq. (8).

As can be easily seen, with respect to its channel unaware version, M-LWDF uses information about the accumulated delay for shaping the behavior of PF, assuring a good balance among spectral efficiency, fairness, and QoS provisioning.

In [51], authors adapt a well-known exponential rule (firstly developed to support multimedia applications in time multiplexed systems [52]) in OFDMA systems, such as LTE, defining the Exponential/PF (EXP/PF). As its name states, EXP/PF takes into account both the characteristics of PF and of an exponential function of the end-to-end delay. Similar to M-LWDF, EXP/PF distinguishes between real-time and best effort flows. For real-time flows the metric is calculated as:

$$m_{i,k}^{EXP/PF} = \exp\left(\frac{\alpha_i D_{HOL,i} - \chi}{1+\sqrt{\chi}}\right) \cdot \frac{d_k^i(t)}{\overline{R^i}(t-1)} \tag{18}$$

where

$$\chi = \frac{1}{N_{rt}} \sum_{i=1}^{N_{rt}} \alpha_i D_{HOL,i} \tag{19}$$

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

Also in this case, proportional fair handles non real-time flows. It is important to point out that both M-LWDF and EXP/PF are based on the assumption that a strictly positive probability of discarding packets is acceptable. Furthermore, taking into account the PF metric of each flow, they both try to guarantee good throughput and an acceptable level of fairness.

Two very promising strategies, i.e., LOG and EXP rules, have been presented in [53]. Sensitivity to channel condition is taken into account with a PF metric that is shaped by means of a function of the head of line packet delay. For the LOG rule such a function is:

$$m_{i,k}^{LOGrule} = b_i \log\left(c + a_i D_{HOL,i}\right) \cdot \Gamma_k^i \ , \tag{20}$$

where b_i, c, and a_i are tunable parameters; Γ_k^i represents the spectral efficiency for the i-th user on the k-th sub-channel.

Authors, moreover, claim that good scheduling performance can be achieved by setting $b_i = 1/E[\Gamma^i]$, $c = 1.1$, and $a_i = 5/(0.99\tau_i)$ [53].

Instead, the EXP rule can be considered as an enhancement of the aforementioned EXP/PF. Its metric is similar to the one of the LOG rule:

$$m_{i,k}^{EXPrule} = b_i \exp\left(\frac{a_i D_{HOL,i}}{c + \sqrt{(1/N_{rt}) \sum_j D_{HOL,j}}}\right) \cdot \Gamma_k^i \ , \tag{21}$$

where, according to [53], the optimal parameter set is:

$$\begin{cases} a_i \in [5/(0.99\tau_i), 10/(0.99\tau_i)] \\ b_i = 1/E[\Gamma^i] \\ c = 1 \end{cases} \tag{22}$$

As we will show next, EXP rule appears to be a more robust solution. Intuitively, LOG rule metric increases with a logarithmic behavior as the head of line delay increases. On the other hand, the exponential function grows much faster with its argument with respect to the logarithmic function. Furthermore, EXP rule also takes into account the overall network status, because that the delay of the considered user is somehow normalized over the sum of the experienced delays of all users.

A different approach is followed in [54], where authors develop a two level framework that guarantees bounded delays to real-time flows. The two levels of the resource allocation procedure work on different time granularity. At the highest level, a discrete time linear control law is applied every LTE frame (i.e., 10 ms) in order to calculate the total amount of data that real-time flows should transmit in the following frame in order to satisfy their delay constraints. Due to the used time granularity, authors refers to it as Frame Level Scheduler (FLS). The lowest layer, instead, works every TTI, and it is in charge of assigning RBs to each flow. In particular, delay constrained flows are allocated following MT policy until the entire amount of data calculated at the highest layer has been transmitted. Then, a PF metric is used to share the spare spectrum among best effort users.

An accurate analysis of the aforementioned scheduling strategies will be presented below. We remark again that these techniques are well-known in literature and often used as a starting point for the development of more complex solutions, such as the ones reported below.

In a very recent work proposed in [55], authors conceived an interesting procedure based on cooperative game-theory that performs resource sharing combining the EXP rule with a virtual token mechanism. It works in two phases: in the first one the game is run to partition available RBs among different groups of flows, populated depending on the type of application they carry. We do not go through the specific game explanation that can be found in [55] and [56]. The second phase uses of EXP rule combined to a virtual token mechanism, in order to meet bounded delay requirements and to guarantee, at the same time, a minimum throughput to all flows [57][58]. In this way, a significant performance gain over the EXP rule is achieved in terms of both packet loss rate and fairness.

In [59], authors describe an opportunistic procedure able to achieve high bandwidth utilization and, at the same time, to guarantee very low Packet Loss Rate (PLR) to delay constrained flows. At first, it evaluates the number of RBs needed by each user as a function of the expected bitrate on the current TTI, of the average past throughput of each user, and of the status of transmission queues. If the overall number of RBs to be allocated is lower than the available bandwidth, spare RBs are distributed to those users having the deadline approaching. Otherwise, users that have less urgency to transmit will be penalized by reducing the number of assigned RBs.

This approach appears heavy from a computational point of view, since after an expensive computation in the first step, it requires a certain number of iterations before reaching the allocation outcome.

A less complex scheme, devised as an evolved version of the previous one, is the Delay-Prioritized Scheduler (DPS) [60]. As a first step, DPS orders candidate users depending on the remaining time before deadline expirations. Once the user with the highest urgency is selected, the frequency allocation step is performed in order to transmit the head of line packet (i.e., the most delayed one). A new iteration is then run on following users in the list, as far as all RBs are assigned.

We note that, with respect to other described routines (e.g., EXP and LOG rules), DPS significantly exploits delay-based priorities, making it similar to EDF.

3) Dynamic Schedulers for VoIP support: According to [61], the maximum acceptable delay for voice is 250 ms. Considering delays introduced by the core network and the delay for RLC and MAC buffering, the tolerable delay at the radio interface should be lower than 100 ms [7] which represents a very strict requirement.

A solution developed for optimizing network performance when there are VoIP and best effort flows is presented in [62]. As in [28], the algorithm is divided in two schemes: TDPS and FDPS. For the time-domain the Required Activity Detection with Delay Sensitivity (RAD-DS) scheme is adopted in order to evaluate the need of a given flow to be scheduled in the current TTI. In particular, the chosen metric for the time domain is obtained as a combination of three different metrics and can be expressed as follows:

$$m_i(t) = m_i^{RR}(t) \cdot RA^i_{traf}(t) \cdot DS^i_{traf}(t) \tag{23}$$

where the RA metric (i.e., the required activity) corresponds to the time share required by i-th user to meet its QoS requirement; for VoIP traffic, for instance, it is calculated as $RA^i_{VoIP}(t) = GBR^i/\overline{R^i_{sch}}(t)$. Furthermore, DS represents a delay sensitive factor that increases with the head of line packet delay (see [62] for details on this shaping).

Finally, RR metric accounts for guaranteeing some fairness in the time resources sharing. In the second scheduling phase (i.e., in the frequency-domain) authors use the PFsch metric expressed by eq. (15).

A scheduling scheme that prioritizes VoIP packets is presented in [63]. It enables periods of VoIP priority mode (VPM) during which only VoIP flows can be served. The length of this priority mode period is kept to a minimum to avoid the degradation of flows other than VoIP, but it can be increased by and adaptive scheme depending on the experienced PLR of VoIP flows. RB allocation is then performed among VoIP flows through a channel sensitive fair queuing as defined in [64].

In conclusion, aforementioned allocation techniques both aim at unbalancing resource allocation in favor of VoIP flows, by increasing the time share for them, rather than modifying the frequency domain allocation. This is more evident in [63], where some TTIs are a-priori entirely dedicated to VoIP flows. In their articles, authors demonstrate how design objectives are achieved by both described algorithms. Unfortunately, results proposed in their papers do not let us to directly compare their performance due to the very different simulation assumptions.

4) Performance evaluation of most relevant channel-aware/QoS-aware strategies: In this section, an accurate analysis of the most promising algorithms based on per-RB metrics (i.e., M-LWDF, EXP/PF, EXP rule, LOG rule, and FLS) will be presented. In addition, the behavior of PF will be also considered as reference.

In the context of QoS provisioning, and especially when dealing with guaranteed delay requirements, a main key performance indicator is the PLR. In general, QoS-aware strategies discard packets that violate the deadline. This is based on the assumption that real-time applications have no advantages in receiving expired packets, and transmitting them after the deadline expiration would represent a waste of resources.

For these reasons, we evaluated the overall PLR of real-time flows by varying the number of users in the cell in the range [10, 40].

Main simulation parameters are summarized in Tab. V, with the difference that, in this case, each user receives three downlink flows (one video, one VoIP, and one best effort flow). In this comparison, we only show performance regarding video flows, given that no significant differences in results have been obtained with VoIP flows. The effectiveness of QoS-aware approaches is demonstrated in Fig. 12, where they outperform PF in terms of PLR for video flows. On the other hand, their ability to limit the PLR decreases as the number of users (i.e., the cell load) increases.

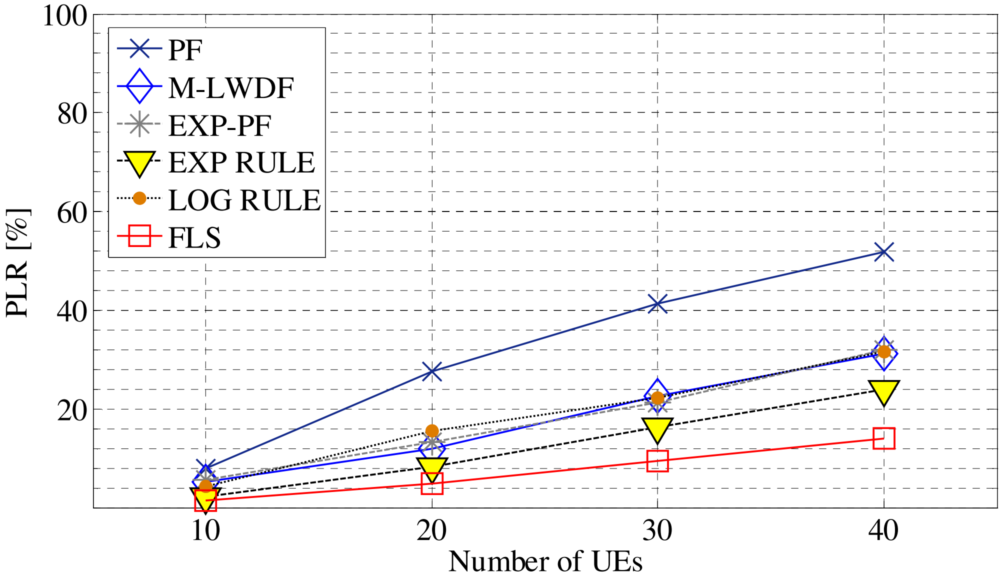
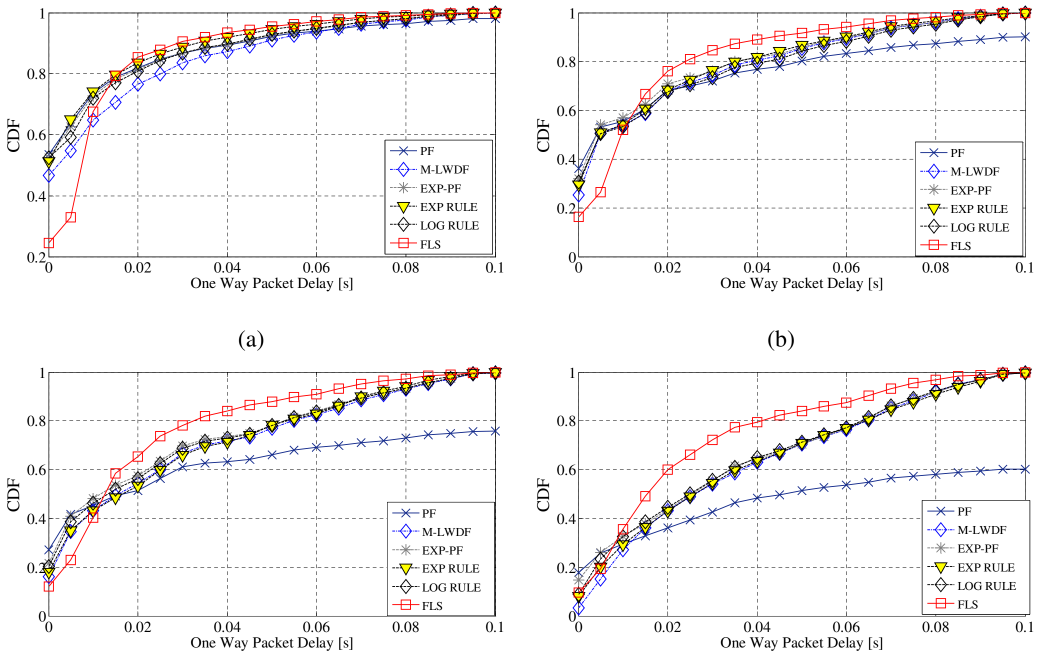


Fig. 12. PLR of video flows with QoS-aware strategies.

Fig. 13 reports the CDF of packet delays for video flows, and shows that all QoS-aware strategies always guarantee packet delivery within the targeted deadline (i.e., 0.1 s). The reason being they drop packets whose delay exceeds the admitted threshold, differently from what PF does.

Our analysis underlines that FLS always reaches the lowest PLR and the lowest delays in all simulated scenarios, but at the cost of reducing resources for best effort flows, as we observe in Fig. 14. Despite FLS appears to be slightly more complex as it introduces additional computational load on top of the structure typically working on a TTI-basis. It is worth to consider that the upper layer only works every 10 TTIs, and, at the same time, the complexity at the bottom layer is quite reduced, since only basic MT and PF are used. In conclusion, these results confirm once more that QoS-unaware solutions, such as



(a)

(b)

(c)

(d)

Fig. 13. CDF of packet delays of video flows with (a) 10, (b) 20, (c) 30, and (d) 40 users with QoS-aware strategies.

PF, are absolutely unsuitable for dealing with delay constrained traffic. Whereas, considered QoS-aware proposals offer tradeoff between QoS requirements and computational load.

5) General considerations on channel-aware/QoS-aware strategies: Delivering packets within a prefixed expiration instant is a fundamental characteristic for a QoS-aware packet scheduler, especially considering that the support of multimedia flows is very important in LTE. Our overview demonstrates the efforts that research community has been doing in this sense. Several interesting works have been presented in literature with the aim of providing to manufacturers better solutions to be implemented in their devices. Nevertheless, despite the huge amount of proposals conceived so far, we believe that some work still need to be done in order to make them ready to be deployed on real devices. In fact, the working rationales of many of them are based on parameter settings (see, for instance, parameters a_i,b_i, and c of LOG and EXP rule) whose optimality might depend on the specific scenario. Further, more evolved approaches risk to be too complex and to waste resources. To complete our discussion and help the reader, all aspects and targets of QoS-aware strategies discussed in this subsection, as well as parameters they use for computing scheduling metrics, have been summarized in Tabs. VI and VII, respectively.

TABLE VI

MAIN ASPECTS AND TARGETS OF QOS-AWARE SCHEDULING STRATEGIES

Name	**Scheduling Target**	**Main Aspects**
PSS/PFsch [46]	Fairness and target bitrate	- Joint TDPS and FDPS structure - PSS at TDPS to populate prioritized subset of users - PFsch at FDPS
[47]	Target bitrate	- Grouping GRB and non-GBR flows - One RB per iteration assigned to GBR users upon meeting GBR requirements - spare resources left to non-GBR
[48]	Target bitrate	- Priority-based ordering of GBR flows - All needed RBs upon meeting GBR requirement are allocated starting from the user with the highest priority
M-LWDF [50]	Bounded Delay	- LWDF scheduler for bounded delay - PF for channel awareness
EXP/PF [51]	Bounded Delay	- Exponential rule for bounded delay - PF for channel awareness
LOG rule [53]	Bounded Delay	- Log rule for bounded delay - PF for channel awareness
EXP rule [53]	Bounded Delay	- Exponential rule for bounded delay - PF for channel awareness
FLS [54]	Bounded Delay	- Double-layer scheduler structure - Control law for resource preemption of real-time flows
[55]	Bounded Delay and minimum guaranteed bitrate	- Cooperative game for distributing resources among different user sets - Resource allocation based on EXP rule with virtual token mechanism
DPS [60]	Bounded Delay	- Prioritization of delay-constrained flows depending on the urgency to be transmitted - All needed RBs upon meeting QoS requirement are allocated starting from the user with the highest priority
RAD-DS/PFsch [62]	VoIP support	- Joint TDPS and FDPS structure - RAD-DS metric at TDPS depends on experienced delay and required QoS - PFsch at FDPS
VPM [63]	VoIP support	- Joint TDPS and FDPS structure - At the TDPS only VoIP flows can be scheduled in *VoIP priority mode* - Channel sensitive WFQ at the FDPS

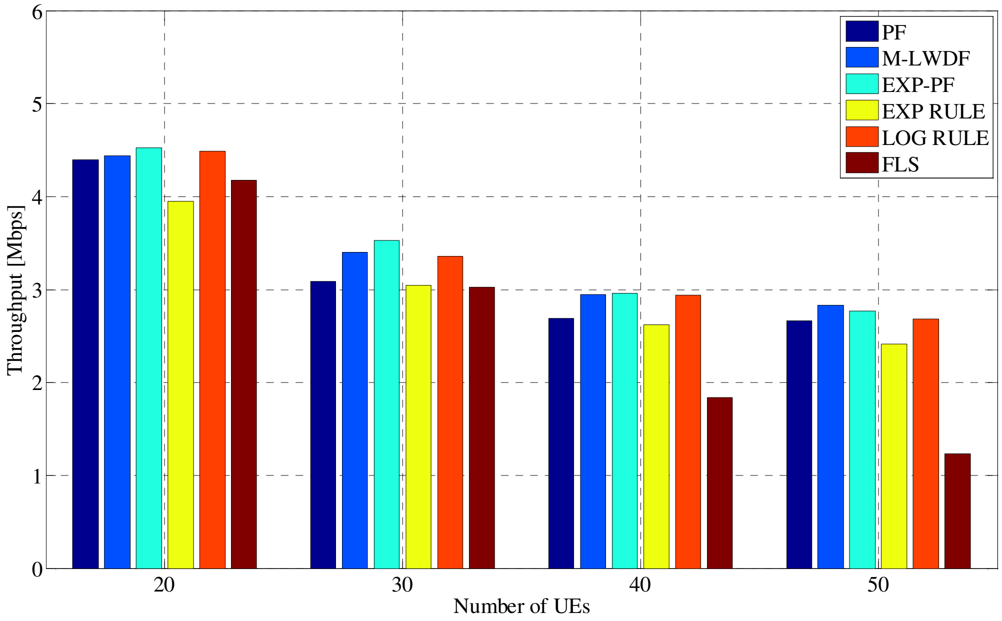


Fig. 14. Aggregate throughput of best effort flows with QoS-aware strategies.

D. Semi-persistent Scheduling for VoIP support

Several solutions have been proposed in literature to supporting high number of VoIP flows, minimizing the impact of signaling overhead. It is important to point out that semi-persistent allocation solutions aim at increasing the VoIP capacity of the network in terms of maximum number of contemporary supported VoIP calls. They are not specifically conceived for improving spectral efficiency or for reducing packet delay and PLR. They can be considered in practice as channel-unaware approaches. Anyway, using them it is possible to indirectly improve performance as the number of scheduled users increases.

In [65], it is shown that the VoIP capacity of the network can be improved with the use of semi-persistent scheme by proposing a slight modification to the generic persistent approach. In particular, the radio resources are divided in several groups of RBs, and each block is pre-configurated and associated only to certain users. In this way a user, even if not actually pre-allocated, will only have to listen to a subset of all the possible RBs. Furthermore, RB groups are associated to each user in contiguous TTIs. Then the resource allocation of each RB group to the associated UEs is performed in a dynamic-like way. This solution appears to be intermediate between the fully dynamic and the persistent scheduling; in fact, the control overhead is reduced with respect to the first one, but it is slightly higher than the one obtained with persistent solutions.

Similarly, in [66] the author proposes to couple VoIP users in pairs, so that they share the same

TABLE VII

Scheduling Input parameters

Name	**Requested bitrate**	**Instantaneous data rate**	**Average data rate**	**Queue size**	**Max delay**	**Head of Line packet delay**	**Max PLR**	**Past PLR**
PSS/PFsch [46]	X	X	X					
[47]	X	X						
[48]	X	X				X		
M-LWDF [50]		X	X		X	X	X	
EXP/PF [51]		X	X		X	X	X	
LOG rule [53]		X	X		X	X		
EXP rule [53]		X	X		X	X		
FLS [54]				X	X			
[55]		X	X		X	X		
DPS [60]		X			X	X		
RAD-DS/PFsch [62]	X	X	X			X		
VPM [63]				X				X

persistently allocated resources. The idea is that pre-allocated resources to each user pair are shared by the same users depending both on the channel conditions and on the experienced PLR. In its simplest version, VoIP users are paired in a random way, without taking into account any user specific information. A slightly more complex method, instead, is designed so that VoIP users experiencing good channel condition are paired with users in bad channel conditions. In this way, users that, due to poor channel quality, require more retransmission opportunities, are favored by pairing users that will tend to scarcely occupy the channel.

A combination of the approaches described in [63] and in [66] is described in [67]. The proposed algorithm, in fact, makes use of a VPM whose duration depends on the channel quality experienced by the users, i.e., if the channel quality improves, the VPM duration reduces. Furthermore, a priority is given to VoIP flow transmission: resources are allocated to VoIP flow pairs in a persistent manner.

1) General considerations on semi-persistent strategies: Even considering the great advantage that persistent solutions give in terms of VoIP capacity and energy savings, they cannot be actually adopted in dynamic contexts such as wireless cellular networks. For this reason, it might be convenient to interleave persistent approaches, specific for VoIP, with dynamic ones that should perform allocation decisions for other flows. The solution in [67], therefore, appears to be the most promising one, as it merges dynamic scheduling for spectral efficiency purposes and VoIP prioritization along with persistent allocation for maximizing performance of VoIP flows.

E. Energy-aware Strategies

Energy saving solutions can be applied to both eNB and UE. For what concern end-user devices, power consumptions can be limited through DRX procedures and (as already discussed in Sec. III) the persistent allocation, which is at the present the only allocation strategy able to meet this goal.

Recent studies reported in [68] and [69] quantify the ecological footprint of future cellular networks. They draw a picture of the cost of the increasing traffic on mobile networks in terms of energy consumed by ICT network infrastructure (e.g., electricity costs are expected to double in the following ten years) and of environmental impact (e.g., total carbon dioxide emissions might triple until 2020). For this reason, green networking related issues are nowadays a hot topic not only for researchers, but also for mobile operators, whose objective is to minimize power consumptions of the network infrastructure to ensure eco-sustainability and to keep low operative costs [70].

A simple way for reaching this target is to maximize the spectral efficiency. In particular, handling a transmission of a given amount of data during a low time interval with high data rates should ensure that an eNB switches more frequently to the sleep mode (i.e., when some parts of the base station, including the radio interface, turn off). This consideration is also confirmed by the study in [71], where it is demonstrated that the MT scheme is more energy efficient than both PF and RR. A different approach is represented by the use of the Bandwidth Expansion Mode algorithm, deployed for achieving energy savings for the eNB in scenarios with low traffic load [72]. It consists of reducing the eNB transmission power by assigning a coding scheme with lower rate to each users, and consequently expanding their spectrum occupation.

Low load conditions can also be exploited, in an energy saving perspective, by compacting the radio resource allocation as much as possible in the time domain, in order to allow the eNB to switch off the transmission equipments. This approach, quite similar to the DRX for the mobile terminal, is referred to as cell discontinuous transmission (DTX). As demonstrated in [73], it can save the 61% of the energy consumed in a realistic scenario. Nevertheless, the main drawback is represented by the need for organizing allocation information inside a very minimum amount of control data.

The main conclusion that we can drawn, after reviewing the state of the art on energy-aware strategies, is that the modification of resource allocation policies on a TTI basis may not have strong impact on energy performance of a cellular network, unless under very low traffic load (i.e., around 50%). In this case, the best choice should be the maximization of the spectral efficiency. In fact, when most of radio resources are free due to lack of data to transmit, fairness and delivering delays do not represent an issue; thus, an allocation based on MT could be used.

V. New Directions and Future Challenges

As already mentioned, the design issues of scheduling techniques strictly follows the evolution of communication technologies. Hence, we conclude our work describing future challenges that could arise with those technologies conceived for the evolution of LTE.

Despite LTE significantly overcomes performances of 3G systems and of other wireless broadband systems currently used, it could be not suitable for future data traffic requirements, which have been recently addressed within IMT-A specifications [74]. Its natural evolution, the Long Term Evolution-Advanced (LTE-A) solution, has been introduced by the 3GPP with the Release 10 for fulfilling and even surpassing IMT-A targets. Innovative technological solutions are standardized or foreseen for LTE-A, such as carrier aggregation, enhanced multi-antenna support, Coordinated Multi-Point (CoMP) transmission techniques, relaying, multi-user Multiple input multiple output (MIMO) communications, and Heterogeneous Networks (HetNets) deployment [75]. In what follows, we provide a brief description on how these aspects can influence the design of the resource allocation schemes, highlighting future related challenges.

1) Carrier Aggregation: A base station can use up to 5 adjacent channels of 20 MHz, thus guaranteeing a notably increase of the network capacity [76]. Considering a broader spectrum utilization, with an allocation policy based on metrics, this means that there are additional computational needs. Moreover, backward compatibility capabilities of LTE-A with LTE [75] imposes to know carrier that LTE-only users should use for receiving/sending packets, because they cannot communicate using more than 20 MHz.

2) Multi-User MIMO: Systems with MIMO features adopt multiple antennas at the transmitter and at the receiver in order to provide simultaneous transmission of several data streams on a single radio link, exploiting the so called spatial diversity gain [77]. In the basic configuration, MIMO transmissions are directed to a single user, thus aiming at increasing its achieved throughput. Nevertheless, spatial diversity can also be exploited by serving different UEs on different spatial streams on the same time/frequency resource, that is, in LTE terms, assigning the same RB to different users. This approach is referred to as Multi-user MIMO (MU-MIMO), and it introduces several issues in the resource allocation policy. In fact, MU-MIMO will allow a base station to pair users for the reception of different downlink flows on a single RB. Hence, a packet scheduler will have to take several new decisions, such as the determination of the best MU-MIMO user pair in terms, for instance, of the overall throughput, and the selection of a single user transmission instead of a MU-MIMO transmission, in case the latter appears to be not convenient [78].

3) Coordinated Multi-Point Transmission: Gain in downlink cell-edge throughput can be achieved in LTE-A with the CoMP architecture. It refers to the possibility to coordinate the downlink transmission towards the same user adopting multiple base stations [79]. The resource allocation process needs, in this case, coordination and synchronization among different eNBs. Let us point out the impact of this requirement. For example, looking at described strategies that use the overall QoS requirements in their metrics for filling up user priority lists, a major challenge is to find the best way to cope with multi-point transmission, identifying whether an eNB should take into account information regarding users of other cells (or a subset of them) or not. We believe that the introduction of CoMP techniques will impose the definition of novel resource allocation techniques, since it will be based on a distributed approach in terms of both information collection and coordinated allocation decision.

4) Scheduling in Heterogeneous Networks: The introduction of new small-range low-power nodes (i.e., micro, pico, femto, and relay nodes) may impact in different ways the design of resource allocation schemes. Micro, pico, and femto nodes can be considered as very small eNBs able to serve a very limited number of users within restricted coverage areas. In particular, micro and pico devices have been conceived for enhancing coverage and capacity in some regions inside a macrocell. Whereas, femto nodes are more suitable for offering broadband services in indoor environments [80]. For this reason, all algorithms presented in previous section could, in general, be directly applied to these technologies, without any modifications. We believe that this apply also to complex strategy, that in principle could not be fitted with low-cost devices such as a femto base station, but whose computational cost would be reduced by the very low number of served users. As a confirmation, recent research works, such as those presented in [81] and [82], do not target issues directly related to packet scheduling, but they focus on other RRM problems, like dynamic frequency allocation and inter-cell interference management by means of dynamic spectrum access. The design of the allocation policies for relay nodes is slightly more complex. Such nodes can be classified as Type I and Type II [83]. The first type works as a range extender through simple packet forwarding, thus not providing any RRM functionality. Relays of type II, instead, implement MAC layer procedures and are in charge of allocating radio resource among registered users. The communication between a relay node and users and the one between relay node and the target eNB, namely the Donor eNB, is organized into a dedicated Multicast Broadcast Single-Frequency Network (MBSFN) frame structure [80]. In details, a relay can exchange packets with the Donor eNB only during proper subframes, leaving the rest of them free for data transmissions with UEs. Therefore, with this kind of relaying, the RRM module has to cope with the presence of both multi-hop connections and MBSFN frame structure (an overview on RRM schemes for OFDMA-based wireless networks with relay nodes can be found in [84]); this has impact on the end-to-end packet delays. Also in this case, all resource sharing approaches, before aforementioned, need to be revised.

VI. Conclusions and Lesson learned

In this paper we provided an extensive survey on downlink packet allocation strategies recently proposed for LTE networks, highlighting at the same time key issues that should be considered when designing a new solution.

LTE is a breakthrough technology with respect to 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, VoIP, online gaming, and everything related to wideband Internet access. Given this ambitious objective, the desired performance can only be achieved by implementing a series of procedures at physical and MAC layers, able to exploit the wireless link capacity up to the Shannon limit.

The most important RRM task is performed by the packet scheduler which is in charge of distributing radio resources among users in an efficient way, taking into account both flow requirements and physical constraints.

From a spectral efficiency point of view, the best solution is to allocate a RB to the user that is expected to exploit it at the best, thus maximizing the cell capacity. However, every other issues, such fairness, computational complexity, cell-edge coverage, QoS provisioning, and energy savings, can be solved always at the cost of reducing the overall cell capacity. In this sense, the design of an allocation strategy often lies in the capacity of finding a good trade-off among the system spectral efficiency and the goals that the network operator wants to reach. A good algorithm should be easily implemented and, above all, should require very low computational cost. For example, if the allocation scheme is based on a complex optimization problem, it would surely guarantee high performance, but it would also lose the fundamental capability of rapidly responding to network changes due to the computational overheads required by each decision.

During our studies, we often found very interesting solutions for theoretical exercises, but they cannot be deployed in real systems due to both the difficulty to be implemented in real devices and the high computational cost required. For these reasons, we often remarked the importance of define a per-RB metric which radio frequency allocation should be based on.

In the first part of the paper, we tried to guide the reader through the understanding of the resource sharing problem in LTE networks, starting from the basics and then adding more and more details in order to explain always complex aspects of the system. The same approach has been followed when surveying the state of the art on allocation policies, classified according to their targets. We showed that, having to deal with wireless environment, we need to to take into account the variable channel conditions. Nevertheless, we showed that metrics already available for operating system and cabled networks, such as those using past throughput and delay sensitiveness, remain useful to shape the behavior of enhanced strategies, according to the desired outcome.

Furthermore, with the introduction of the need for strong QoS support, existing solutions have been demonstrated to be unsuitable for dealing with bounded delays and minimum data-rate requirements. This leads to the introduction of QoS-aware solutions that, from our point of view, are very interesting and promising. They are in general characterized by a strong use of mathematical models such as those belonging to control theory as well as to game theory, or simple mathematical functions on top of the classic metric-based resource allocation. Moreover, they are able to describe flows requirements and to meet the desired performance targets. However, also in this case, it was possible to note a strong presence of channel-unaware and basic channel-aware variables.

Nevertheless, the dependence of the scheduler working rationales on parameter settings is a problem that need to be carefully addressed. We think that a robust strategies should guarantee the ability to work in very different scenarios. Therefore, it should require no strong parameter settings, or it should at least dynamically adapt such parameters to environmental changes.

A main issue we found while studying literature on this topic, is the lack of a common reference scenario that can be used to compare different solutions. We think that it is of fundamental importance to an effective comparison of innovative solutions with existing ones. For this reason, we tried to identify a reference scenario and we used it to compare performance of the most representative solutions. In proposed simulation campaign, we used the PF scheme as a reference strategy. Thanks to this approach, we were able to demonstrate also the real need of sophisticated strategies for facing the problem of QoS provisioning.

REFERENCES

[1] 3GPP, *Tech. Specif. Group Radio Access Network - Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN), 3GPP TS 25.913.*

[2] D. McQueen, “The momentum behind LTE adoption,” *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 44–45, Feb. 2009.

[3] International Telecommunication Union (ITU), *Overall network operation, telephone service, service operation and human factors*, ITU-T Recommendation E.800 Annex B, Aug. 2008.

[4] T. I. Sesia S. and B. M., *LTE, The UMTS Long Term Evolution : From theory to practice*. John Wiley & Sons, 2009.

[5] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution HSPA and LTE for Mobile Broadband*. Academic Press, 2008.

[6] H. Ekstrom, "QoS control in the 3GPP evolved packet system," *IEEE Commun. Mag.*, vol. 47, pp. 76–83, Feb. 2009.

[7] 3GPP, *Tech. Specif. Group Services and System Aspects - Policy and charging control architecture (Release 9), 3GPP TS 23.203*.

[8] ——, *Tech. Specif. Group Radio Access Network - Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Radio Resource Control (RRC); Protocol specification (Release 9), 3GPP TS 36.331*.

[9] ——, *Tech. Specif. Group Radio Access Network - Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Packet Data Convergence Protocol (PDCP) specification (Release 9), 3GPP TS 36.323*.

[10] ——, *Tech. Specif. Group Radio Access Network - Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Radio Link Control (RLC) protocol specification (Release 9), 3GPP TS 36.322*.

[11] ——, *Tech. Specif. Group Radio Access Network - Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Medium Access Control (MAC) protocol specification (Release 9), 3GPP TS 36.321*.

[12] ——, *Tech. Specif. Group Radio Access Network - Physical Channel and Modulation (Release 8), 3GPP TS 36.211*.

[13] N. Kolehmainen, J. Puttonen, P. Kela, T. Ristaniemi, T. Henttonen, and M. Moisio, "Channel Quality Indication Reporting Schemes for UTRAN Long Term Evolution Downlink," in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Marina Bay, Singapore, May 2008, pp. 2522 –2526.

[14] R. Love, R. Kuchibhotla, A. Ghosh, R. Ratasuk, B. Classon, and Y. Blankenship, "Downlink Control Channel Design for 3GPP LTE," in *Proc. of IEEE Wireless Comm. and Net. Conf., WCNC*, Las Vegas, Nevada, USA, Apr. 2008, pp. 813 –818.

[15] F. Capozzi, D. Laselva, F. Frederiksen, J. Wigard, I. Kovacs, and P. Mogensen, "UTRAN LTE Downlink System Performance under Realistic Control Channel Constraints," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Anchorage, Alaska, USA, Sep. 2009.

[16] 3GPP, *Tech. Specif. Group Radio Access Network - Multiplexing and Channel Coding, 3GPP TS 36.212*.

[17] R. Kwan, C. Leung, and J. Zhang, "Multiuser scheduling on the downlink of an LTE cellular system," *Rec. Lett. Commun.*, vol. 2008, pp. 3:1–3:4, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1155/2008/323048

[18] H. Yang, F. Ren, C. Lin, and J. Zhang, "Frequency-Domain Packet Scheduling for 3GPP LTE Uplink," in *Proc. of IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1 –9.

[19] L. Ruiz de Temino, G. Berardinelli, S. Frattasi, and P. Mogensen, "Channel-aware scheduling algorithms for SC-FDMA in LTE uplink," in *Proc. of IEEE Personal, Indoor and Mobile Radio Comm., PIMRC*, Cannes, France, Sep. 2008, pp. 1 –6.

[20] K. Pedersen, G. Monghal, I. Kovacs, T. Kolding, A. Pokhariyal, F. Frederiksen, and P. Mogensen, "Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Baltimore, USA, 2007, pp. 1792 –1796.

[21] 3GPP, *Tech. Specif. Group Radio Access Network - Max #UEs/Subframe for Optimum E-UTRA DL Performance (5-20 MHz), 3GPP TSG-RAN WG1 R1-070792.*

[22] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Commun. Mag.*, vol. 47, no. 6, pp. 48 –55, Jun. 2009.

[23] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, and L. Chen, "Performance Analysis of Power Saving Mechanism with Adjustable DRX Cycles in 3GPP LTE," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Calgary, Alberta, Sep. 2008, pp. 1 –5.

[24] D. Laselva, F. Capozzi, F. Frederiksen, K. Pedersen, J. Wigard, and I. Kovacs, "On the Impact of Realistic Control Channel Constraints on QoS Provisioning in UTRAN LTE," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Anchorage, Alaska, USA, Sep. 2009, pp. 1 –5.

[25] 3GPP, *Tech. Specif. Group Radio Access Network - Persistent Scheduling in E-UTRA, 3GPP TSG-RAN WG1 R1-070098.*

[26] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Proc. of Wireless Commun., Net. and Mobile Comput., WiCom*, Shanghai, China, Sep. 2007.

[27] A. S. Tanenbaum, *Modern Operating Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2007.

[28] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio, "Dynamic packet scheduling performance in UTRA Long Term Evolution downlink," in *Proc. of International Symposium on Wireless Pervasive Comput.,*, Santorini, Greece, May 2008, pp. 308 –313.

[29] D. Liu and Y.-H. Lee, "An efficient scheduling discipline for packet switching networks using Earliest Deadline First Round Robin," in *Proc. of Interntional Conf. on Computer Commun. and Net., ICCCN*, Dallas, USA, Oct. 2003, pp. 5 – 10.

[30] A. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," *Annals of Aplied Probability*, vol. 11, pp. 1–48, 2001.

[31] R. Kwan, C. Leung, and J. Zhang, "Proportional Fair Multiuser Scheduling in LTE," *IEEE Signal Process. Let.*, vol. 16, no. 6, pp. 461 –464, Jun. 2009.

[32] F. Calabrese, C. Rosa, K. Pedersen, and P. Mogensen, "Performance of proportional fair frequency and time domain scheduling in LTE uplink," in *Proc. of IEEE European Wireless Conf., EW*, Lucca, Italy, May 2009, pp. 271 –275.

[33] C. Wengerter, J. Ohlhorst, and A. von Elbwart, "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, vol. 3, Stockholm, Sweden, May 2005, pp. 1903 – 1907.

[34] M. Proebster, C. Mueller, and H. Bakker, "Adaptive fairness control for a proportional fair LTE scheduler," in *Proc. of IEEE Personal Indoor and Mobile Radio Commun., PIMRC*, Istanbul, Turkey, Sep. 2010, pp. 1504 –1509.

[35] X. Li, B. Li, B. Lan, M. Huang, and G. Yu, "Adaptive pf scheduling algorithm in lte cellular system," in *Proc. of International Conf. on Information and Commun. Tech. Convergence, ICTC*, Jeju Island, Korea, Nov. 2010, pp. 501 –504.

[36] A. Pokhariyal, K. Pedersen, G. Monghal, I. Kovacs, C. Rosa, T. Kolding, and P. Mogensen, "HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution," in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Dublin, Ireland, Apr. 2007, pp. 2761 –2765.

[37] K. C. Beh, S. Armour, and A. Doufexi, "Joint Time-Frequency Domain Proportional Fair Scheduler with HARQ for 3GPP LTE Systems," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Calgary, Alberta, Sep. 2008, pp. 1 –5.

[38] H. Fattah and H. Alnuweiri, "A cross-layer design for dynamic resource block allocation in 3G Long Term Evolution System," in *Proc. of IEEE Mobile Adhoc and Sensor Systems, MASS*, Macau, China, 2009, pp. 929 –934.

[39] P. Liu, R. Berry, and M. Honig, “Delay-sensitive packet scheduling in wireless networks,” in *Proc. of IEEE Wireless Commun. and Net. Conf., WCNC*, vol. 3, Atlanta, Geogia, USA, Mar. 2003, pp. 1627 –1632 vol.3.

[40] G. Song and Y. Li, “Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks,” *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127 – 134, Dec. 2005.

[41] J. Huang and Z. Niu, “Buffer-Aware and Traffic-Dependent Packet Scheduling in Wireless OFDM Networks,” in *Proc. of IEEE Wireless Commun. and Net. Conf., WCNC*, Hong Cong, China, Mar. 2007, pp. 1554 –1558.

[42] Y. Lin and G. Yue, “Channel-Adapted and Buffer-Aware Packet Scheduling in LTE Wireless Communication System,” in *Proc. of Wireless Commun., Net. and Mobile Comput., WiCOM*, Dalian, China, Oct. 2008, pp. 1 –4.

[43] G. Piro, L. Grieco, G. Boggia, F. Capozzi, and P. Camarda, “Simulating LTE Cellular Systems: An Open-Source Framework,” *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498 –513, Feb. 2011.

[44] T. Camp, J. Boleng, and V. Davies, “A survey of mobility models for ad hoc network research,” *Wireless Commun. and Mobile Comput.*, vol. 2, pp. 483–502, 2002.

[45] R. Jain, *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, 1991.

[46] G. Monghal, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, “QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution,” in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Marina Bay, Singapore, May 2008.

[47] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, “Multi-qos-aware fair scheduling for lte,” in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, May 2011, pp. 1 –5.

[48] D. Skoutas and A. Rouskas, “Scheduling with qos provisioning in mobile broadband wireless systems,” in *Wireless Conference (EW), 2010 European*, April 2010, pp. 422 –428.

[49] 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.

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

[51] R. Basukala, H. Mohd Ramli, and K. Sandrasegaran, “Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system,” in *Proc. of First Asian Himalayas International Conf. on Internet, AH-ICI*, Kathmundu, Nepal, Nov. 2009, pp. 1 –5.

[52] J.-H. Rhee, J. Holtzman, and D.-K. Kim, “Scheduling of real/non-real time services: adaptive EXP/PF algorithm,” in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, vol. 1, Jeju, Korea, Apr. 2003, pp. 462 – 466.

[53] B. Sadiq, R. Madan, and A. Sampath, “Downlink scheduling for multiclass traffic in lte,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2009, pp. 9–9, 2009.

[54] G. Piro, L. Grieco, G. Boggia, R. Fortuna, and P. Camarda, “Two-level Downlink Scheduling for Real-Time Multimedia Services in LTE Networks,” in *IEEE Trans. Multimedia, to be published*, vol. 13, no. 5, Oct. 2011, pp. 1052 –1065.

[55] M. Iturralde, A. Wei, and A. Beylot, “Resource allocation for real time services using cooperative game theory and a virtual token mechanism in lte networks,” in *IEEE Personal Indoor Mobile Radio Communications, PIMRC*, Jan. 2012.

[56] M. Iturralde, T. Yahiya, A. Wei, and A. Beylot, “Resource allocation using shapley value in lte,” in *IEEE Personal Indoor Mobile Radio Communications, PIMRC*, Sept. 2011.

[57] ——, “Performance study of multimedia services using virtual token mechanism for resource allocation in lte networks,” in *Vehicular Technology Conference (VTC Fall), 2011 IEEE 74th*, Sept. 2011.

[58] P. L. L. and D. B. S., *Computer Networks: A Systems Approach, 3rd Edition*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[59] H. Adibah Mohd Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, M. Xue, and C.-C. Lin, "Resource allocation technique for video streaming applications in the lte system," in *Wireless and Optical Communications Conference (WOCC), 2010 19th Annual*, May 2010, pp. 1 –5.

[60] K. Sandrasegaran, H. A. Mohd Ramli, and R. Basukala, "Delay-prioritized scheduling DPS for real time traffic in 3gpp lte system," in *Wireless Communications and Networking Conference (WCNC), 2010 IEEE*, Apr. 2010, pp. 1 –6.

[61] J. Puttonen, T. Henttonen, N. Kolehmainen, K. Aschan, M. Moisio, and P. Kela, "Voice-over-IP performance in UTRA long term evolution downlink," in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Marina Bay, Singapore, May 2008.

[62] G. Monghal, D. Laselva, P. Michaelsen, and J. Wigard, "Dynamic Packet Scheduling for Traffic Mixes of Best Effort and VoIP Users in E-UTRAN Downlink," in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Marina Bay, Singapore, May 2010, pp. 1 –5.

[63] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "MAC Scheduling Scheme for VoIP Traffic Service in 3G LTE," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Baltimore, MD, USA, Oct. 2007.

[64] L. Wang, Y.-K. Kwok, W.-C. Lau, and V. Lau, "Channel adaptive fair queueing for scheduling integrated voice and data services in multicode cdma systems," in *Proc. of IEEE Wireless Commun. and Net., WCNC*, vol. 3, New Orleans, Louisiana, USA, Mar. 2003, pp. 1651 –1656.

[65] Y. Fan, P. Lunden, M. Kuusela, and M. Valkama, "Efficient Semi-Persistent Scheduling for VoIP on EUTRA Downlink," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Calgary, Alberta, Sep. 2008, pp. 1 –5.

[66] Y.-S. Kim, "An Efficient Scheduling Scheme to Enhance the Capacity of VoIP Services in Evolved UTRA Uplink," *EURASIP J. Wirel. Commun. Netw.*, vol. 2008, 2008.

[67] S. Saha and R. Quazi, "Priority-coupling-a semi-persistent MAC scheduling scheme for VoIP traffic on 3G LTE," in *Proc. of Conf. on Telecommunications, ConTEL*, Zagreb, Croatia, 2009, pp. 325 –329.

[68] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55 –62, Aug. 2011.

[69] V. Mancuso and S. Alouf, "Reducing costs and pollution in cellular networks," *Communications Magazine, IEEE*, vol. 49, no. 8, pp. 63 –71, August 2011.

[70] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524 –540, Apr. 2011.

[71] D. Sabella, M. Caretti, and R. Fantini, "Energy efficiency evaluation of state of the art packet scheduling algorithms for lte," *in Proc. of IEEE European Wireless Conf., EW*, pp. 1–4, Apr. 2011.

[72] S. Videv and H. Haas, "Energy-efficient scheduling and bandwidth-energy efficiency trade-off with low load," in *Proc. IEEE International Conf. on Comm.*, Jun. 2011, pp. 1 –5.

[73] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing Energy Consumption in LTE with Cell DTX," in *Proc. of IEEE Vehic. Tec. Conf., VTC-Spring*, May 2011, pp. 1 –5.

[74] ITU-R , *Requirements Related to Technical Performance for IMT-Advanced Radio Interfaces, Rep. M.2134, 2008.*

[75] S. Parkvall, A. Furuskar, and E. Dahlman, "Evolution of LTE toward IMT-advanced," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 84 –91, Feb. 2011.

[76] K. I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. Garcia, and Y. Wang, "Carrier aggregation for LTE-advanced: functionality and performance aspects," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 89 –95, Jun. 2011.

[77] K. C. Beh, A. Doufexi, and S. Armour, “On the performance of SU-MIMO and MU-MIMO in 3GPP LTE downlink,” in *In Proc. Of IEEE Personal, Indoor and Mobile Radio Communications, PIMRC*, Sep. 2009, pp. 1482 –1486.

[78] B. Mondal, T. Thomas, and A. Ghosh, “Mu-mimo system performance analysis in lte evolution,” in *In Proc. Of IEEE Personal, Indoor and Mobile Radio Communications, PIMRC*, sept. 2010, pp. 1510 –1515.

[79] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel, “Coordinated multipoint: Concepts, performance, and field trial results,” *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102 –111, Feb. 2011.

[80] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, “A survey on 3GPP heterogeneous networks,” *IEEE Wireless Com.*, vol. 18, no. 3, pp. 10 –21, Jun. 2011.

[81] V. Capdevielle, A. Feki, and E. Temer, “ Enhanced Resource Sharing Strategies for LTE Picocells with Heterogeneous Traffic Loads ,” in *In Proc. Of IEEE Vehic. Tech. Conf., VTC Spring*, May 2011, pp. 1–5.

[82] T. Lan, k. Sinkar, L. Kant, and K. Kerpez, “Resource Allocation and Performance Study for LTE Networks Integrated with Femtocells,” in *In Proc. Of IEEE Global Telecommunications Conference, GLOBECOM*, Dec. 2010, pp. 1 –6.

[83] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, “LTE-advanced: next-generation wireless broadband technology,” *IEEE Wireless Com.*, vol. 17, no. 3, pp. 10 –22, Jun. 2010.

[84] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y.-D. Kim, E. Kim, and Y.-C. Cheong, “An Overview of Radio Resource Management in Relay-Enhanced OFDMA-Based Networks,” *IEEE Commun. Surveys Tuts.*, vol. 12, no. 3, pp. 422 –438, Apr. 2010.