NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES
Communication Systems & Networks

MASTER THESIS

A Scheduling and Resource Allocation Algorithm for LTE Networks Using Tree Structures

Ioannis A. Lionas

Ledion Z. Sotiri

Supervisors: **Lazaros Merakos,** Professor

ATHENS

JUNE 2014

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
Συστήματα Επικοινωνιών & Δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμος Χρονοπρογραμματισμού και Διαχείρισης Πόρων σε Δίκτυα LTE με Χρήση Δενδρικών Δομών

Ιωάννης Α. Λιόνας

Λεντιόν Ζ. Σωτήρη

Επιβλέποντες: **Λάζαρος Μεράκος,** Καθηγητής ΕΚΠΑ

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2014

MASTER THESIS

A Scheduling and Resource Allocation Algorithm for LTE Networks Using Tree Structures

Ioannis A. Lionas

ID: M1235

Ledion Z. Sotiri

ID: M1197

SUPERVISORS: **Lazaros Merakos,** Professor

ADVISORY COMMITEE: **Lazaros Merakos,** Professor
Stathes Xadjiefthymiades, Associate Professor

June 2014

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αλγόριθμος Χρονοπρογραμματισμού και Διαχείρισης Πόρων σε Δίκτυα LTE με Χρήση Δενδρικών Δομών

Ιωάννης Α. Λιόνας

Α.Μ.: Μ1235

Λεντιόν Ζ. Σωτήρη

Α.Μ.: Μ1197

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Λάζαρος Μεράκος,** Καθηγητής ΕΚΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Λάζαρος Μεράκος,** Καθηγητής ΕΚΠΑ
Ευστάθιος Χατζηευθυμιάδης, Αναπληρωτής Καθηγητής

Ιούνιος 2014

ABSTRACT

Long Term Evolution (LTE) has been designed by 3GPP with the target to meet the ever increasing demands in broadband wireless access. Techniques such as OFDMA multiple access scheme, MIMO and Adaptive Modulation and Coding (AMC) have been adopted in order to boost the achieved data rates and improve spectral efficiency. However, the development of sophisticated scheduling algorithms is required so that the full potential of those techniques is exploited. Even though 3GPP has fully standardized the control signaling required to perform scheduling, the algorithms that need to be executed to make efficient decisions are left to vendor implementation.

Therefore significant research effort has been dedicated to this direction and several scheduling algorithms have been proposed. The main conclusion drawn from the study of the literature is that scheduling in a multicarrier system with the restrictions of LTE constitutes a multidimensional problem. Among the multiple dimensions, those that are mostly considered in the proposed algorithms are throughput, fairness and QoS guarantee.

The main contribution of this thesis is to propose a new scheduling algorithm that addresses most of the issues that define the overall performance of an LTE scheduler. The proposal focuses mainly on the complexity involved in making a scheduling decision. It attempts to resolve this issue by the introduction of a sophisticated tree structure that enables the efficient storage of all the parameters that are considered essential in the scheduling decision process. Thus the scheduler can have immediate access to this information. A full time domain scheduling algorithm that utilizes this tree structure is described and two new resource allocation algorithms are proposed. These algorithms also utilize some additional tree structures derived from appropriate preprocessing actions that take place before the actual scheduling decision. The first algorithm has low complexity and satisfactory performance while the second has improved performance with the cost of some additional complexity. Extensive simulation results confirm the capability of the proposed scheme in satisfying the strict QoS requirements of LTE, while the performance of the newly proposed algorithms is compared with well-known scheduling techniques. The proposed solutions are applicable to the downlink case, while the same concept may be adapted properly to provide an efficient solution for the uplink case.

SUBJECT AREA: Wireless Communication Networks, LTE, Scheduling, Resource Allocation

KEYWORDS: scheduling, resource allocation, resource block, algorithm, throughput, multiuser diversity, adaptive modulation, coding, channel state, tree structure

ΠΕΡΙΛΗΨΗ

Το σύστημα LTE σχεδιάστηκε από τη 3GPP με στόχο την ικανοποίηση των ολοένα αυξανόμενων αναγκών για ασύρματη ευρυζωνική πρόσβαση. Τεχνικές όπως το σχήμα πολλαπλής πρόσβασης OFDMA, το MIMO και η Προσαρμοστική Διαμόρφωση και Κωδικοποίηση υιοθετήθηκαν προκειμένου να αυξήσουν τους επιτεύξιμους ρυθμούς μετάδοσης και να βελτιώσουν τη φασματική απόδοση. Ωστόσο, απαιτείται η ανάπτυξη εξελιγμένων αλγορίθμων χρονοπρογραμματισμού προκειμένου να αξιοποιηθεί η πλήρης δυναμική αυτών των τεχνικών. Παρά το γεγονός ότι η 3GPP έχει προτυποποιήσει πλήρως τη σηματοδοσία ελέγχου που απαιτείται για την εκτέλεση του χρονοπρογραμματισμού, οι αλγόριθμοι που χρειάζεται να εκτελεστούν προκειμένου να ληφθούν αποδοτικές αποφάσεις έχουν αφεθεί στους κατασκευαστές για υλοποίηση.

Ως εκ τούτου, σημαντική ερευνητική προσπάθεια έχει καταβληθεί προς αυτή την κατεύθυνση και έχουν προταθεί αρκετοί αλγόριθμοι χρονοπρογραμματισμού. Το κύριο συμπέρασμα που εξάγεται από τη μελέτη της βιβλιογραφίας είναι ότι ο χρονοπρογραμματισμός σε ένα σύστημα πολλών φερουσών με τους περιορισμούς του LTE αποτελεί ένα πολυδιάστατο πρόβλημα. Ανάμεσα στις πολλές διαστάσεις του, αυτές που κυρίως λαμβάνονται υπόψη στους προτεινόμενους αλγορίθμους είναι η ρυθμαπόδοση, η δικαιοσύνη και η εξασφάλιση εγγυημένης ποιότητας υπηρεσίας.

Η κύρια συμβολή της παρούσας διατριβής είναι η πρόταση ενός νέου αλγόριθμου χρονοπρογραμματισμού και διαχείρισης πόρων ο οποίος αντιμετωπίζει τα περισσότερα από τα θέματα που καθορίζουν τη συνολική απόδοση μίας οντότητας χρονοπρογραμματισμού του LTE. Η πρόταση εστιάζει κυρίως στην πολυπλοκότητα που εισάγεται κατά τη λήψη μίας απόφασης χρονοπρογραμματισμού. Επιχειρεί δε να επιλύσει αυτό το πρόβλημα με την εισαγωγή μίας εξελιγμένης δενδρικής δομής η οποία επιτρέπει την αποδοτική αποθήκευση όλων των παραμέτρων που θεωρούνται ουσιώδεις στη διαδικασία λήψης μίας απόφασης χρονοπρογραμματισμού. Με αυτό τον τρόπο η οντότητα χρονοπρογραμματισμού έχει άμεση πρόσβαση σε αυτές τις πληροφορίες. Στην εργασία περιγράφεται ένας πλήρης αλγόριθμος προγραμματισμού στο πεδίο του χρόνου που αξιοποιεί αυτή τη δενδρική δομή και επιπλέον προτείνονται δύο νέοι αλγόριθμοι κατανομής πόρων. Οι αλγόριθμοι αυτοί επίσης αξιοποιούν κάποιες επιπρόσθετες δενδρικές δομές οι οποίες παράγονται ύστερα από κατάλληλη προ-επεξεργασία που λαμβάνει χώρα πριν τη λήψη της απόφασης χρονοπρογραμματισμού. Ο πρώτος αλγόριθμος παρουσιάζει χαμηλή πολυπλοκότητα και ικανοποιητική απόδοση ενώ ο δεύτερος βελτιωμένη απόδοση με το κόστος κάποιας επιπρόσθετης πολυπλοκότητας. Αποτελέσματα εκτεταμένων προσομοιώσεων επιβεβαιώνουν την ικανότητα του προτεινόμενου σχήματος στην ικανοποίηση των αυστηρών απαιτήσεων του LTE σε ότι αφορά την ποιότητα υπηρεσίας, ενώ ταυτόχρονα η απόδοση των προτεινόμενων αλγορίθμων συγκρίνεται με γνωστές τεχνικές χρονοπρογραμματισμού. Οι προτεινόμενες λύσεις είναι εφαρμόσιμες μόνο στην περίπτωση της κατωφερούς ζεύξης, ωστόσο η ίδια ιδέα μπορεί να προσαρμοστεί κατάλληλα για να παρέχει μία αποδοτική λύση και στην περίπτωση της ανωφερούς ζεύξης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Δίκτυα Ασύρματων Επικοινωνιών, Χρονοπρογραμματισμός, Διαχείριση Πόρων, LTE.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: χρονοπρογραμματισμός, διαχείριση πόρων, αλγόριθμος, ρυθμαπόδοση, προσαρμοστική διαμόρφωση, κωδικοποίηση, κανάλι, δενδρική δομή

Στο γιο μου Αντώνη

Ιωάννης Λιόνας

Στους γονείς μου

Λεντιόν Σωτήρη

ACKNOWLEDGMENTS

The completion of a task not only triggers to humans a sense of joy and satisfaction but also serves so they acquire further knowledge and experiences, which are not limited only to the main subject but collateral knowledge is also acquired. But any personal achievement in any field in fact indicates that others have also contributed, consequently we feel obliged to express our gratitude to them.

We would like to thank Prof Lazaros Merakos for giving us the opportunity to participate in a research project as well as Dr. Nikos Passas who was always willing to guide us and resolve any situation that occurred. We are grateful to Dr. Spyros Xergias whose work was a stepping stone for our work.

Ledion also would like to thank his colleague Ioannis with whom he had an excellent collaboration during those few months which resulted in the completion of this joint dissertation.

Finally Ledion would like to give grateful thank to his lovely parents for their unlimited support and encouragement, not only during this dissertation but, from the beginning of his academic education.

Ioannis would like to thank his colleague Ledion for the impeccable cooperation and the fruitful knowledge interchange.

Finally Ioannis would like to thank his family for their support and patience that helped him work undistracted on this thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This thesis was conducted from December 2013 until June 2014 under the supervision of Dr Spyridon Xergias and Dr Nikos Passas. Attending the lectures of the master course “Advanced Topics in Wireless and Mobile Networks” was our initial motivation to conduct this thesis. Our research effort would be incomplete without the simulation results produced using the WiMAX simulator that Dr Spyridon Xergias kindly provided us. The simulator was modified properly according to the LTE standard, producing a simulation environment where we could program and evaluate the proposed algorithms.

1. INTRODUCTION

1.1 Mobile Communications

Mobile communications as we know them today constitute a relatively new branch of communications, considering its origins at the beginning of mobile telephony its life is less than half a century [1]. Despite its youngness, its development has been so rapid in many areas that an attempt to describe it requires a multidimensional approach: over the offered services the supported technologies the standardizing agencies etc. For the time being, mobile communications possess a great portion in the market of telecommunications worldwide and its penetration in the developed world is remarkably high [2]. But what really does the term mobile communication includes? A general definition may be given as: with the term mobile communications we refer to the ability an entity has to communicate with another one any time and from any place mainly through wireless cellular networks.

1.2 Challenges and Demands for Mobile Communication

If one attempted to note down the social, individual, industrial, business demands and applications of mobile communications would realize that these demands and applications were not always planned or predefined but usually arise as a result of the deployment of mobile communications systems. In this context it is difficult to determine and predict which will be the applications and the provided services in the next generation networks. However we can note down some applications, provided by existing mobile and non mobile systems, which require more advanced features than those supported by existing mobile systems such as LTE, therefore we may predict at least some demands that next generation mobile systems should support.

Applications such as mobile gaming, mobile video, mobile video conferencing, high quality services and other relevant activities require high bit rates and very low latency, features that are not met or at least are not fully met by current 3G systems. Besides the challenges that LTE – Advanced must overcome it has to confront with limited resources like bandwidth or to deal with the rapidly growing number of users which tends to become exponential.

Some of the aforementioned demands can be fulfilled only by 4G systems. The most prominent and realistically affordable 4G system for mobile communication that not only meets the required features but in many cases exceeds those is LTE-Advanced.

1.3 Services Aimed to be Provided in LTE – Advanced Systems

LTE-Advanced is qualified as 4G technology which aims to further expand and enrich existing features supported by LTE systems providing thus a wide variety of new services and solving still-remaining issues from 3G systems. In this topic we will make a concise description of the services provided by LTE-Advanced and the corresponding applications.

As we referred earlier 4G does not provide only new services but it refines existing ones. In our description we do not distinguish between existing and new services.

Services to be provided by LTE-Advanced:

- Office: Agenda, contacts, email, information browser, fax.
- Media: MP3 Player, TV receiver, camera, video, game console.

- Gateway: internet gateway, voice gateway, messaging gateway.
- Navigation: Maps, Location Information, in car navigation, pedestrian navigation.
- Wallet: Payments, keys, access, passes.
- Remote: Entertainment systems, garage, information appliance.

The range of applications for the provided services can be categorized according to the addressing sector:

- Applications for Individuals: This includes all the social or professional activities an individual can involve, such as: e-learning, social networking, mobile blogging, mobile chatting and gaming. As professionals: medical personnel, service technician, delivery people etc.
- Applications in Business Environment: In business, mobile communications have various applications such as: information services, transactions and banking services, online commerce, remote costumer care, call management, online address book and directory services.
- Industrial applications: In industry among many applications mobile communications are used in trading, payment, mobile airline ticket, city or country navigating, mobile high definition TV, yellow pages etc.

1.4 System Requirements to Meet the Target Services

LTE-Advanced systems were designed to complement and eventually replace the LTE systems. The requirements for 4G systems include accessing services from anywhere, anytime, reliably and with high speed [3]. But to meet all the designed targets LTE-Advanced must support special specifications. In this paragraph we provide a detailed description of all the features that LTE-Advanced systems must provide in order to support the services referred in the previous paragraph.

Pertaining its compatibility to other systems LTE-Advanced should meet:

- A high degree of functionality worldwide while retaining the flexibility to support a wide range of local services and applications in a cost efficient manner.
- Compatibility of services with other systems and capability of interworking.
- High quality mobile services, user friendly applications, services and equipment.
- User equipment suitable for worldwide use, and worldwide roaming capability.

The features that LTE- Advanced supports are [4]:

- Peak spectral efficiency in b/s/Hz 15 using up to 4x4 MIMO and 30 using up to 8x8 MIMO.
- Downlink cell spectral efficiency in b/s/Hz (at 3 km/h with 500 m inter-site distance) 2.4, 2.6, 3.7 for 2x2, 4x2, 4x4 MIMO correspondingly.
- New bands in addition to those used in LTE (Release 8): 450 – 470 MHz, 698 – 862 MHz, 790 – 862 MHz, 2.3 – 2.4 GHz, 3.4 – 4.2 GHz, 4.4 – 4.99 GHz some of those have formally been included in Releases 9 and 10.
- New User Equipment – UE categories were defined in [5] in order to accommodate LTE – Advanced capabilities.

- Peak data rates 3 Gbps in downlink and 1.5 Gbps in uplink.
- Increased Number of simultaneously active subscribers.
- Improved performance related to LTE at cell edges.
- Support of asymmetrical and larger bandwidths with maximum at 100 MHz.
- Enhanced multi-antenna transmission techniques DL 8x8 MIMO, UL 4x4 MIMO.

1.5 Thesis Outline

This Thesis objective is to propose, analyze and deploy a tree-based QoS aware scheduling and resource allocation algorithm for mixed type traffic in LTE-Advanced wireless networks. We concentrate mainly on downlink leaving uplink for future work, although its implementation is a straightforward application with minor changes of the proposed schemes. The main feature in our proposal is the usage of a structure called COFRTS – Colored OFDMA Frame Registry Tree which stores information (metadata from received packets in the enhanced Node Bases – eNB station queues) in an efficient way that when retrieved, the complexity of the calculations made in order to extract useful information is reduced significantly.

The work represented in this thesis is divided into 7 chapters:

In chapter 1 we make a concise description of mobile communications, the necessities that imposed the introduction of LTE-Advanced and the services and features that 4G systems support. In paragraph 1.5 we describe the thesis motivation.

In Chapter 2 the overall LTE system architecture is covered along with the stack of protocols implemented to support the fully packet switched functionality of the network. Additionally, LTE standardization regarding the QoS management is presented and analyzed.

Chapter 3 contains an overview of the LTE physical layer parameters focusing mostly on those that affect the scheduling and resource allocation decisions. The chapter also includes an introduction to the theory of the two LTE features that contribute the most in its enhanced performance: the OFDMA multiple access scheme and the multiple antennas techniques, also known as MIMO.

A concise summary of the most known scheduling and resource allocation techniques proposed for LTE in the literature is contained in Chapter 4. Lessons learnt from the adoption of each one of the proposed solutions are identified and analyzed.

In Chapter 5 a description and in depth analysis of two newly proposed solutions for the LTE scheduling and resource allocation problem is given. The ideas behind the design of the algorithms are highlighted and a full evaluation of them in terms of complexity is carried out.

Chapter 6 contains the implementation part of this thesis. An overview of the simulation environment is presented along with all the specific simulation parameters and assumptions and the final results from the comparison of the proposed algorithms with some widely used techniques.

Finally, generic conclusions drawn from the simulation results are presented in Chapter 7.

2. LTE ARCHITECTURE

2.1 LTE Network Architecture

The increased requirements for higher peak data rates, lower delays and better quality of service led 3GPP to design and adopt a new more flexible architecture in the network deployment compared to its predecessor systems. The main characteristic of this new approach is the simplification the mobile terminal's access to the network by decreasing the number of network components required. Additionally the network was designed so that it can support a fully packet switched functionality using the IP protocol. No circuit switched services are offered by the LTE and the network architecture provides easy access to any external packet switched network such as the Internet, IP Multimedia Subsystem (IMS) e.t.c.

LTE Network architecture has two distinct components: the access network and the core network ([6]). The access network, called Evolved UMTS Terrestrial Radio Access Network (E-UTRAN), handles the radio communication between the mobile station, called User Equipment (UE), and the base station, called e-NodeB (eNB). The main functions of eNB involve transmitting and receiving data to and from the UE along with providing the UE with all the appropriate control information. This information includes control signaling to indicate a handover to a neighboring cell, power control settings, scheduling and resource allocation commands. Therefore eNB combines the earlier functionalities of nodeB and Radio Network Controller (RNC), thus reducing the latency that arises when UE is communicating with the network. Every eNB could be connected to nearby eNBs by means of an optional dedicated interface, named X2, as depicted in figure 2.1.

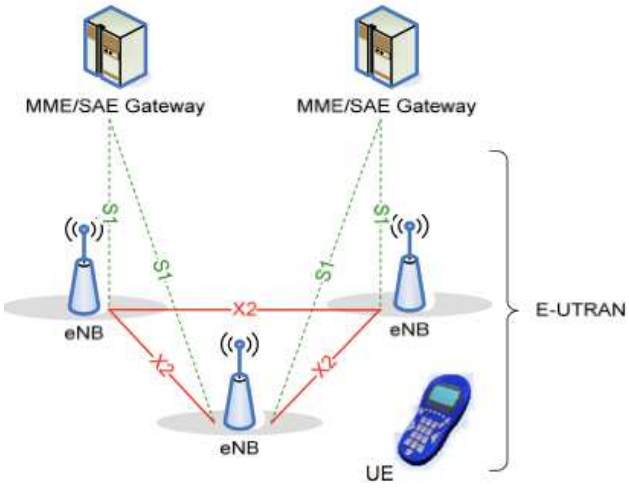

Figure 2.1: LTE network architecture

This new interface is mainly used for exchanging control signaling and forwarding data packets during handover, which results in a limited packet loss due to intra-eNB user mobility.

All eNBs are connected to the core network, called Evolved Packet Core (EPC), through S1 interface. The main components of EPC are shown in figure 2.2 along with their interconnectivity ([7], [8]). It consists of the following entities:

- Home Subscriber Server (HSS)
- Packet Data Network Gateway (P-GW)
- Serving Gateway (S-GW)

- Mobility Management Entity (MME)

The only component that has been carried forward to LTE from its predecessor systems, GSM and UMTS, is the HSS. It is the central database that stores all information related with the network's subscribers. All other core network components are new and each one serves its unique purpose.

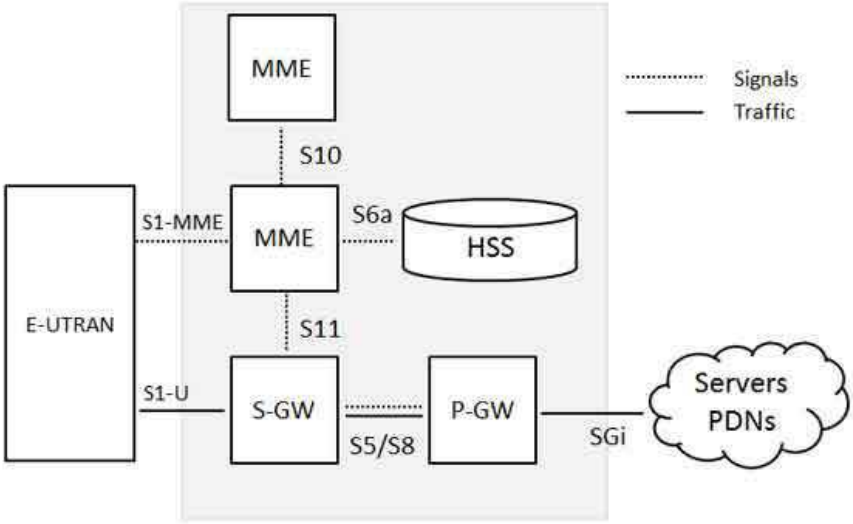

Figure 2.2: Main components of EPC

The P-GW provides connectivity with external packet data networks such as the Internet and IP Multimedia Subsystem (IMS). Every UE is assigned to a default P-GW when it switches on. Later on, it can connect to a different P-GW if it wishes to connect to a different external packet data network than its default.

The S-GW forwards data between the UE and P-GW. It acts as a router and it is responsible of all UEs of a certain geographical region. It also supports active session mobility between LTE and 2G or 3G systems. Each UE is connected to a unique S-GW at every time, but it may be assigned to a different one if it moves sufficiently far. The interconnection with E-UTRAN is implemented through S1 interface.

MME is the entity that performs session and mobility management. It connects with E-UTRAN via S1 interface and its main functions include:

- Authentication and authorization of UEs
- Performing handovers inside the LTE networks
- Selecting S-GW for the UE
- Allocating temporary IDs to UEs
- Handling mobility to other access networks.

In addition to S1 interface, MME connects to the other EPC entities through dedicated interfaces. Specifically it uses S6 interface to communicate with the HSS and S11 interface to connect with the S-GW.

MME and S-GW can be deployed as one physical entity or as two different physical boxes. In the latter case S1 interface is split in two separate parts. S1-U (User plane) carries user data between E-UTRAN and S-GW and S1-C (Control Plane) carries control information between E-UTRAN and MME.

2.2 LTE Network Protocol Architecture

2.2.1 LTE End-to-End Layer Stack

All the above mentioned functions of each LTE network node are implemented by an appropriately designed OSI-like layered model. The end-to-end stack of sublayers, from the UE up to the application server, is depicted in figure 2.3. A very large set of transport protocols may run in the application layer of both the UE and the application server, including UDP, TCP and RTP (Real Time Protocol).

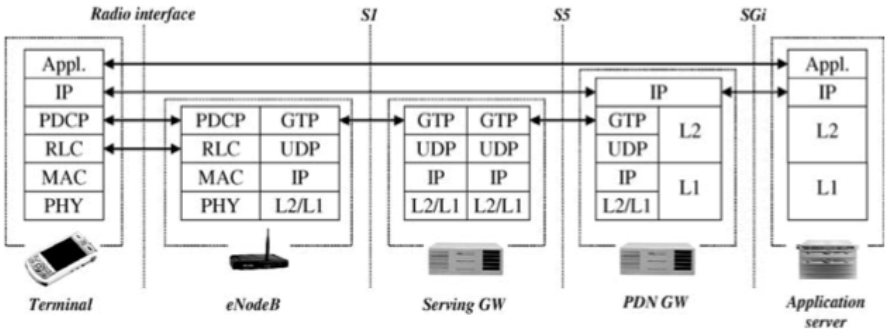

Figure 2.3: Representation of LTE protocol stack

The part of the transmission path between eNB and the P-GW is the fixed network part of the LTE architecture. The largest amount of research focuses on the radio interface between the UE and eNB, which is the mobile part. It distinguishes from all the other parts of the transmission path due to the scarcity of physical resources, as well as the higher error rates. The design of the protocol stack at this radio access part of network is very specific and supports the fully packet-switched operation of the LTE network. It can be divided in two planes, the user plane and the control plane.

At the user plane side the application creates data packets that are processed using common protocols such as UDP, TCP and IP. Control plane, on the other hand, uses Radio Resource Control (RRC) protocol to produce the required by lower level protocols configuration signaling. Both planes share the same underlying link and physical layer.

2.2.2 Link Layer Architecture

The main challenge for the LTE Link Layer Architecture ([9]) is to fulfill the Quality of Service requirements imposed by IP protocol. The characteristics of different IP data flows vary and therefore the link layer protocol overhead must scale. For example Voice over IP (VoIP) data flows can tolerate delays in the order of 100ms and packet losses up to 1%. On the other hand, real time video data flows cause high bandwidth consumption and are very sensitive to packet losses. BE traffic is tolerant to latency and packet losses while bandwidth consumption appears to be variable.

LTE link layer was designed so that all the above mentioned considerations are taken into account. It consists of three sublayers each one serving specific purposes. The Packet Data Convergence Protocol (PDCP) [10] is the upper sublayer as depicted in figure 2.7. The main functions of this sublayer are the following:

- It is responsible mainly for IP header compression and decompression so that packets that are transmitted through the radio access network have considerably

lower size. The protocol used for this task is the Robust Overhead Compression (ROHC) protocol. It's not mandatory for a UE to support this function except for the case of VoIP. A VoIP packet contains a payload of 32 bytes and the appended header is 60 bytes for IPv6 and 40 bytes for IPv4 resulting in an overhead of 188% and 125% respectively. The ROHC process compresses the header down to a size between 4 and 6 bytes which is equal to a relative overhead between 12.5% and 18.8%.

- It performs security functions such as ciphering and deciphering for both user plane and control plane data and integrity protection for the control plane data. A PDCP Protocol Data Unit (PDU) counter (COUNT) with a length of 32 bits is used as an input to the security algorithms. It is initiated at the beginning of the connection and is incremented with every PDCP PDU. The least significant bits (7 or 12 bits) form a field called PDCP Sequence Number (PDCP SN) and are appended as a header to the PDCP data PDU. This field enables the detection of a possible PDU loss. Integrity protection for the control packets is performed by adding a 32-bit field called "Message Authentication Code for Integrity" (MAC-I).
- It supports lossless mobility in case of inter-eNB handovers through appropriate reordering and in-sequence delivery of packets to the upper layers. At handover the header compression process is reset which means the COUNT value is set to zero. In a seamless handover, this means that some PDCP SDUs that have not been acknowledged yet by the source eNB would be lost. However, PDCP sublayer supports lossless handover by storing all the PDCP SDUs that have not been acknowledged and retransmitting them to the target eNB.
- It discards packets due to timeout. A timer starts for every PDCP Service Data Unit (SDU) received by the higher layers, and if its transmission has not yet been initiated by the expiration of the timer the SDU is dropped. The timer is set to a value that depends on the related QoS characteristics of the traffic flow that the SDU is part of. Discarding packets at this level may be an unavoidable solution in cases where the traffic rate of a service is considerably higher than the supported by the radio interface data rate, resulting in possible buffer overflow at the UE.

A typical PDCP PDU format is shown in the figure 2.4. The D/C field is used to distinguish between Data and Control PDUs.

Figure 2.4: PDCP PDU format

The Radio Link Control (RLC) sublayer [11] performs segmentation and concatenation on service data units (SDUs) that are received from PDCP. This means that RLC sublayer reformats the PDCP PDUs in order to fit them into the size of the allocated physical resources as indicated by the MAC sublayer. So the process of packet segmentation and concatenation is part of the scheduling decision process. It is also responsible for a part of the ARQ functionality, the in sequence delivery of packets and the duplicates detection. There are three modes of RLC functionality:

- The Transparent Mode (TM), in which the PDU passes through this sublayer without any processing. Since no overhead is added, a RLC SDU is directly mapped to a RLC PDU. The use of this mode is very limited and it is employed only in several control messages that do not need RLC configuration.

- The Unacknowledged Mode (UM), in which one or more RLC SDUs or segments of them are concatenated to form an RLC PDU. The size of the RLC PDU is indicated by the underlying MAC sublayer and a specific header is added that contains a Sequence Number (SN) and the boundaries of the specific SDUs or SDU segments within the PDU. The receiving RLC entity reorders the PDUs according to the SN and checks for duplicates. Some amount of out of order delivery is unavoidable due to errors in the transmission. Those that are received out of order are stored until all the previous PDUs are received. The duplicates are simply discarded. To avoid excessive reordering delays, a timer is initiated for every RLC PDU received out of order. If the timer expires before the missing PDUs are received, the out of sequence PDU is discarded. The RLC entity of the receiver reassembles the RLC SDUs from the in order received PDUs and delivers them to upper layers. Two PDUs might be needed to assemble an SDU as segments of it might be contained in different PDUs. If one of the PDUs that contains a SDU segment is discarded, the SDU is not assembled.
- The Acknowledged Mode (AM) which supports a bidirectional data service. The most significant feature of this mode is the retransmission of a PDU that was not received correctly. This feature makes the AM mode suitable for error sensitive and delay tolerant applications. The functionality is identical to the UM mode with the exception of the additional retransmission property. The retransmission of a PDU is triggered by an NACK message sent by the receiver side. Along with the RLC data PDUs, the RLC entity in the AM mode transmits control PDUs containing ACK or NACK messages depending on the reception status. The RLC control PDUs are created in the RLC sublayer itself unlike the RLC data PDUs that their content is forwarded from upper layers. A resegmentation of the original PDU may be required if the available resources at the time of retransmission are not adequate.

A typical RLC PDU format of the UM is shown in figure 2.5. The Frame Indicator (FI) is a 2-bit field that indicates whether the first and last SDUs contained in the data field are complete or partial SDUs. The Sequence Number (SN) field has a length of 5 or 10 bits and enables the reordering process of PDUs. The Length Indicator (LI) field contains the size information of a SDU inside the PDU. The number of these fields is variable as the number of the respective SDUs is variable. Each LI field corresponds to a certain SDU except for the first and last SDU that their size information may be deduced from the whole PDU size information contained in the respective MAC PDU. The presence of this field enables the reassembly of the SDUs at the receiver.

Figure 2.5: RLC PDU format (UM)

Finally the MAC sublayer [12] is the lowest sublayer of the link layer in the LTE radio interface protocol stack. It communicates with the underlying physical layer through transport channels and with the RLC sublayer above through logical channels. Multiplexing and demultiplexing between logical channels and transport channels is one of the most important functions of the MAC sublayer. It is implemented by constructing MAC PDUs which are also known as Transport Blocks. There is a one-to-one correspondence between MAC SDUs and MAC PDUs.

The main MAC sublayer function, which is also the subject of this thesis, is the scheduling of UEs. While the algorithms used to perform the scheduling are left to vendor implementation, the signaling that carries the scheduling commands is standardized. There are numerous algorithms proposed in the literature concerning how UEs are scheduled to transmit and how resources are allocated to them. Some of the most representative are analyzed in Chapter 4 while some newly proposed solutions are presented in Chapter 5. Scheduling algorithms are executed at this sublayer and their outcome determines what UE will receive in which downlink subchannel and how the resource grant requests for uplink transmissions will be satisfied. They can be very sophisticated and take into account several parameters such as channel quality, buffer status, QoS characteristics etc. Once the algorithm is executed, MAC sublayer produces the appropriate control signaling that is sent in the downlink to inform UEs about how the resources are allocated both in uplink and downlink directions and about the Modulation and Coding Scheme that will be used.

MAC sublayer is also responsible for managing the Hybrid Automatic Repeat re-Quest (HRAQ) function, which is an automatic retransmission of a physical layer Transport Block when an error is identified. More specifically the HARQ process is performed in combination by MAC sublayer and physical layer. Physical layer attaches a 24-bit Cyclic Redundancy Checksum (CRC) at the end of the transmitted transport block, allowing the receiver to detect probable transmission errors. MAC sublayer of the receiver produces a NACK message when the received transport block fails CRC. This causes a retransmission of the message using different channel coding by the transmitter side.

Finally, other MAC functions include the control of random access process which is the initial UE action to request resource allocation when it has data to transmit, the uplink timing alignment which makes sure that a UE uplink transmission reaches the eNB without overlapping with other UEs transmissions and the control of the Discontinuous Reception (DRX) which is a method of saving UE's battery by inhibiting reception process when monitoring downlink channels is not required. A conceptual block diagram depicting the MAC sublayer functionality is shown in figure 2.6.

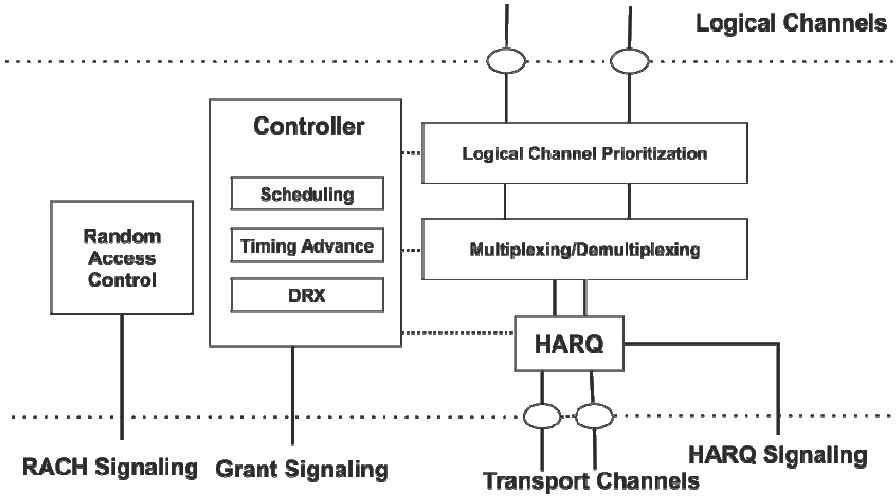

Figure 2.6: MAC sublayer block diagram

Figure 2.7 shows the data flow of an IP packet through the link level sublayers down to the physical layer. PDCP sublayer receives IP packets from the application level as Service Data Units (SDUs), compresses their header, adds a PDCP header and delivers the resulting packets to RLC sublayer as RLC SDUs. The RLC sublayer

segments and concatenates various RLC SDUs depending on the available radio resources and attaches a 4-byte header forming the RLC Protocol Data Unit (PDU). Finally MAC sublayer adds another 1-byte header to form the resulting transport block which is forwarded to the physical layer for transmission. The net overhead reduction that is achieved by the link level performing the abovementioned processing can be 1.5% for TCP/IP data packets, 55% for TCP/IP ACK packets and 42% for VoIP packets.

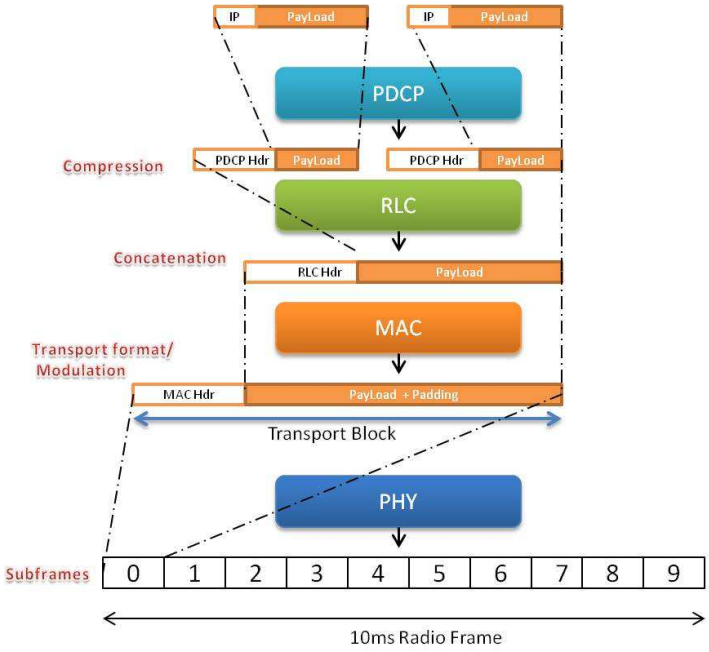

Figure 2.7: Life of an LTE packet

2.3 Quality of Service (QoS) in LTE

Quality of Service (QoS) expresses network's ability or probability to offer a desirable level of services for a specific amount of traffic. The QoS can be estimated using the following measures ([13]):

- Throughput: it is the achieved long term mean data rate. Regarding this feature, the various types of services in LTE are divided in Guaranteed Bit Rate (GBR) and non GBR. A Guaranteed Bit Rate, which is a target for its long term average data rate, is defined for every service of the former type. To achieve this, the system must make sure that adequate resources are allocated for this service at any given time. A Maximum Bit Rate (MBR) is also associated with every GBR service and it is the highest bit rate with which it is expected to receive. For non GBR services, a UE – Aggregate Maximum Bit Rate (UE-AMBR) is defined that limits the data rate on all the non GBR services of a UE, along with an Access Point Name – Aggregate Maximum Bit Rate (APN-AMBR) which sets an upper bound on the data rate of all the non GBR services of a UE that use a specific APN.
- Delay: it is an upper limit for the delay between the UE and the P-GW that a packet of a certain service can tolerate. In LTE, the lowest acceptable delay is 50ms and the highest 300ms.
- Packet loss: it is defined as the mean rate of packets that are lost due to different types of errors. In LTE there are services that tolerate various values of maximum packet loss rate ranging from 10^{-6} to 10^{-2}.

- Priority: for every LTE service a priority level is set my means of a parameter called Allocation/Retention priority (ARP). The ARP determines whether a request for resources may be rejected in cases of traffic congestion in the network. The scheduler for example may reject a certain traffic flow with low ARP to free up capacity, in order to serve a traffic flow with higher ARP.

Different values of the abovementioned bounds are assigned to different services in LTE. The network can identify a data stream of an individual service and classify it to a specific class from a set of predetermined classes ([14]). All the standardized classes for LTE are listed in Table 1. Every one of them has a unique QoS Class Identifier (QCI). It is an 8-bit identifier that distinguishes the various classes. The table also contains the specific values of the abovementioned QoS characteristics for every class, along with some example services that can be assigned to each one.

Table 1: Standardized QCI characteristics

QCI	Resource type	Priority	Packet delay budget	Packet error loss rate	Example services
1	GBR	2	100 ms	10^{-2}	Conversational voice
2		4	150 ms	10^{-3}	Conversational video (live streaming)
3		3	50 ms	10^{-3}	Real time gaming
4		5	300 ms	10^{-6}	Non-conversational video (buffered
5	Non-GBR	1	100 ms	10^{-3}	IMS signaling
6		6	300 ms	10^{-6}	Video (buffered streaming), TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
7		7	100 ms	10^{-6}	Voice, Video (live streaming), Interactive gaming
8		8	300 ms	10^{-3}	Video (buffered streaming), TCP-based (e.g., www, e-mail, chat, ftp, p2p file sharing, progressive video, etc.)
9		9	300 ms	10^{-6}	

To implement QoS, 3GPP has defined an extensive bearer model, the basic concepts of which are depicted in figure 2.8 ([15], [16]). A bearer is an established traffic flow between the UE and P-GW and constitutes the level of granularity for the QoS in LTE. Bearers provide a logical end-to-end transmission path and all packets that are delivered through a specific bearer have the same QCI and as a result receive the same QoS treatment (e.g. scheduling policy, queue management policy e.t.c.) [17]. A bearer that is established between a UE and a P-GW is called EPS bearer. As shown in the figure, separate bearers are also established in every part of the end-to-end transmission path. The one that is set in the over-the air connection between the UE and the eNB is called radio bearer.

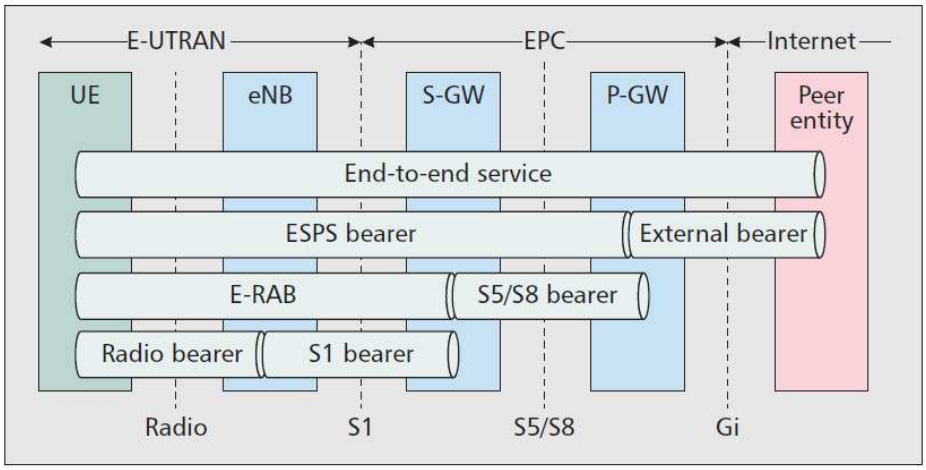

Figure 2.8: Bearers employed in LTE end-to-end service delivery

Bearers are separated into default bearers and dedicated bearers. A default bearer is set when a UE is initially attached to the LTE network. It is a non-GBR bearer, which means that QCI from 5 to 9 is assigned to it. Additionally an IP address is assigned to every default bearer. A UE can establish additional default bearers and each one of them will be assigned a separate IP address. The default bearer remains active as long as the UE is attached to the network. A dedicated bearer is established when there is a requirement for QoS by a specific UE service. The dedicated bearers may be either GBR or non-GBR. They are established on top of the default bearer. They are not assigned an IP and therefore they are always linked with an already established default bearer.

3. LTE-A PHYSICAL LAYER

The LTE specifications for high peak data rates, spectral efficiency and scalable bandwidth require a new approach in the design of the physical layer compared to the older systems. As a result of this, a new multiple access scheme is adopted which substitutes the older W-CDMA scheme that was used in 3G systems. Orthogonal Frequency Division Multiple Access (OFDMA) was selected as the new access scheme due to its improved spectral efficiency. In addition to this, LTE also implements multiple antenna techniques, known as MIMO, as a means of increasing channel capacity and improving signal robustness. OFDMA and MIMO are the two key technologies that LTE exploits to significantly boost the achieved data rates and will be briefly discussed in this chapter among other physical layer parameters.

3.1 Multicarrier Modulation and Multiple Access

3.1.1 OFDM

The multiple access scheme used in LTE is implemented through the use of a popular multicarrier modulation technique called Orthogonal Frequency Division Multiplexing (OFDM). It is based on dividing the available bandwidth to multiple narrow sub-bands. In each of the sub-bands, a subcarrier modulated using any of the available modulation schemes (BPSK, QPSK, 16QAM, 64QAM) carries an information symbol. Therefore symbols in OFDM are transmitted in parallel, which makes the total OFDM symbol length longer that the typical single carrier systems.

The parallel transmission of symbols is achieved by converting time samples to the frequency domain using FFT. The exact procedure of producing the OFDM signal is depicted in the following block diagram.

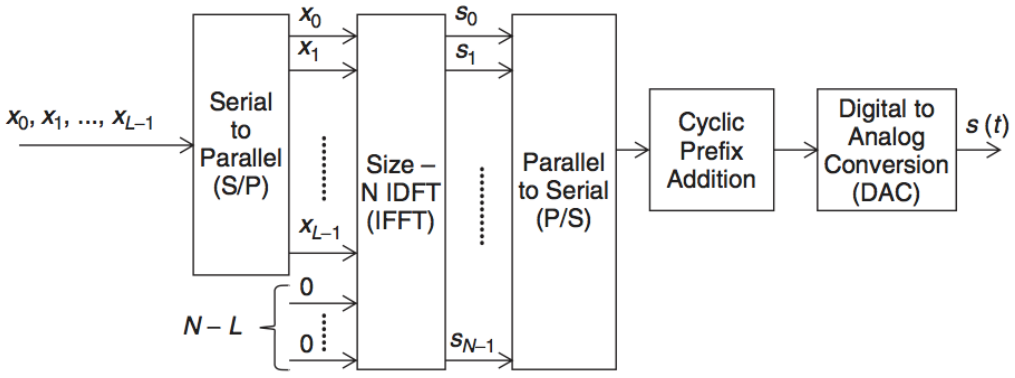

Figure 3.1: OFDM modulation block diagram

A series of modulated symbols are converted into parallel blocks of symbols. An IFFT is applied to each of these blocks. If the size N of the IFFT is larger than the length L of the block, appropriate zero padding is performed. The result of the transform is converted again to a series stream, and a cyclic prefix (CP) is added for reasons that will be explained later. This CP is usually a simple repetition of the last few samples of those produced by the IFFT stream, as shown in figure 3.2.

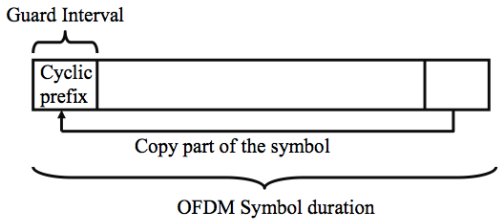

Figure 3.2: Cyclic prefix

Finally, a Digital-to Analog Converter derives the transmitted signal. The resulting spectrum of the OFDM signal is depicted in figure 3.3.

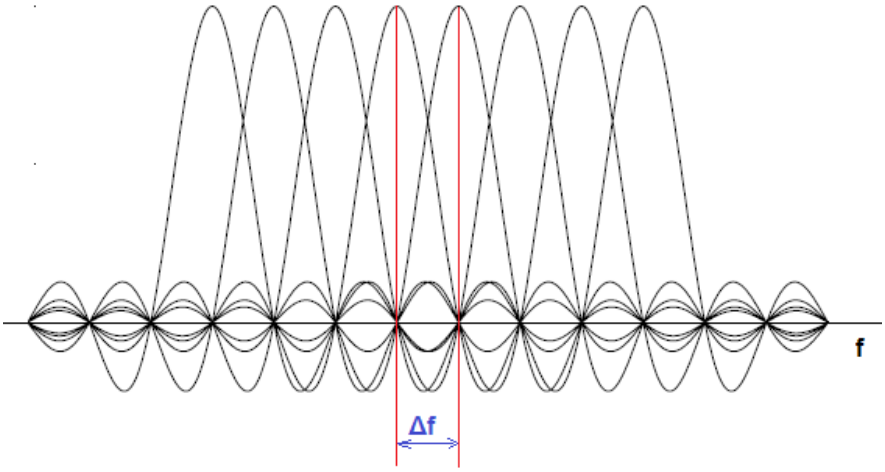

Figure 3.3: OFDM signal spectrum

There are several significant advantages of OFDM that led to its selection as the LTE modulation technique. The most important of them is its great performance in a frequency selective fading channel environment. In mobile wireless communications systems, the signal that reaches the receiver is usually distorted due to the multipath phenomenon. According to this, different copies of the transmitted signal reach the receiver, each one having its own delay and attenuation. This results in a frequency selective fading channel, meaning that the frequency spectrum of the channel impulse response is not flat. OFDM copes with this situation by dividing the bandwidth into narrow sub-bands each one having range lower than what is called channel's coherence bandwidth. Coherence bandwidth is the range of frequencies within which the channel response can be considered flat. Therefore the effect of the channel in each of the subcarriers separately can be considered flat in the frequency domain. This eliminates the need for channel equalization processing at the receiver, as it happens in single carrier systems.

In addition to this, the delay spread due to multipath introduces intersymbol interference (ISI). By this term, we mean that a delayed copy of the transmitted symbol interferes with a subsequent symbol causing possible errors in the demodulation process. The problem is solved by the insertion of the cyclic prefix, a guard interval introduced at the beginning of the transmitted symbol. This interval has sufficient length that is longer than channel's maximum delay spread. This prevents the distortion of the symbol's useful part from a delayed signal, making the OFDM symbol robust to ISI.

Another major advantage of OFDM is its high spectral efficiency. This is achieved by aggregating a high amount of subcarriers into a small bandwidth range. However this involves the risk of introducing inter-carrier interference (ICI), which is a cross-talk between subchannels. In OFDM this is avoided due to the orthogonality of the subcarriers that FFT imposes.

On the other hand, OFDM has two worth mentioning drawbacks. The first is its sensitivity to carrier frequency shifts. Small errors in the oscillators may cause loss of orthogonality between the subcarriers resulting in ICI (Inter-Carrier Interference). The second is that the resulting OFDM signal has very high Peak-to-Average Power Ratio. This requires, in the transmitter side, the implementation of an RF Power Amplifier with improved linearity properties and high power consumption. This may not be much of an issue at the eNB, but it is inefficient for a handheld mobile device. For this reason the multiple access scheme chosen for the uplink is not OFDMA, as it is for the downlink, but a variation of it called Single Carrier-FDMA (SC-FDMA).

3.1.2 OFDMA

OFDM modulation technique enables the allocation of different parts of the available subcarriers to different users. As a consequence, more than one UE can simultaneously transmit data during one OFDM symbol. This multiple access technique is called OFDMA and is used in LTE downlink.

The main advantage of OFDMA is the dynamic allocation of resources amongst all the UEs in a cell. The eNB scheduler can therefore implement channel dependent scheduling exploiting any probable knowledge about each UE's fading pattern. This means that it has the flexibility to assign to a UE only those subcarriers where the measured Signal-to-Noise Ratio (SNR) is relatively high. Those subcarriers, where the same UE receives weak signal, may be assigned to a different UE with better SNR. The resource allocation decisions are taken on a periodic basis so the system can also adapt to possible channel changes over time.

The concept of allocating different subcarriers of an OFDM symbol to different users compared to the regular assignment of the whole OFDM signal to an individual user is depicted in figure 3.4. It should also be noted that a UE could be assigned non-contiguous subcarriers in OFDMA. This is the case in LTE downlink, unlike the LTE uplink where, due to the use of SC-FDMA, a UE can only be assigned a contiguous part of the spectrum.

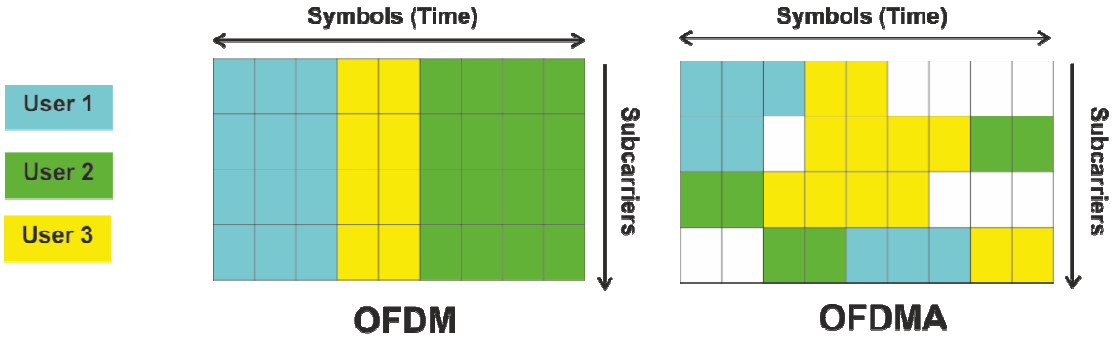

Figure 3.4: OFDM and OFDMA subcarriers allocation

3.1.3 SC-FDMA

For reasons that were previously analyzed, the multiple access technique used in LTE uplink is Single Carrier FDMA. The block diagram of the following figure shows how the SC-FDMA signal is produced in a UE. Compared to the OFDMA system, a single processing stage is added in each UE just before the application of the IFFT. This process is an L-point Discrete Fourier Transform (DFT) where L is the number of subcarriers that the UE has been assigned by the eNB. The DFT is applied on the modulated symbols that are to be transmitted and the resulting values are fed to the IFFT process. By adding this DFT stage, each symbol's power is spread over the whole number of available subcarriers.

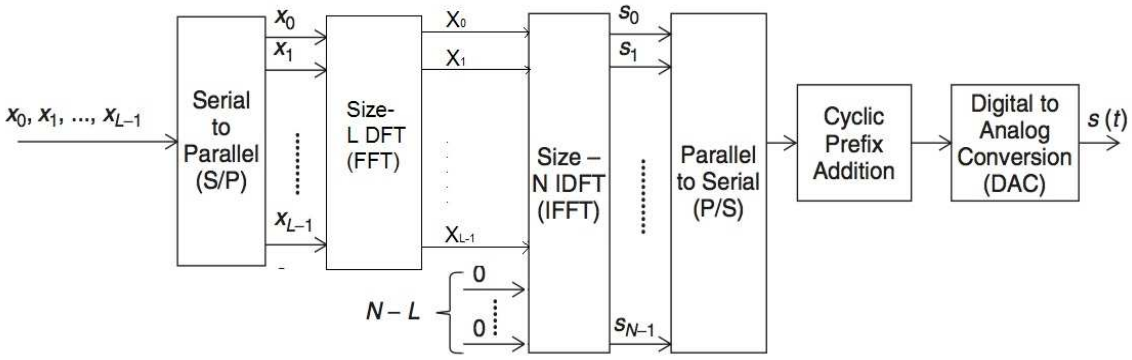

Figure 3.5: SC-FDMA block diagram

As a result, the final transmitted signal has lower PAPR and therefore the power consumption requirements for the UE are lower. The SC-FDMA signal lacks the multicarrier properties of the OFDMA signal and therefore is more susceptible to frequency selective fading. However the implementation of efficient channel equalization at the eNB is easier.

The employment of SC-FDMA imposes another restriction in the uplink resource allocation process since only contiguous parts of the spectrum can be allocated to the various UEs. Therefore LTE uplink scheduling becomes less flexible compared to the downlink case. Figure 3.6 illustrates the concepts of OFDMA and SC-FDMA.

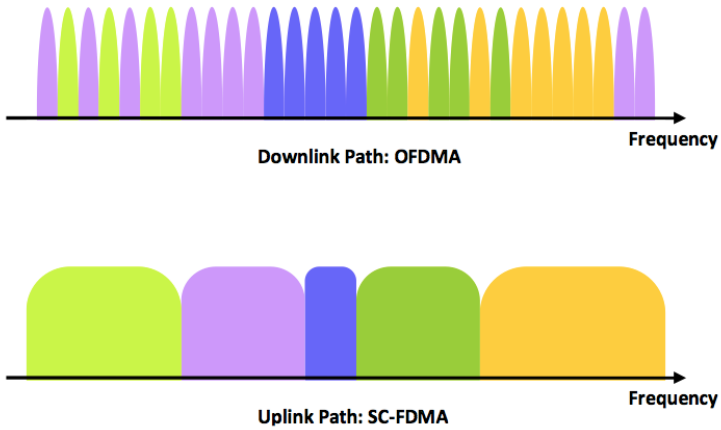

Figure 3.6: OFDMA and SC-FDMA resource allocation

3.2 LTE Physical Layer Parameters

3.2.1 LTE Generic Frame Structure

Based on OFDMA and SC-FDMA multiple access schemes, LTE physical layer is designed so that it exploits all the advantages of multicarrier transmission ([18]). The transmissions, both in downlink and uplink, are organized into frames ([5]). The duration of each frame is 10msec and is divided into 10 subframes of equal size. Each subframe is further divided into two slots with 0.5msec duration each. The following figure depicts this frame structure.

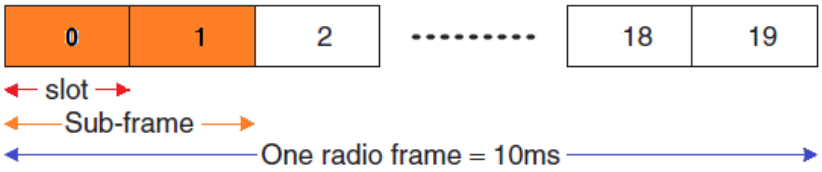

Figure 3.7: LTE frame structure

7 or 6 OFDM symbols are transmitted in each slot, depending on the cyclic prefix mode selected. There are two modes known as normal and extended CP mode respectively. As mentioned previously the length of the cyclic prefix must be longer than the delay spread caused by the multipath phenomenon. In normal mode the cyclic prefix has a length that is adequate for most of the delay spread scenarios that may be encountered in practice. For all the OFDM symbols the length of the CP is 4.69μsec except for the first symbol that has a length of 5.2μsec. Since the slot duration is equal to 0.5msec and each one of the 7 OFDM symbols has a constant length of 66.7μsec, the CP length of the first OFDM symbol had to be designed slightly longer. In exceptional cases where longer delay spreads are encountered, extended CP mode is used. The slots then contain 6 OFDM symbols so that the CP has a longer duration of 16.67μsec.

The bandwidth in LTE is scalable ranging from 1.4MHz to 20MHz, which means that the number of subcarriers is variable. The size of the FFT that produces the subcarriers varies from 128 to 2048. The subcarriers spacing is 15kHz. The higher sampling rate is achieved in the case of 20MHz bandwidth and calculated as the product of FFT size (2048) and subcarrier spacing, which is 30.72MHz. The inverse of this quantity (T_u = 1/30720000) is a basic time unit and almost all LTE time parameters can be expressed as multiples of it. For example a frame is $307200T_u$, a subframe is $30720T_u$ and a slot $15360T_u$. It should also be noted that scheduling decisions are taken on a subframe basis, and for this reason the subframe is also referred to as Transmission Time Interval (TTI).

Both Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) are supported by LTE. In FDD, uplink and downlink communication are performed simultaneously using different frequency bands while in TDD uplink and downlink use the same frequencies by time multiplexing. Two different types of frames are defined for each case. Type 1 frame is used in FDD where each communication direction uses its own frame. Type 2 frame is used in TDD and is shared by uplink and downlink as shown in figure 3.8. Half of the frame is dedicated to DL transmission while the other half to UL transmission. Type 2 frames also contain two special subframes. The special

subframe provides adequate guard interval between downlink and uplink transmission so that interference is avoided. It is always situated at the switching point from downlink to uplink. It contains three fields. The Downlink Pilot Time Slot (DwPTS) and Uplink Pilot Time Slot (UpPTS) are used for regular downlink and uplink transmission respectively. Both DwPTS and UpPTS have configurable duration. The Guard Period (GP) lies in between them and it is the field that essentially provides the required guard interval.

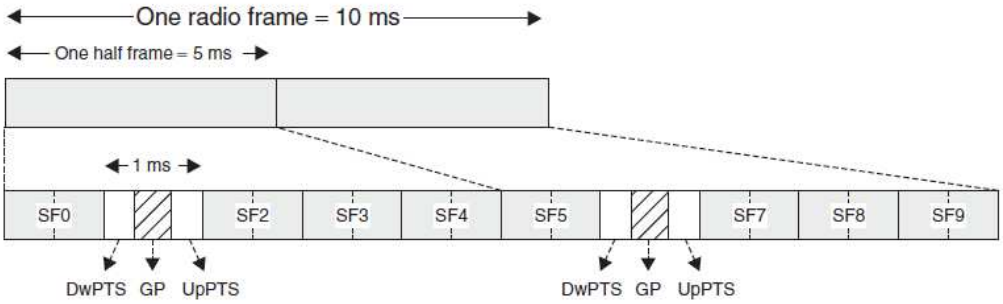

Figure 3.8: Frame structure type 2

3.2.2 Resource Grid

The subcarriers are grouped in frequency domain into Resource Blocks ([5]). A Resource Block (RB) contains 12 adjacent subcarriers and has duration equal to one slot as shown in figure 3.9. This means that a Resource block contains in the time domain 6 or 7 OFDM symbols depending on the cyclic prefix mode.

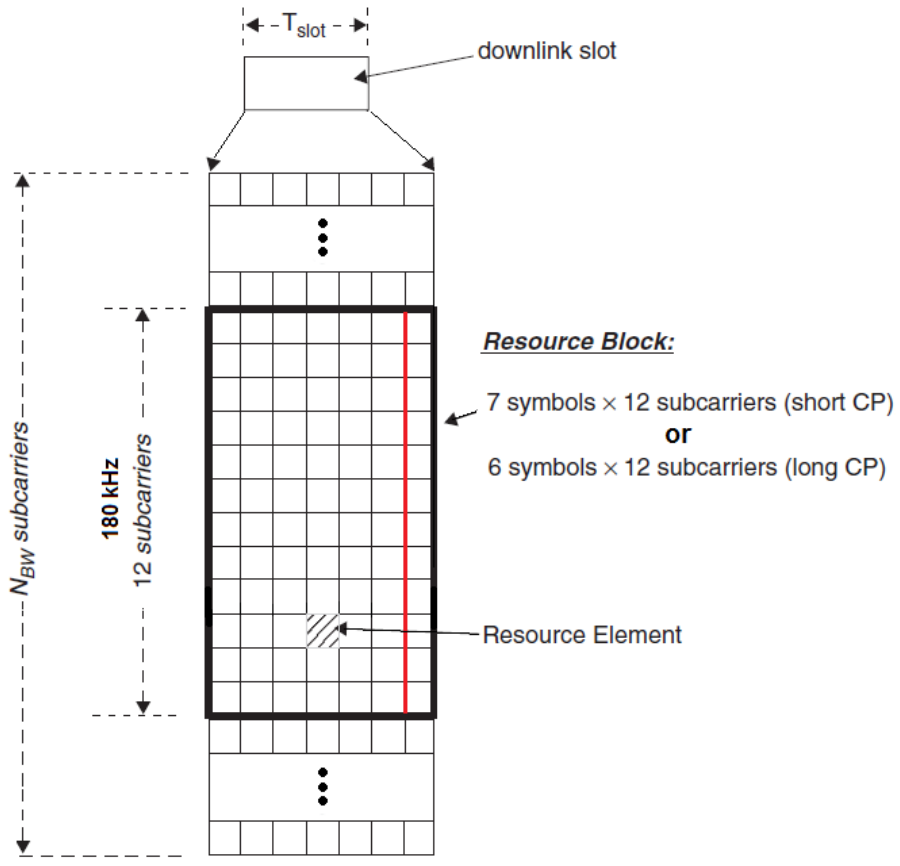

Figure 3.9: Resource grid

The smallest resource unit formed by one subcarrier and one OFDM symbol is called resource element and carries one information symbol.

Not all subcarriers produced by FFT are used. A specific amount of subcarriers that carry useful symbols is defined for each bandwidth, so that adequate number of subcarriers is left for guard frequencies in the two spectrum edges. The DC subcarrier remains also unused as it may be subject to undesirable leakage from the local oscillators. This results in a variable number of useful Resource Blocks that range from 6 RBs in the case of 1.4MHz to 100RBs in case of 20MHz.

Table 2 below summarizes the most important physical layer parameters for all different bandwidth options:

Table 2: Physical Layer parameters

Bandwidth (MHz)	1.4	3	5	10	15	20
Frame Duration (ms)	10					
Subframe Duration (ms)	1					
Subcarrier Spacing (kHz)	15					
Sampling Rate (MHz)	1.92	3.84	7.68	15.36	23.04	30.72
FFT size	128	256	512	1024	1536	2048
Number of Occupied Subcarriers (including DC subcarrier)	76	151	301	601	901	1201
Number of Guard Subcarriers	52	105	211	423	635	847
Number of Resource Blocks	6	12	25	50	75	100
Occupied Bandwidth (MHz)	1.14	2.265	4.515	9.015	13.515	18.015
Bandwidth Efficiency	77.1%	90%	90%	90%	90%	90%
OFDM symbols/subframe	7/6 (normal/extended CP)					
CP Length (normal mode) (μsec)	5.2 (first symbol)/4.69(6 following symbols)					
CP Length (extended mode) (μsec)	16.67					

3.3 Channel State Information

A very significant feature of LTE and all the modern mobile wireless communication systems is Adaptive Modulation and Coding (AMC). Its primary function is to adapt the transmission rate to the current channel conditions by changing the modulation scheme and the coding rate. To achieve this in LTE, a carefully designed channel sensing and reporting mechanism has been designed. It involves the use of pilot signals for channel estimation and a feedback mechanism that the receiver employs to inform the transmitter via what is called Channel Quality Indicator (CQI).

3.3.1 Reference Signals

Reference signals ([5]) are pilot signals that are inserted into the resource block grid and are used by the receiver to perform channel estimation. They are arranged in the time-frequency domain so that they allow correct interpolation of the channel. There exist different reference signals for the downlink and the uplink.

Three different types of downlink reference signals are defined: cell-specific reference signals, UE specific reference signals and MBSFN (Multicast-Broadcast Single Frequency Network) specific reference signals. Cell-specific reference signals will only be discussed as the others are out of the scope of this thesis. They are called cell specific because the reference symbols have complex values that depend on the cell identity along with the position of the symbol.

For the case of a single antenna at the transmitter, cell specific reference signals are inserted within the first and the third last OFDM symbol of each slot as shown in figure 3.10.

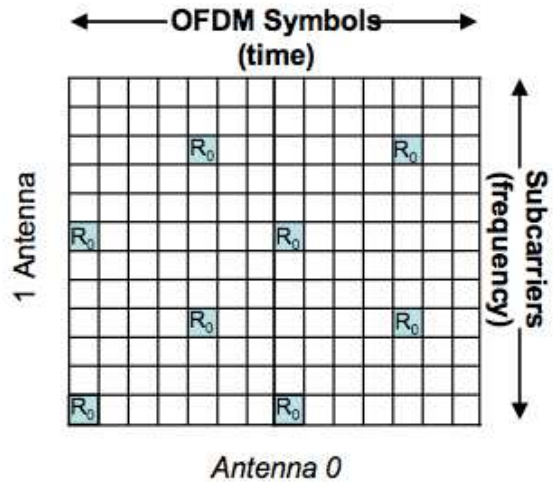

Figure 3.10: Reference signals for single antenna

For multiple antennas at the transmitter, additional reference symbols are added enabling the receiver to estimate the link quality of each individual antenna port. This means that the reference symbols are antenna port specific as well. Figure 3.11 depicts the positions of the reference symbols for the cases of 2 and 4 antennas. It should be added that at the time frequency location of the reference symbol of an antenna nothing is transmitted from the other antennas. This allows the receiver to make accurate estimations of each channel without the interference caused by the simultaneous transmission of other antennas reference signals. The presence of additional antennas (up to 8) in LTE-Advanced is supported by some special pilot symbols called UE-specific reference signals. These are only transmitted in the resource blocks where a multilayered transmission using these additional antenna ports is scheduled, and not in the whole spectrum. The exact position of these reference signals inside the grid can be found in [5], par 6.10.3.

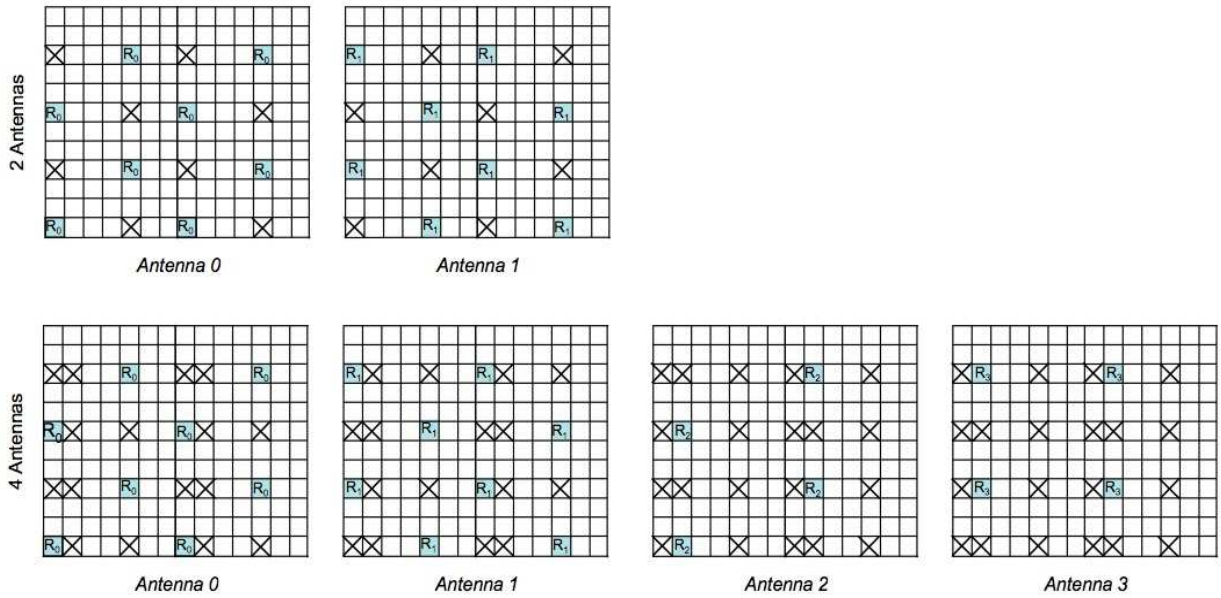

Figure 3.11: Reference signals for two and four antennas

Regarding the uplink reference signals, there are two types: the Demodulation Reference Signals (DM-RS) and the Sounding Reference Signals (S-RS). Both DM-RS and S-RS are shown in figure 3.12.

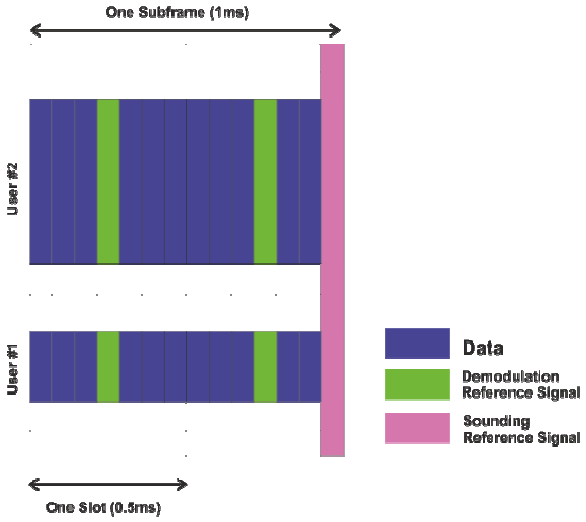

Figure 3.12: Uplink reference signals

The former are used for coherent signal demodulation at the eNB and the latter enable channel dependent scheduling for the uplink. The DM-RS are time multiplexed with the data and are transmitted on the fourth or third SC-FDMA symbol of the uplink signal for normal or extended cyclic prefix respectively. The S-RS is transmitted in a wider band, typically in the last SC-FDMA symbol of a subframe. It is an optional feature and a UE is configured by the eNB to transmit it. The transmission of this signal enables eNB to acquire channel knowledge about the whole bandwidth for every UE, since DM-RS signals are only transmitted in the allocated portion. When it is enabled, no data transmission takes place in this symbol.

3.3.2 Channel Quality Indicator

Channel Quality Indicator (CQI) is a link quality indicator and is essentially the information that is reported back by the UE to eNB ([19]). UE calculates this value through measurements using the abovementioned downlink reference signals. Several parameters, that constitute link quality measures, are estimated and the resulting CQI value is used by eNB to perform efficiently various functions such as time scheduling, resource allocation and mobility management. Some of those parameters are the following:

- SNR – Signal to Noise Ratio
- SINR – Signal to Interference Ratio
- SNDR – Signal to Noise Distortion Ratio

In LTE standard, CQI is 4-bit quantity that takes integer values between 0 and 15 as shown in Table 3. The table also shows how these integer values are interpreted in terms of downlink modulation and coding scheme (MCS). The last column shows the mean number of information bits per symbol (or per resource element) for each MCS. It is calculated by multiplying the number of bits per symbol, as defined by the modulation order (2,4,6), with the coding rate. The exact definition of CQI can be found in [19], par 7.2.3. According to this definition, a transport block with a modulation and coding scheme that corresponds to the reported CQI value can be received with a block error rate lower that 10%.

Table 3: 4-bit CQI table

CQI index	modulation	code rate x 1024	efficiency
0	out of range		
1	QPSK	78	0.1523
2	QPSK	120	0.2344
3	QPSK	193	0.3770
4	QPSK	308	0.6016
5	QPSK	449	0.8770
6	QPSK	602	1.1758
7	16QAM	378	1.4766
8	16QAM	490	1.9141
9	16QAM	616	2.4063
10	64QAM	466	2.7305
11	64QAM	567	3.3223
12	64QAM	666	3.9023
13	64QAM	772	4.5234
14	64QAM	873	5.1152
15	64QAM	948	5.5547

The UE measures the receiver channel quality and reports the channel dependent Channel Quality Indicator (CQI) in uplink to provide time and frequency variant channel information. UE reports the estimated CQI in various modes as configured by the eNB. There is the wideband reporting mode according to which UE reports a single CQI value for the whole downlink band. For the sub-band reporting, UE divides the whole band into sub-bands, estimates CQI for each of the sub-bands and sends all the values to the base station. This allows eNB to perform channel dependent scheduling. The last mode is the UE selected sub-band reporting. In this mode, UE selects the sub-bands with the

best channel quality and reports them back to the eNB along with a CQI value that spans them and a wideband CQI. A special case is when UE reports two CQIs and takes place when it can support the reception of two transport blocks during MIMO transmission.

Referring to channel state information, UE can also report two more indicators that are related with the sensed channel environment. These are the Rank Indicator (RI) and the Precoding Matrix Indicator (PMI) and will be briefly discussed in the next section.

3.4 Multiple Antennas Techniques (MIMO)

Along with OFDMA multiple access technique, the use of multiple antennas is a key technology in LTE systems that has a significant effect on data throughput maximization. Therefore deployment and optimization of MIMO techniques is an issue that gathers a huge amount of research effort. MIMO stands for Multiple Input Multiple Output and involves the usage of more than one antenna, in both the transmitter and the receiver, to send data on the same frequency resources. While traditional Single Input Single Output communications provide higher data rates under line-of-sight conditions, MIMO thrives under rich multipath conditions. The presence of multiple antennas in the transmitter or the receiver can be exploited in order to achieve different benefits such as:

- Multiple antennas at the transmitter and/or the receiver can be used to provide diversity against radio channel fading. The channels sensed by the different antennas should be uncorrelated in order to achieve the desirable diversity. So antennas should be placed in a sufficiently large distance (spatial diversity) or different polarization direction should be applied in each antenna (polarization diversity).
- Multiple antennas can also be used to shape the antenna pattern of either the transmitter or the receiver so as to achieve beamforming. This results in a higher antenna gain in the direction of the receiver, which can improve the achieved peak data rate.
- The simultaneous availability of multiple antennas in both the receiver and the transmitter can also be used for the creation of what can be seen as multiple channels that share the same space. As a result of this, multiple data streams can be transmitted from different antennas using the same bandwidth and the same space, which can also boost the achieved data rate.

Various transmission modes have been standardized for LTE and LTE-A and each one of them employs a different MIMO technique depending on the experienced channel environment ([20],[21]). Those techniques are transmit diversity, receive diversity, beamforming and spatial multiplexing and will be analyzed more in depth in the following paragraphs.

3.4.1 Transmit Diversity

Transmit diversity is the default transmission mode when multiple antennas exist at the transmitter. It aims to improve the Signal to noise Ratio (SNR) at the receiver and makes the transmission more robust to channel fading. In this mode, copies of the same signal are sent via all antennas. The receiver receives all copies of the signal and

applies diversity processing techniques to recover the actual signal as depicted in the figure 3.13.

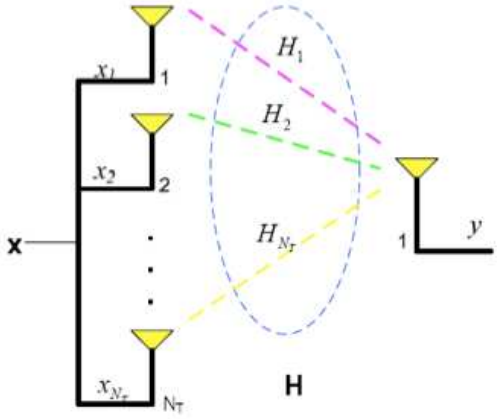

Figure 3.13: Transmit diversity principle

The two or more copies of the signal x that are transmitted may reach the receiver in a destructive way, meaning that their phases may be shifted in such a way that the signal y that is derived from the superposition of the received copies may be negligible. To cope with this situation some preprocessing in the transmitter is applied that makes sure the signals reach the receiver with the appropriate phases. There are two ways to apply this preprocessing. The first one requires feedback from the receiver and is called closed loop transmit diversity and the second does not require any feedback and is called open loop transmit diversity.

In the closed loop transmit diversity, the phases of the transmitted copies are shifted according to a reported by the receiver Precoding Matrix Indicator (PMI). The receiver uses the antenna specific reference symbols to estimate each individual channel response and then selects a precoding matrix indicator that is fed back to the transmitter. For example in the case of two transmit antennas a very simple PMI would indicate either that the two transmitted copies of the signal should be in phase or that one of them should be transmitted with a phase shift of 180^o. The disadvantage of this technique is that the feedback loop introduces time delays making it inappropriate in cases where the channel impulse response changes fast. Such cases are for example when the UE is moving so fast that the reported PMI may be out of date by the time it is used.

In the open loop transmit diversity no feedback from the receiver is required. However a specific preprocessing is executed in the transmitter by means of what is known as Alamuti's technique [22]. According to this technique, the transmitter sends two different symbols s_1 and s_2 at a specific moment using different antennas. At the next moment the same symbols shifted appropriately $-s_2^*$ and s_1^* are sent by the same antennas. The receiver can then make two successive measurements of two different combinations of the same symbols, solve the derived equations and recover the transmitted information. The only requirement is that the channel impulse response remains roughly stable during this interval.

3.4.2 Receive Diversity

In this mode, the receiver uses more than one antennas to detect two or more copies of the same signal, as shown in the figure 3.14. The signals arrive at the receiver with different phase shifts. However the receiver can remove those shifts by performing antenna specific channel estimation. After equalizing each individual channel separately, the receiver can add the derived signals avoiding the risk of destructive interference between them.

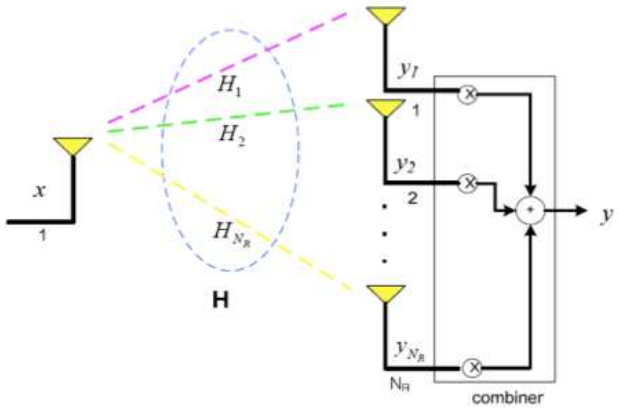

Figure 3.14: Receive diversity principle

As in the case of transmit diversity, the antennas in the receiver should be placed sufficiently apart so that the respective channels are uncorrelated. In such a case, it is unlikely that the channels that correspond to the different antennas will undergo fading at the same time. This makes the total received signal more robust and improves the mean experienced SNR.

3.4.3 Spatial Multiplexing

A totally different technique that can be deployed when both transmitter and receiver have multiple antennas is spatial multiplexing. This technique offers a huge leap in system's throughput potential as it enables the simultaneous transmission of more than one data streams from different antennas. The number of different data streams that can be supported by such a system is called rank in LTE. The peak data rate that can be achieved, compared to a Single Input Single Output (SISO) system, is almost multiplied by the rank value.

A simple spatial multiplexing model with N_T transmit antennas and N_R receive antennas is shown in figure 3.15. A $N_T \times N_R$ matrix H describes the total channel effect in the transmitted symbols. This means that H_{ij} expresses the total attenuation and phase shift that symbols transmitted from antenna j and received by antenna i undergo.

In a 2x2 spatial multiplexing system the output symbols r_1 and r_2 would be given by the following equations:

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \quad \Rightarrow \quad \begin{aligned} r_1 &= H_{11}s_1 + H_{12}s_2 + n_1 \\ r_2 &= H_{21}s_1 + H_{22}s_2 + n_2 \end{aligned}$$

where n_1 and n_2 represent the additive noise components. The receiver can estimate the matrix H by measuring the reference symbols sent by each antenna appropriately. As a consequence, the recovery of the different transmitted symbols s_1 and s_2 would be feasible if the abovementioned set of equations is solvable. This depends of course on how well-conditioned matrix H is, which would indicate how uncorrelated the various matrix rows and lines are and as a consequence how uncorrelated the different multipath channels are. A measure of how well conditioned matrix H is, would be its condition number. The receiver based on its estimations of matrix H, can estimate this condition number and derive a Rank Indicator (RI) that is reported back to the transmitter. This value represents the number of different data streams the receiver can separate. Easily one can conclude that the maximum reported value would be the minimum dimension of matrix H or $RI = min\{N_T, N_R\}$.

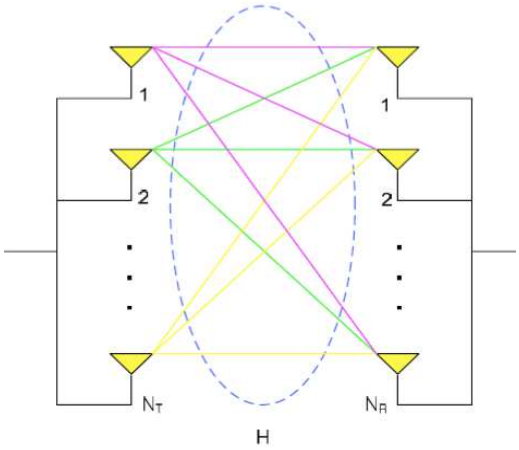

Figure 3.15: Spatial multiplexing principle

So if the channel conditions are favorable the receiver can recover more than one data streams. This technique is standardized in the LTE and is called open loop spatial multiplexing. Otherwise the system can fall back to the diversity mode where transmitter sends a single data stream from all its antennas. However, there can be cases where channel matrix elements are so badly behaved that it is impossible for the receiver to recover even one data stream. Such a case appears, for example, when the determinant of a square matrix H is zero. In this case, a technique called closed loop spatial multiplexing offers a solution.

As in the case of transmit diversity, an appropriate preprocessing of transmitted information can take place in the transmitter which can transform the initially bad behaved set of equations to a well behaved one, combined this time with a respective postprocessing. More specifically, the solution emanates from expressing the channel matrix H in terms of its eigenvectors and eigenvalues as follows:

$$H = P^{-1}\Lambda P$$

where P is the matrix produced from matrix H eigenvectors and Λ the diagonal matrix formed by matrix H eigenvalues. If the transmitted vector is multiplied by P^{-1} and the received vector is multiplied by P, the resulting set of equation will be the following:

$$r = PHP^{-1}s + Pn = \Lambda s + Pn$$

Ignoring the noise term, all the transmitted symbols can be recovered as long as the eigenvalues of matrix H are non-zero. In case that some eigenvalues are zero, the maximum number of different symbols that can be recovered equals the number of non-zero eigenvalues. If matrix H is not square, a similar processing can be executed using Singular Value Decomposition (SVD).

The receiver reports the matrix that should be used for the preprocessing, as a Precoding Matrix Indicator. Reporting all the possible matrices that are derived from the eigenvectors of every possible channel matrix H would require a large amount of control information. This is why the reported matrices are selected from a predetermined finite set of PMIs which are indicated by appropriate indices. UE sends the index that corresponds to the selected PMI instead of transmitting all the matrix coefficients, reducing the control overhead.

3.4.4 Beamforming

With this technique, the transmitter uses its multiple antennas to increase antenna gain by shaping the antenna pattern in the direction of the receiver. To perform this, the data and the reference symbols are transmitted from the different antennas using appropriate weights. The weights are chosen so that the received signals can be added constructively at the direction of the receiver and destructively in all other directions. Same technique can be used if there exist multiple receive antennas and one transmit antenna. By applying the appropriate weightings at the receiver antennas, we can make sure that the received signals of the different antennas add together in phase increasing the total gain.

Unlike the transmit diversity case, beamforming performs best when the antennas are close together, with a separation distance comparable to the wavelength of the signals. This ensures that the transmitted signals from the different antennas are highly correlated. This is in contradiction with transmit diversity and spatial multiplexing, that perform better when antennas are located far apart producing the desirable uncorrelated signals.

What is more interesting is that this technique enables the simultaneous communication with two different UEs using the same frequency resources by applying different weights in each data stream. This way eNB transmitter creates two different antenna beams each one pointing to the direction of the respective UE. Each antenna beam carries its own data stream and interference between them is avoided. This technique is called dual layer beamforming and is standardized in the latest releases of LTE.

3.4.5 LTE Transmission Modes

3GPP has standardized several transmission modes that implement the various multiple antenna techniques that were previously analyzed. The following table lists those predefined transmission modes as they are contained in Release 11 of the LTE standard ([19], [23], [24]):

Table 4: Transmission modes

Transmission mode	Transmission scheme
1	Single antenna
2	Transmit diversity
3	Open loop spatial multiplexing
4	Closed loop spatial multiplexing
5	Multi-user MIMO
6	Closed loop spatial multiplexing using a single transmission layer
7	Beamforming
8	Dual-layer beamforming
9	Eight-layer spatial multiplexing
10	Coordinated scheduling/ Coordinated beamforming

LTE, up to release 8, limits the number of antennas to 4 for the transmitter and 4 for the receiver, while LTE advanced, up to release 11, increases this limit to 8 antennas for both the receiver and the transmitter. This means that the maximum reported RI value would be 4 for LTE and 8 for LTE-A. However the maximum number of different codewords (transport blocks) that can be transmitted for both LTE and LTE-A is 2. Each codeword has its separate modulation and coding scheme and for that reason UE can report 2 separate CQI values. When the rank is higher than the number of codewords, each codeword is separated into what is called layers and this enables the construction of larger codewords. Each layer is then transmitted from a separate antenna. For example if one codeword is to be transmitted and rank value is 2, the codeword is divided into two layers by assigning the even symbols of the codeword to the first layer and the odd symbols to the second layer. All the possible combinations of layers, rank values and number of codewords are contained in [5], par 6.3.

3.5 Physical Channels

Control signaling and data flow between the different sublayers of LTE through channels. LTE has defined three different types of channels: logical channels, transport channels and physical channels.

Logical channels carry data and signaling between RLC sublayer and MAC sublayer ([12]). There are several logical channels distinguished by the type of data they transfer. They are also classified as control channels and traffic channels. Control channels are: Broadcast Control Channel (BCCH), Paging Control Channel (PCCH), Common Control Channel (CCCH), Dedicated Control Channel (DCCH) and Multicast Control Channel (MCCH). Traffic channels are: Dedicated Traffic Channel (DTCH) and Multicast Traffic Channel (MTCH).

Transport channels transfer data and signaling messages from MAC sublayer to Physical Layer ([25]). There are different transport channels for the downlink and the uplink. Channels used for the downlink are: Broadcast Channel (BCH), Downlink Shared Channel (DL-SCH), Paging Channel (PCH) and Multicast Channel (MCH).

Transport channels for the uplink direction are: Uplink Shared Channels (UL-SCH) and Random Access Channel (RACH).

Physical Channels are classified depending on the information they carry and on the way they are mapped to physical resources. The following Physical Channels are defined in LTE:

Physical Downlink Shared Channel (PDSCH): It is the channel that carries the information data that have to be delivered in the downlink to every user. It is used for delivering control messages and signaling, such as paging messages, as well. The data are delivered to the Physical Layer in blocks, called Transport Blocks (TB), which are formed every TTI (1 msec). Each TB corresponds to a MAC PDU and has variable size. It is transmitted with its own Modulation and Coding Scheme that depends mainly on the reported CQI of the UE that is destined for. The determination of the MCS is also a part of the scheduling decision. The size is determined by the amount of Resource Blocks that are allocated for its transmission and the selected MCS. The TB undergoes further processing in the physical layer that involves channel coding, rate matching and mapping to the allocated resources before being transmitted [25].

Physical Uplink Shared Channel (PUSCH): This channel carries data information that UEs transmit in the uplink. Data are also delivered from link layer to physical layer in TBs and undergo the same processing before being transmitted in the allocated RBs. The difference is that, due to the SC-FDMA used in uplink, those allocated RBs must be contiguous. Control data, such as UE channel state information reports, may also be transmitted in this channel multiplexed with the information data.

Physical Broadcast Channel (PBCH): This channel carries the so-called Master Information Block (MIB), a block of information that is broadcast to all UEs of a cell and is essential for the initial access. The MIB contains the downlink system bandwidth info, the most significant bits of System Frame Number and the structure of PHICH channel. Its transmission is spread over 4 frames and it is transmitted continuously with the updated information. Due to the fact that UE is unaware of the used bandwidth before the initial access, MIB is transmitted in the 6 center Resource Blocks which would constitute the RBs used in the minimum allowable bandwidth. These RBs are situated around the DC subcarrier and are common in all bandwidths.

Physical Control Format Indicator Channel (PCFICH): This channel carries the Control Format Indicator (CFI) message which determines the number of OFDM symbols of a subframe that contain the downlink control information. This number varies from 1 to 3 symbols except for the 1.4MHz bandwidth case where it varies from 2 to 4 symbols. CFI is mapped to very specific resource elements of the first OFDM symbol of each subframe. Every UE decodes this word and becomes aware of how many OFDM symbols of the current subframe contain control information, so that it can search this region for possible control messages that are destined for it.

Physical Downlink Control Channel (PDCCH): This channel delivers all the required control messages and configuration commands that need to be sent to UEs. These messages are known as Downlink Control Indicator (DCI) messages. The scheduling and resource allocation commands are included in these messages among other information such as power control, CQI report requests, multiple antenna transmission mode selection etc. Multiple DCI messages are transmitted in the first few OFDM symbols of each subframe, as determined by the information contained in PCFICH channel. The DCI messages undergo channel coding and a (Cyclic Redundancy Check) CRC field is generated using the Radio Network Temporary Identifier (RNTI) of the UE that the message is intended for. The resulting codewords are mapped to groups of resource elements called Control Channel Elements (CCE). The UE decodes the

PCFICH, contained in the first OFDM symbol, and determines the number of OFDM symbols that carry DCI messages. Then it performs a blind search over these OFDM symbols to find whether there is a DCI message for it. If there is one, it decodes it successfully using its own RNTI, otherwise UE is not scheduled in the current subframe. From the decoded DCI message, UE receives the information about what RBs are allocated to it. Then it can retrieve the transport block that is transmitted in these RBs. The following figure shows the resource grid of one subframe with three OFDM symbols used for control signals.

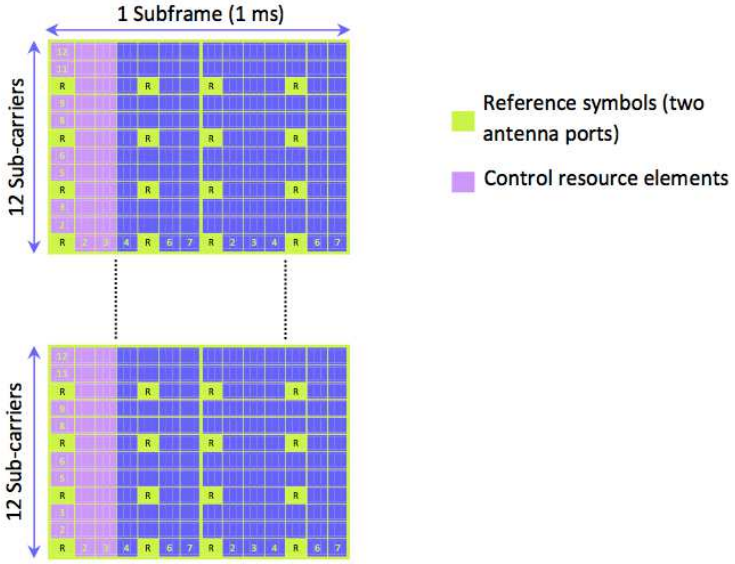

Figure 3.16: PDCCH signaling region

It should be added that a DCI for a specific UE can be transmitted in subcarriers where the UE may not have strong signal. This is the reason why the modulation scheme used to transmit PDCCH is QPSK and channel coding used to code DCI messages has a low rate.

Physical Hybrid ARQ Indicator Channel (PHICH): This channel contains HARQ ACK/NACK response messages that indicate whether previous PUSCH transmission has been successfully received or not. Multiple PHICH messages for different UEs can be mapped to the same resource elements and are multiplexed using code division. All PHICH information can be found in the first OFDM symbol of each subframe.

Physical Multicast Channel (PMCH): It is the channel used to carry the data from Multicast-Broadcast Single Frequency Network (MBSFN). MBSFN is a communication channel that can deliver services like mobile TV. PMCH is optional and if active it uses specific subframes where PDSCH is not transmitted.

Physical Uplink Control Channel (PUCCH): Uplink control data are carried in this channel such as the already mentioned Channel State Information (CQI, RI, PMI) and the HARQ messages. In addition, a UE can use this channel to send a scheduling request to the eNB. The RBs used by this channel are located at the two edges of the used bandwidth. The number of RBs that are dedicated to PUCCU is predetermined and increases as the bandwidth increases. Figure 3.17 shows a resource grid with 4 pairs of RBs dedicated to PUCCH transmission, 2 in each edge. A UE uses the PUCCH resources to transmit control information in a specific subframe, only if it is not allocated

PUSCH resources in it ([19], par 10.1). Otherwise it transmits control info multiplexed with data in the resources that has been assigned for PUSCH.

Physical Random Access Channel (PRACH): It is the channel that carries the initial uplink message that a UE sends to request access to the network.

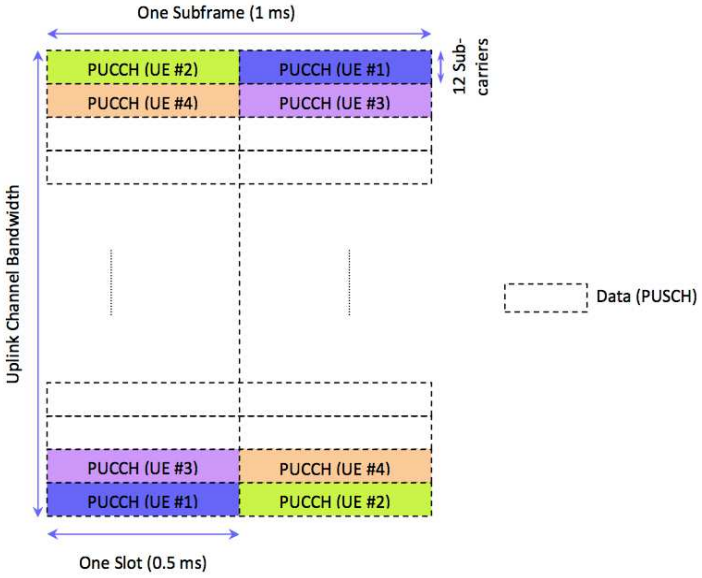

Figure 3.17: PUCCH signaling region

Depending on the content and purpose, logical channels are mapped to transport channels and transport channels mapped to physical channels. Figure 3.18 depicts the way different channels from different levels interact with each other in LTE.

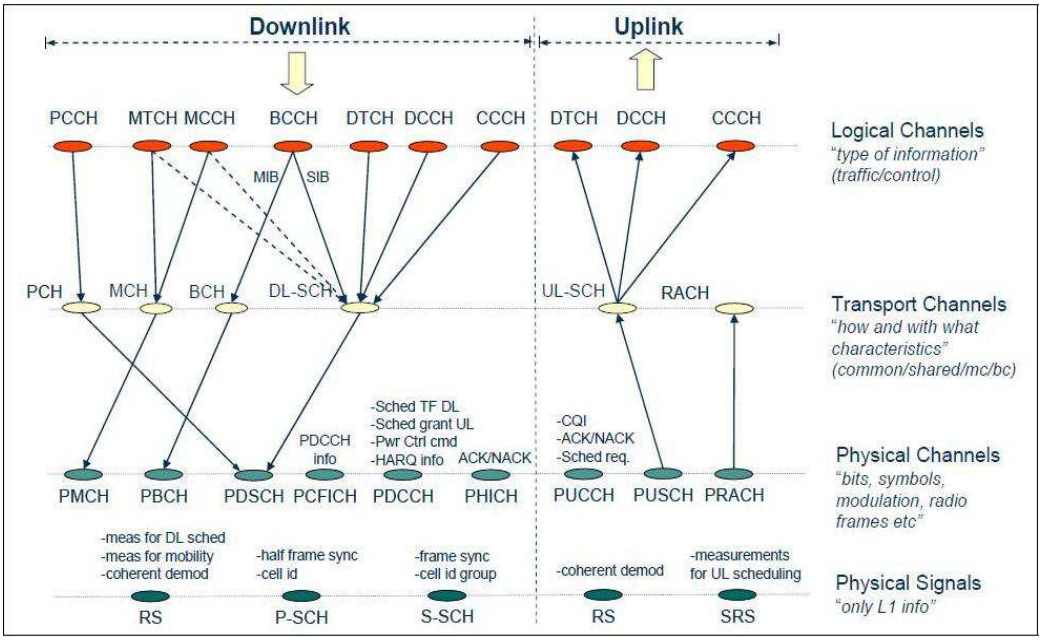

Figure 3.18: LTE channels structure and mapping

3.6 Resource Allocation Types

The purpose of resource allocation in LTE is to assign dynamically resource blocks to UEs in a way that maximizes system's performance. Efficient resource allocation can

significantly boost achieved data rate or other metrics that evaluate the services that are offered to users. The minimum resource allocation unit in LTE is a pair of resource blocks often referred to as a scheduling block. The pair consists of a resource block from the first slot of a subframe and a resource block from the second slot of the same subframe, usually the one that consists of the same subcarriers. However there is a case where the second resource block of a scheduling block contains different subcarriers for frequency diversity purposes.

There are three different resource allocation types in LTE known as resource allocation type 0, 1, 2 ([19] par. 7.1.6). These types have been defined so that the control overhead required to transmit a resource allocation command is minimized. As an example, consider the case of 20MHz bandwidth, where an eNB scheduler decides to allocate 10 out of the 100 available scheduling blocks to a specific UE. It then needs to send a 100-bit word with 1's in the positions that correspond to the 10 allocated scheduling blocks and 0's everywhere else. In general, it must use words with length equal to the number of available resource blocks so that it can identify any possible combination of resource blocks that can be allocated. This increases the control overhead, especially as the bandwidth increases.

3.6.1 Type 0 Resource Allocation

In type 0 resource allocation mode, resource blocks are grouped into Resource Blocks Groups (RBGs). The number of resource blocks that each RBG contains varies depending on the bandwidth. The following table shows the exact size of RBGs for every different bandwidth value:

Table 5: Number of RBs and RBGs

System bandwidth (MHz)	Number of RBs	RBG size (RBs)
1.4	6	1
3	12	2
5	25	2
10	50	3
15	75	4
20	100	4

The total number of RBGs can be calculated by dividing the total number of RBs by the RBG size. For example in the 20MHz bandwidth case the number of RBGs allocated is 100/4 = 25. If the remainder of the division is not zero, then the number of RBGs equals the smallest following integer of the result. In this case the last RBG would contain fewer RBs that equal the remainder value. For example in the 10MHz case, the number of RBGs is $N_{RBGs} = ceil(50/3) = 17$ and the last RBG contains $50\%3 = 2$ RBs. The figure below illustrates the type 0 resource allocation method for the case of 10MHz.

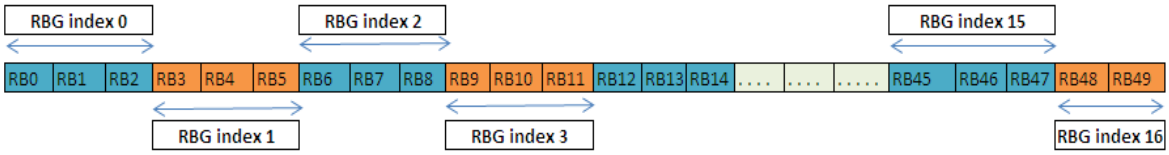

Figure 3.19: Type 0 resource allocation

The reduction in the control overhead is obvious. A 25-bit word is adequate to allocate any combination of RBGs to a UE in the 20MHz case, which is a significant overhead reduction compared to the 100-bit word that would be required if the granularity of 1 RB was used. The obvious disadvantage of using this type is that there would be a chance that some RBs be wasted. For example in the 20MHz case again, if 1 RB is enough for a UE to send its data, it would be allocated 1 RBG that contains 4RBs wasting 3RBs.

3.6.2 Type 1 Resource Allocation

The granularity limitations of type 0 resource allocation led to the standardization of type 1 resource allocation. The concept of RBGs is also used in this mode, however in this mode RBGs are also grouped into subsets. The size of RBGs remains the same and the number of subsets is also defined for every bandwidth value as shown in the following table:

Table 6: Number of subsets in Type 1 resource allocation

System bandwidth (MHz)	RBG size (RBs)	Number of subsets
1.4	1	1
3	2	2
5	2	2
10	3	3
15	4	4
20	4	4

The subsets are assigned an index p from 0 to $P-1$, where P is the number of subsets in the system. A subset with index p contains every Pth RBG starting from RBG p. Figure below shows the various subsets formed in the 20MHz case.

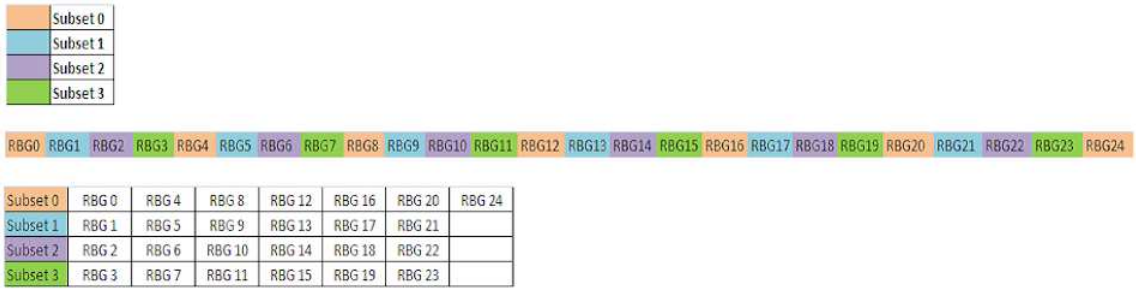

Figure 3.20: Type 1 resource allocation subsets

The words that identify the specific resource blocks to be allocated are different in type 1 resource allocation. The word in type 0 was a simple bitmap where each bit corresponds to a RBG. In type 1 the resource allocation word consists of three fields. The first field identifies the subset and requires $log_2 P$ bits. The second field is one bit that indicates whether a shift in the bitmap of the third field is used or not. The third field is a bitmap that represents the actual resource allocation within the subset. The length N_{RB} of this field is given by the formula:

$$N_{RB} = ceil(N_{RB}^{DL}/P) - log_2(P) - 1$$

where N_{RB}^{DL} is the total number of resource blocks. This number is smaller than the number of RBs of the subset $N_{RB}^{RBGsubset}$. The difference between the number of RBs in the subset $N_{RB}^{RBGsubset}$ and the number of RBs of the bitmap N_{RB} represents the value of the shift. This means that if the shift is active according to the second field, then the bits of the third field map to the RBs of the subset shifted by $N_{RB}^{RBGsubset} - N_{RB}$. If the shift is not active, then the bits of the third field map to the RBs of the subset without a shift, starting from the first RB of the subset.

The following example illustrates how type 1 resource allocation command is formed. Figure below shows the RBs of subset 0 in the case of 20MHz bandwidth. It consists of all the RBGs starting from RBG 0 and including every 4th RBG, since the number P of subset is 4 and each RBG consists of 4 RBs.

Subset 0	RBG0				RBG4				RBG8				RBG12				RBG16				RBG20				RBG24			
Subset 0	RB0	RB1	RB2	RB3	RB16	RB17	RB18	RB19	RB32	RB33	RB34	RB35	RB48	RB49	RB50	RB51	RB64	RB65	RB66	RB67	RB80	RB81	RB82	RB83	RB96	RB97	RB98	RB99

Figure 3.21: Type 1 resource allocation example

The number $N_{RB}^{RBGsubset}$ of RBs in the subset is 28. The length of the bitmap in the third field is $N_{RB} = ceil(100/4) - log_2(100/4) - 1 = 22$. So the amount of shift is 28-22 = 6. It should be noted that not all the subsets have the same length. For example, it is easy to show that all other 3 subsets in the case of 100MHz bandwidth consist of 6 RBGs, which means 24RBs. The shift value is then set accordingly, in this case for all other subsets is 24-22 = 2. For subset 0 with shift not active the bits in the bitmap of the third field represent the highlighted RBs shown in the figure below.

Subset 0	RBG0				RBG4				RBG8				RBG12				RBG16				RBG20				RBG24			
Subset 0	RB0	RB1	RB2	RB3	RB16	RB17	RB18	RB19	RB32	RB33	RB34	RB35	RB48	RB49	RB50	RB51	RB64	RB65	RB66	RB67	RB80	RB81	RB82	RB83	RB96	RB97	RB98	RB99

Figure 3.22: Type 1 resource allocation example

If the shift is active, then the bitmap is shifted by 6 RBs as illustrated below.

Subset 0	RBG0				RBG4				RBG8				RBG12				RBG16				RBG20				RBG24			
Subset 0	RB0	RB1	RB2	RB3	RB16	RB17	RB18	RB19	RB32	RB33	RB34	RB35	RB48	RB49	RB50	RB51	RB64	RB65	RB66	RB67	RB80	RB81	RB82	RB83	RB96	RB97	RB98	RB99

Figure 3.23: Type 1 resource allocation example

The length of the resource allocation word is 25, since 2 bits are used to indicate the subset, 1 bit for the shift and 22bits for the bitmap.

Using type 1 resource allocation the granularity decreases to 1RB, compared to 1RBG of type 0. Therefore resources are not wasted as it happens in type 0 resource allocation.

3.6.3 Type 2 Resource Allocation

Despite the considerable reduction in control overhead, type 0 and type 1 resource allocation modes still impose a control overhead that might be considered unacceptable especially in cases of small bandwidth values. To address this issue, type 2 resource allocation is also defined in LTE.

For type 2 resource allocation, the concepts of Physical Resource Block (PRB) and Virtual Resource Block (VRB) are introduced. Physical Resource Blocks are the actual Resource Blocks, as already presented, and they are numbered from 0 to N_{RB}, where N_{RB} is the total number of Resources Blocks of the current bandwidth. A single pair of PRBs is the minimum resource allocation unit as it has already been stated. VRBs are resource blocks that are mapped to PRBs in a certain way. There are two ways this mapping is performed resulting in two types of VRBs, the localized VRBs and the distributed VRBs.

The mapping from localized VRBs to PRBs is direct as illustrated in the figure 3.24 for the 5MHz bandwidth case. In this case, consecutive VRBs are mapped to consecutive PRBs.

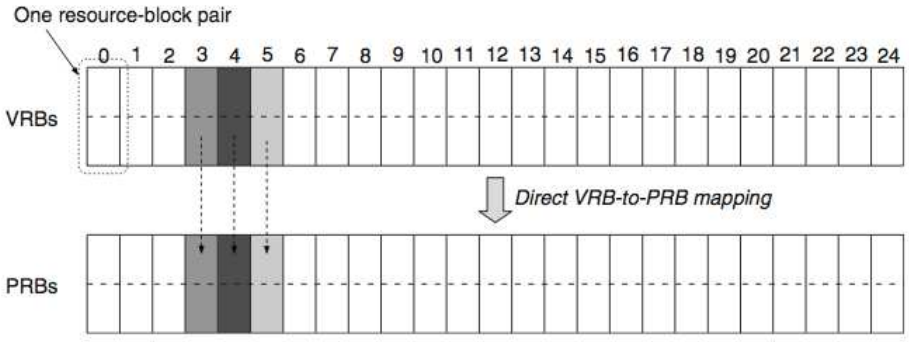

Figure 3.24: Localized VRBs mapping

In the distributed VRBs case, the mapping is not direct so that consecutive PRBs do not correspond to consecutive VRBs. In addition to this, a pair of VRBs corresponds to a pair of PRBs where each RB does not share the same subcarriers. This can be considered as an inter-slot hoping that provides frequency diversity. In fact there is a constant predetermined distance between the subcarriers of the first PRB of a pair and the second PRB. The exact mapping and inter-slot hoping are defined in [5], par 6.2.3. The basic idea of distributed VRBs is illustrated in the figure 3.25. The exact size of the distance between the two PRBs of a pair in the distributed VRBs case depends on the bandwidth. Its value is selected in the order of half the system's bandwidth, so that adequate frequency diversity is provided, and is also a multiple of P^2 [26], where P is the RBG size as previously explained. This constraint is imposed so that type 2 resource allocation commands could be sent in the same subframe with type 0 and type 1 resource allocation commands without any conflict. The table below lists the resource block distances for every bandwidth value:

Table 7: Gap distance for different bandwidths in type 2 resource allocation (RBs)

Bandwidth (RBs)	P (RBs)	Distance (RBs)
6	1	3
15	2	8
25	2	12
50	3	27
75	4	32
100	4	48

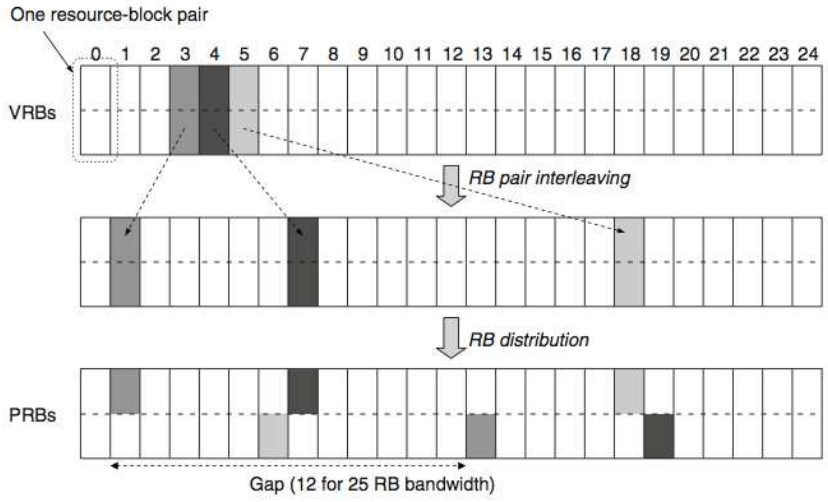

Figure 3.25: Distributed VRBs mapping

In addition to the distances specified in the table above, for the large bandwidth values there are second smaller distance values defined that can be used as shown in Table 8:

Table 8: Second gap distances for type 2 resource allocation (applicable only for RBs≥50)

Bandwidth (RBs)	P (RBs)	Distance (RBs)
50	3	9
75	4	16
100	4	16

The existence of smaller distance values allows the scheduler to restrict the distributed resource allocation to a smaller part of the bandwidth.

The DCI word that includes a type 2 resource allocation command contains special fields which denote all the above mentioned settings such as localized or distributed mode and distance value. For the actual resource allocation a single field, called Resource Indication Value (RIV) is used. The RIV is a function of the first VRB RB_{start} of the allocated VRBs and the length of the contiguously allocated VRBs L_{CRBS}. The specific formula used to calculate the RIV is:

$$RIV = N_{RB}(L_{CRBS} - 1) + RB_{start} \text{ if } \quad (L_{CRBS-1}) \leq \lfloor N_{RB}/2 \rfloor$$

$$RIV = N_{RB}(N_{RB} - L_{CRBS} + 1) + (N_{RB} - 1 - RB_{start}) \quad \text{otherwise}$$

where N_{RB} is the total number of RBs. The RIV value calculated by this formula uniquely identifies a pair of values, the one denoting the RB_{start} and the other the L_{CRBS}. A method that the UE can employ to reverse the RIV value and determine the allocated resource blocks can be found in [27].

The number of all available combinations is calculated as follows. If the starting VRB is the 0th then the number of combinations is N_{RB}, if $RB_{start} = 1$ then the number of possible combinations is $N_{RB} - 1$ and so on. So the total number of combinations is given by the sum:

$$N_{RB} + N_{RB} - 1 + N_{RB} - 2 + \cdots + 1 = N_{RB}(N_{RB} + 1)/2$$

The number of bits required to indicate all these combinations is $ceil\left(log_2\left(N_{RB}(N_{RB} + 1)\right)/2\right)$. It can easily be concluded that in Type 2 resource allocation the control overhead is significantly reduced. In the case of 10MHz bandwidth for example, with $N_{RB} = 50$, the number of bits required is $ceil\left(log_2(50(50 + 1)/2)\right) = 11$. It is reminded that 17 bits are used in both type 0 and type 1 allocation. So, using type 2 resource allocation results in considerable control overhead reduction. Finally, it should be added that type 2 resource allocation using localized VRBs produces contiguous allocated PRBs, which makes it ideal for the uplink case.

4. SCHEDULING IN LTE – ADVANCED

4.1 Introduction

In the first chapter we made a short review of the challenges that LTE-Advanced faces, where we pointed out the importance of improving its performance. Many considerable efforts and research has been made in order to improve the involved components that determine system's performance, among others: refining the used algorithms and technology. One of the most important factors that drastically affects system's performance is the management of radio resources and the scheduling of UEs, in other words the algorithms used in eNBs both for scheduling and resource allocation in order to satisfy the determined QoS requirements.

The objective of this chapter is to represent and address the scheduling problem in LTE-Advanced. The structure of the chapter is organized as follows: firstly we address the scheduling problem and all of its aspects for multiple users scheduling in LTE systems, then we make an illustration of the proposed algorithms found in the literature as well as the strategies that are being used in order to solve or to approach as much as possible the optimal scheduler. For each algorithm we highlight the parameters that are taken into account, as well as the pros and cons of each approach.

It is clarified that we deal only with scheduling algorithms that fit to the LTE and LTE-Advanced specifications, which as it has been aforementioned uses OFDMA schemes in downlink. Moreover we focus only in general purpose algorithms i.e. algorithms which do not specialize in increasing the performance for a certain traffic type such as video, voice etc.

4.2 Packet Scheduling in LTE-Advanced

In this paragraph we make a concise description of the packet scheduling problem, firstly a formal definition of the scheduler is given thereafter an analysis concerning the considerations and challenges that must be taken into account.

4.2.1 A Formal Definition of the Scheduler

In LTE-Advanced systems the scheduler is found in the MAC layer and its primary responsibility is scheduling and resource assigning, i.e. qualifying among the active UEs which are able to send and assigning to them spectrum appropriately.

Although in chapters 2 and 3 a detail description of the scheduling procedures was done, we did not define even in an abstract context what a scheduler is nor did we give a mathematical formulation. Below we attempt to give a formal definition for the scheduler of wireless LTE networks.

In wireless LTE-Advanced systems the scheduler can be considered as a rule which assigns resources to entities that make requests for those.

From a mathematical perspective the scheduler can be defined as a system $\boldsymbol{\mathcal{F}}[\cdot]$ which contains a set of resources i.e. SBs denoted by $\boldsymbol{S}$ and given a set of UEs $\mathbf{U} = [U_1, \ldots\ldots, U_N]$ a set of MCSs $\mathbf{M}$ and a set of preferences $\boldsymbol{P} = [P_1, \ldots\ldots, P_N]$, then it produces as a result the MCS and a subset of $\boldsymbol{S}$ for each UE for time period of T .i.e.:

$$[(S_1, m_1), \ldots\ldots, (S_N, m_N)] = \boldsymbol{\mathcal{F}}[\mathbf{U}, \mathbf{M}, \mathbf{P}, \mathrm{T}] \qquad 4.1$$

$m_j \in \boldsymbol{M}$

$S_j \subseteq \boldsymbol{S}, j = 1, \ldots., N \; and \; S_j = S_{j1} \cup \ldots. \cup S_{jN_j}$: $S_{jk} \; are \; SBs \; of \; U_j \; for \; connection \; k.$

$S_j \cap S_i = \emptyset \;\; \forall \, i \neq j, \;\; i, j = 1, \ldots., N.$

As a preference P_j of U_j we define: $P_j = (b_j, T_j, C_j)$ where b_j is a vector or a scalar representing the number of bytes per bearer that U_j wants to send within time period T_j, and C_j is a vector or a scalar of the QCI for each bearer of UE U_j.

4.2.2 Addressing the Scheduling Problem in LTE-Advanced

The above definition raises some fundamental questions: what should $\boldsymbol{\mathcal{F}}[\cdot]$ be, what the objectives it should meet are and what are the challenges needed to overcome.

We start our analysis by explaining how $\boldsymbol{\mathcal{F}}[\cdot]$ is usually derived for downlink. In order to derive a close type or a set of rules that fully determine $\boldsymbol{\mathcal{F}}[\cdot]$ three steps have to be done:

1. Determination of all the possible variables it has to take into consideration.
2. Complete determination of all the possible relations between the above variables as well as the possible constraints imposed by the system's specifications.
3. Determination of the objectives that it has to meet, i.e. what targets it should achieve.

The determination of all the variables involved in the scheduling process is done by taking into account the transmission medium and the quality of the radio channels, the different types of traffic specified in LTE-Advanced and system's specifications concerning scheduling. In the following we note down all the possible variables a scheduler takes into consideration:

- CQI report (covered in detail in chapter 3) which for each UE determines radio channel conditions, the received information contributes in calculating the rate a UE can achieve in a specific SB reliably i.e. BLER lower than a certain value.
- QCI for each active bearer which determines the QoS parameters (delay, packet error rate, GBR or NGBR etc), contributes in meeting the QoS requirements.
- Traffic volume for each bearer of each UE, the variables that are mostly used in order to determine traffic volume are: number of bits per UE or per bearer, the buffers fullness is another variable indicator of the data volume.
- HARQ retransmissions, suggest a good indicator for verifying the validity of the reported CQIs. Large number of retransmissions suggests that despite good CQI received values the actual conditions regarding the quality of the channel indicate that it is prone to errors.
- Transmission history for each UE, gives information about the achieved data rates. There are a variety of approaches on how to manipulate transmission history in order to achieve a certain policy.
- Limitations concerning transmission power, it is a significant parameter since it can affect interference between eNBs a factor that can increase or decrease systems overall capacity.

After the determination of all the variables that affect scheduler's design follows the step of extracting relations that occur from the specifications. We do not concern with describing the relations that occur for all the variables, instead we describe what the objectives that a good scheduler should meet are:

- Desired Complexity and Scalability: According to the specifications the scheduler takes a decision at the beginning of every TTI that means that the complexity issue is very significant making inappropriate solutions that perhaps find an optimal solution at the expense of complexity.
- Achieving performance as much as possible near to the optimal that is determined by a performance metric. There is a variety of performance metrics such as: throughput, fairness in time or frequency domain, experienced data rates, average or maximum packet delay experienced by a UE etc.
- Compliance with the QoS requirements: QoS provisioning is essential in LTE-Advanced networks so that it is important that the scheduler considers QoS constraints.

4.2.3 Modeling the Scheduler

The solutions to the scheduling problem in LTE are addressed in many different ways, the most prominent and rigorous ones found in the literature model it as a linear or non linear optimization problem with constraints. As it was clarified earlier as an optimization objective of the optimization problem a certain parameter is selected such as throughput, fairness etc. From the theory of linear or non linear optimization [28] we know that for an optimization problem with constraints it is possible that no feasible region is found that satisfies all the constraints, which implies that no solution exists. In this case there are different approaches, which lead to a suboptimal solution, most of which suggest that not all the variables are taken into consideration e.g. omitting power consumption constraints.

Another significant constraint even in case that a solution does exist is the complexity of the scheme, considering the number of factors it has to take into consideration and the tight time constraint imposed by the system [12]. For the time being we do not provide any specific formulation of the problem nor do we provide any solution, the interested reader may find some interesting formulations and solutions in [29] and [30]. Furthermore in [31] it has been proved that in OFDMA systems the problem of two-dimensional mapping of the incoming request to system's resources matrix (figure 4.1) even in its simplest form is NP-Hard so the problem of scheduling that is more complex than the previous one is NP-Hard.

In the light of the previous conclusion that the problem of scheduling itself is NP-Hard makes no surprise that instead of solving the optimization problem most of the approaches suggested in the literature confront the scheduling problem by separating the scheduling in the following discrete steps:

1. Determination whether a UE can be scheduled.
2. Choosing which UEs to be scheduled.
3. Prioritizing what need to be sent for each UE.
4. Allocating resources for each UE.

The decision pertaining which UEs are capable to transmit is based on whether the quality of the pilot signals is decoded correctly i.e. the received CQI values are acknowledged correctly. However the decision concerning which UEs should be scheduled and what data to transmit from each bearer is more complicated and is based on most of the variables mentioned above i.e. semi persistent scheduling, HARQ retransmissions., availability of data to send and QoS requirements. Different approaches can be made in order to make such decisions, even solving it as a simple optimization problem. The final step is to perform allocation of SBs to each UE: usually this is based on comparing a metric for each SB which aims in optimizing a utility function whose objective can be: fairness, throughput maximization etc.

The first three steps of the scheduler are usually referred as Time Domain – TD scheduling whereas the last step is being recognized as Frequency Domain – FD scheduling. In figure 4.1 we have depicted the combination of TD and FD for the LTE-Advanced scheduler.

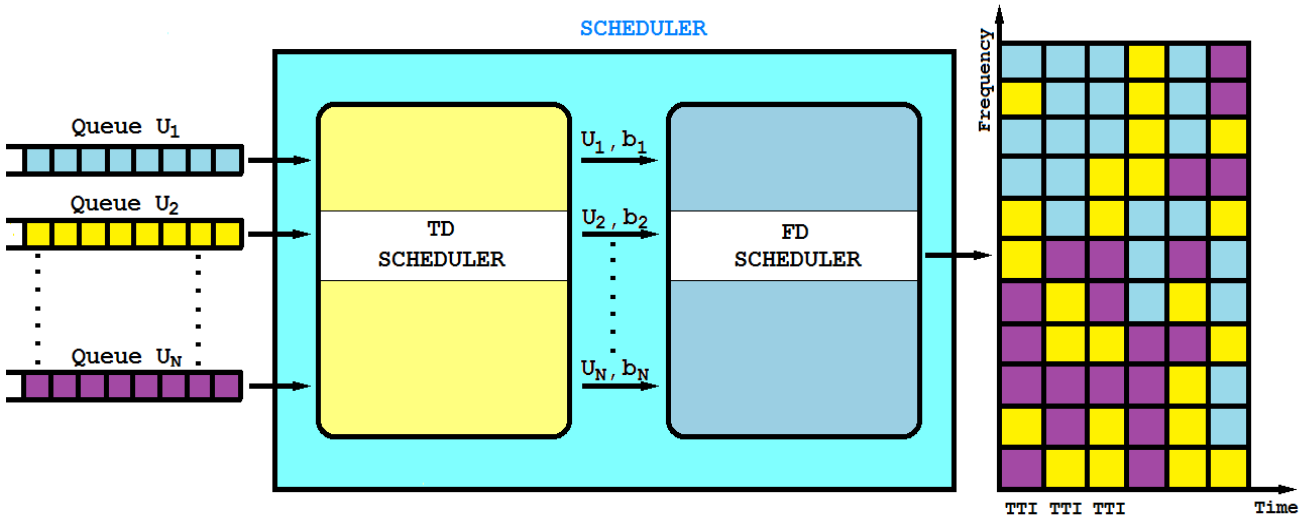

Figure 4.1: LTE-Advanced Scheduler Model.

4.3 Scheduling Strategies Found in Literature for LTE Downlink

In this section we describe a variety of scheduling and resource allocation strategies found in the literature for LTE and LTE-Advanced downlink wireless networks. Due to the fact that the proposed strategies take into account different variables such as input parameters, service targets and objectives, it is considered appropriate to classify them in different classes in order to simplify reading. Following [32] we group the strategies into the following classes:

- Channel Unaware.
- Channel Aware and QoS Unaware.
- Channel Aware and QoS Aware.

Firstly we choose to present channel unaware algorithms, which historically have been adopted in order to provide mostly fairness in the sense that all UEs are granted to be scheduled, subsequently channel aware but not QoS aware algorithms are presented and lastly channel aware and QoS aware algorithms are introduced and deeply analyzed.

4.3.1 Channel Unaware Algorithms

Algorithms classified in this class do not take into account the properties of the medium in which the signals are transmitted, thereafter it constitutes no surprise that the algorithms presented in this paragraph were initially proposed for wired networks [33]. Their application in LTE is not realistic but due to their simplicity and in combination with more sophisticated channel aware approaches, which improve their performance, render their use possible. Due to the vast channel unaware algorithms found in literature we focus only on the approaches that are considered more probable in combination with others to achieve realistic application in LTE-Advanced networks.

4.3.1.1 Round Robin – RR

The Round Robin – RR scheduler perhaps is the simplest scheduler found in the literature. The main idea in its approach is to be fair in the amount of time that each UE occupies the resources. Long term fairness can be achieved by assigning the available resources sequentially among the active flows, i.e. those connections which have a non-empty queue. The allocation of SBs is done as follows: if the number of SBs is greater than the number of active flows, then all the active flows are allocated in the same subframe. Otherwise not all the flows can be scheduled in a given subframe, in this case in the next subframe the allocation starts from the last not allocated flow. In case of using retransmission with HARQ the algorithm assigns the same SBs and MCS as in the first transmission. UEs designated for retransmissions are not considered for transmitting new data if a transmission opportunity is available in the same TTI.

In this way each user is guaranteed to be scheduled maximizing fairness, an obvious drawback of this approach is that QoS and CQI parameters are not taken into account resulting with high probability in low and unfair throughput. The advantage of this algorithm is its implementation simplicity. There are other two versions of RR, each of them adding a specific feature in the logic of RR, Weighted Round Robin and Deficit Weighted Round Robin.

4.3.1.2 Blind Equal Throughput – BET

This algorithm uses as metric the past average throughput achieved by each UE i.e. let $\bar{R}_i(t)$ be the past average throughput achieved by UE i until time t and $R_i(t)$ be the data-rate achieved by UE i until time t, then the current past average throughput is calculated by $\bar{R}_i(t) = \beta \cdot \bar{R}_i(t-1) + (1-\beta) \cdot R_i(t), \quad 0 \leq \beta \leq 1$

The metric for UE i and SB k is defined as: $m_{i,k} = \frac{1}{\bar{R}_i(t-1)}$ and it is calculated every TTI.

The algorithm uses the metric to allocate resources to flows that have been served with lower average throughput in the past. This policy allows UEs with low average past throughput to be served until they reach other users average throughput. Consequently UEs experiencing bad channel conditions are assigned SBs more frequently than UEs with good channel quality. Therefore throughput fairness is achieved in long term.

4.3.1.3 Guaranteed Delay

The main idea in this approach is to avoid packet drop, therefore each packet has to be served within a certain period of time, hence the name guaranteed delay. In order to achieve this objective a metric is defined that takes into account the time instant when the packet was created and the deadline determined by its QoS requirements. In [34] and [35] two different metrics were introduced whose objective is deadline expiration avoidance, the Largest Weighted Delay First – LWDF and Earliest Deadline First – EDF correspondingly.

The LWDF metric for UE i and SB k is defined as $m_{i,k} = a_i \cdot D_{HOL,i}$ where $a_i = -\frac{\log \delta_i}{\tau_i}$. δ_i is system's parameter which represents the acceptable probability for UE i that a packet is dropped due to deadline expiration. $D_{HOL,i}$: is the Head of Line Delay i.e. the delay of first packet to be transmitted by UE i and τ_i is delay threshold.

The a_i is a weight in the metric so that the UE with strongest intolerance in acceptable loss rate and deadline expiration will be favored for allocation.

The EDF metric for UE i and SB k is defined as $m_{i,k} = \frac{1}{\tau_i - D_{HOL,i}}$. This metric schedules the packet with the closest deadline expiration.

4.3.2 Channel Aware and QoS Unaware Algorithms

Algorithms classified in this category are those that take into consideration the properties of the medium in which signals propagate. Since the propagation environment is wireless that means that the channel is time-variant and prone to errors.

In LTE-Advanced this is done by reading the CQI feedbacks that UEs periodically send back to eNB, predicting in this way the MCS that a UE can receive with a specific BLER.

4.3.2.1 Maximum Throughput

This is another simple scheduling algorithm. The main idea in its strategy is to assign a particular SB to the UE with the best CQI in that SB. As a result the throughput of the system is maximized at the expense of fairness, since UEs with low CQI therefore with poor channel quality (typically those at the edge of the cell) will suffer from starvation as SBs will be scheduled mainly to those near the eNB's antenna. Specifically let $\mathrm{R_i(k,t)}$ be the achievable rate for UE $\mathrm{U_i}$ on SB k at subframe t, and if $\underline{i}_k(t)$ is the UE to which SB k is assigned, then $\underline{i}_k(t) = \mathrm{argmax}_{j=1\ldots.N} R_j(k,t)$. All the calculations are done by using the best MCS a UE can achieve.

4.3.2.2 Proportionally Fair – PF

The literature contains a vast number of papers and a variety of versions of the PF algorithm based on the variables they take into account. In this paragraph we will describe only the ones that are channel aware and for LTE-Advanced downlink only. The main concept of the PF algorithm is to maintain a balance between two competing interests: throughput maximization and fairness.

We describe an algorithm for LTE proposed in [36]. In this paper the authors propose a PF algorithm for multiuser scheduling, they approach the problem of scheduling by

formulating it as an integer linear problem with constraints. The linear problem which they formulate is a joint optimization problem for allocation of SBs and determination of MCS. Basically their approach finds a set of SBs for each UE (a SB allocated to a UE is not allowed to be assigned to other) and a MCS for each UE that maximizes the overall rate i.e. throughput. The authors mention that the problem can be solved by integer programming techniques, implying implicitly that its complexity is significant.

For that reason a suboptimal PF algorithm is also proposed which is formulated as well as a linear integer programming problem, the idea on the suboptimal algorithm is firstly to perform single user optimization i.e. for each UE to determine the MCS and a set of SBs that maximizes the rate without concerning that other UEs may choose the same SBs. Subsequently for each user a metric is constructed as follows $\varphi_i = \frac{r_M}{\bar{R}_i(t)}$ where r_M is the maximum bit rate achieved by UE i given that it has chosen $M_{optimal}$ as the MCS that maximizes the rate and $\bar{R}_i(t) = (1-a)\bar{R}_i(t-1) + a\, R_i(t-1)$ where $\bar{R}_i(t)$ is the mean rate for UE i in subframe t and $R_i(t)$ is the rate achieved by i in subframe t. Then the resource allocation is done in sequential fashion i.e. repeatedly assigns to UE that achieves the greatest value of φ_i the required number of SBs and removes it from the set of UEs. As in the previous algorithm no complexity evaluation has been done, it is just mentioned that it has lower complexity compared to the Joint optimization.

The assessment of the performance of the algorithms is done by comparing the two proposed algorithms with a Maximum rate algorithm which does not concern fairness. As comparing variables two metrics are used: Jain's fairness Index and Total average bit rate. With the assumption that UEs average SINR's are uniformly distributed the simulations show that the Joint optimization algorithm achieves a superior performance compared to the Maximum rate concerning fairness with a modest loss in throughput, whereas the suboptimal PF scheduler achieve fairness similar to Joint optimization with lower complexity sacrificing throughput.

4.3.2.3 Throughput to Average – TTA

This scheduling scheme [32] aims to achieve a performance intermediate between the two most used comparing and competing metrics in wireless systems: throughput and fairness. It can be considered as an intermediate algorithm between PF and Maximum Throughput.

Its main concept is in every TTI i.e. subframe to divide the available resources utilizing the following metric: $m_{i,k} = \frac{R_i(k)}{R_i}$ where $R_i(k)$ represents the achieved rate for UE U_i at SB k and R_i represents the expected achievable rate in the current TTI for U_i. The purpose of this metric is to quantify the allocation of a particular SB contributing in the allocation of best SBs to each UE, favoring that way fairness on a time window of one TTI. From the definition of the metric, it is obvious that if R_i is high then, its value will be low on a single SB enabling the scheduler to guarantee a minimum level of service to every UE exploiting channel awareness.

The TTA algorithm does not choose which UE will be scheduled making it appropriate only for resource allocation and not scheduling.

4.3.2.4 Delay Sensitivity

In this category we present delay sensitive schedulers, i.e. schedulers that take as key performance indicator the average data delivery delay.

A representative scheme that belongs in the above category is proposed in [30] where the authors propose a cross-layer algorithm to minimize the overall delay. Their strategy is to find a set of SBs and a MCS for each UE so as to minimize overall average packet delay taking into consideration channel conditions and traffic load. The problem of scheduling is formulated as a non linear optimization problem with constrains, regarding queue stability power limitations etc. Due to the fact that the optimization problem is hard to be solved with known optimization techniques a suboptimal algorithm is proposed. The suboptimal scheme distinguishes the resource allocation phase from MCS determination, each of which optimizes a given parameter independently. In the first phase SBs assignation is performed, the main idea in this stage is to spread SBs allocation to each UE among TTIs as much as possible selecting simultaneously the best channel for a UE. In the second phase a second scheme is used in order to determine MCS for each UE.

The authors do not make any estimation on the complexity of the suboptimal algorithm. Regarding its assessment they compare the proposed scheme with a conventional static one in which the number of SBs allocated to a UE satisfies rate proportionality meaning that the number of SBs allocated to a UE is $\left\lfloor \frac{w_n R}{\sum_i w_i} \right\rfloor$, as an assessing metric a rather unusual one is used: Overall Average Packet Delay versus Arrival Rate. From the simulation it is demonstrated that the proposed algorithm in various scenarios achieves substantial reduction in average delay compared to the conventional one.

4.3.2.5 A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-Advanced Technologies Using Neural Networks

In this [37] paper the authors address the problem of scheduling by proposing a scheduler aiming to assure a dynamic tradeoff between fairness and throughput during the transmission session taking into consideration the channel conditions for each UE. The innovation concept in their approach is that different scheduling rules are applied at each TTI in order to achieve convergence to the required tradeoff level between fairness and throughput i.e. achieving overall satisfaction.

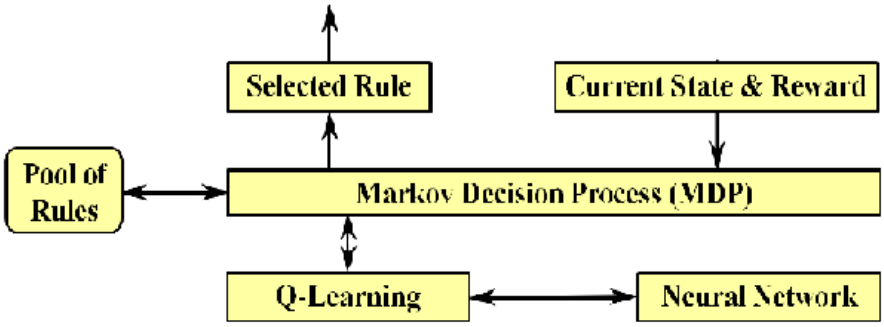

Figure 4.2: Framework of the Proposed Selecting Rule Scheme.

In order to achieve this: a Q-Learning algorithm with Neural Network whose diagram is depicted in figure 4.2 is proposed as the scheme that performs the scheduling rule adaption and refinement at each TTI. The scheme works as follows: it defines fairness, throughput, and percent of users located in different classes as state space (where Average Jain Fairness Index until time t is used as fairness metric, Average Normalized System Throughput until time t is used as throughput index and the classification in a specific class is done by calculating an average value of a channel quality metric for each UE which takes into account all the CQIs of the UE in the current TTI denoted by

t). Current state reflects the results of the applied rules in the previous TTI. Also a reward value is defined in order to quantify how the last action was from an aggregate objective point of view. It takes into account percentage of users that belong to a specific class as well as fairness and throughput as defined earlier. The transition from one state to another is done every TTI and is considered as a Markov process meaning that the current state depends only on the previous state. The reward value is determined by the transition from previous state to current.

The Q-learning algorithm, which is a type of temporal (works every TTI) difference reinforcement learning based on value iteration, aims in collecting as many rewards as possible in order to determine a policy of selecting the scheduling rules without any model of the environment. The mathematical approach in order to calculate the Q values is the same used in neural networks in order to converge to the target solution.

The authors make no estimation over the complexity of the proposed scheme. Considering the performance of the algorithm the fairness and the throughput described above are used as comparison metrics. The assessment of the proposed scheme is done considering different types of policies so as to highlight the tradeoff between system throughput and fairness. For this, six types of policies QP1 through QP6 (where QP1 aims to maximize the fairness metric and QP6 the throughput metric) are compared to each other and with the RR and PF algorithms. The results show that from the fairness point of view the QP1 policy shows the best result on the contrary QP6 shows the worst performance while the PF and the RR are found between those two policies with the PF exceeding RR. Regarding the throughput metric the QP6 outperforms all the others while QP1 shows the worst performance whereas PF and RR again are found between the two extreme cases.

4.3.3 Channel Aware and QoS Aware Algorithms

In this last class we classify algorithms that can be considered in a sense *complete*, since they take into consideration not only the properties of the channel success in achieving spectral efficiency but the QoS requirements for each bearer of a specific UE. The last parameter extends the concept of fairness, introduced in algorithms belonging to previous classes, among users not only in amount of time or rate that they did guarantee but in taking into account another variable that is related to different types of requirements that each connection demands. It is possible to quantify the requirements for each UE so as to introduce the concept of fairness and to guarantee some minimum performance requirements. This does not necessarily mean that QoS requirements will be met since it is possible, especially in a system with deficient admission control procedures, satisfying the QoS requirements to be unfeasible

In this section we give a comprehensive review of the QoS and channel aware algorithms found in the literature.

4.3.3.1 Modified – Largest Weighted Delay First

Modified- Largest Weighted Delay First (M-LWDF) is an algorithm that has been studied extensively and can be adjusted accordingly so that users with special QoS requirements are favored. It is a general scheduling algorithm extensively studied for single carrier systems ([38]) but it can be also applied in LTE.

The application of this algorithm is quite straightforward. At each TTI, the following metric is calculated for every scheduling block i and for every traffic flow k:

$m_{i,k} = a_k HOL_k \frac{r_{i,k}(t)}{\overline{R_k(t)}}$ where HOL_k is the current delay of the head of line packet of traffic flow k, $r_{i,k}(t)$ is the feasible rate that user of traffic flow k can achieve in scheduling block i and $\overline{R_k(t)}$ is the average rate of traffic flow k user up to time t. The coefficient a_k is a weighting factor that can be used to differentiate the various types of traffic flows by assigning properly chosen values. One popular choice for a_k is the following: $a_k = -\frac{\log \delta_k}{\tau_k}$ where δ_k is the maximum tolerated probability to violate the maximum allowed delay of traffic flow k and τ_k is the maximum allowed delay. Scheduling Block i is allocated to the user with the maximum value $m_{i,k}$.

It can be easily concluded that users with large HOL delay values are more likely to be served. At the same time fairness is also favored since the metric is proportional to the common PF metric. Simulation results in [39] prove that M-LWDF algorithm outperforms the common scheduling techniques in terms of providing the required quality to real time services such as real time video.

4.3.3.2 Exponential/Proportional Fair Algorithm

Another metric based algorithm, popular in single carrier systems and designed to provide QoS performance is the Exponential/Proportional Fair algorithm (EXP/PF). Though it has been initially designed and analyzed for single carrier systems, it can be also applied in multicarrier systems ([40], [41]), such as LTE. The associated metric has a different expression for real time and non-real time services as indicated by the following formulas:

$$m_{i,k} = \begin{cases} \exp\left(\frac{a_k HOL_k(t) - \overline{\alpha \cdot HOL(t)}}{1+\sqrt{aHOL(t)}}\right)\frac{r_{i,k}(t)}{R_k(t)}, k \in RT \\ \frac{w(t)}{M(t)}\frac{r_{i,k}(t)}{R_k(t)}, k \in NRT \end{cases}$$

$$w(t) = \begin{cases} w(t-1) - \varepsilon, HOL_{\max} > \tau_{\max} \\ w(t-1) + \frac{\varepsilon}{c}, HOL_{\max} < \tau_{\max} \end{cases}$$

where $M(t)$ is the average number of RT packets in the buffer of the eNB at time t, ε and c are constants, $w(t)$ is an adaptive factor defined for non-real time users, HOL_{max} is the maximum head of line packet delay among all RT traffic users and τ_{max} the maximum delay constraint among all RT traffic users. The coefficient a_k is as defined in the M-LWDF metric and the ratio $r_{i,k}(t)/R_k(t)$ is the common PF metric. The application of this algorithm in LTE is accomplished by simply allocating scheduling block i to user k that maximizes metric $m_{i,k}$. This results in providing RT users with priority over NRT users when their HOL packets delays are approaching the maximum allowed limit.

4.3.3.3 Logarithmic – LOG and Exponential – EXP Rule Algorithm

Scheduling according to logarithmic (LOG) or exponential (EXP) rule is a very common QoS aware technique. According to the logarithmic rule, Scheduling Block i is assigned to user k that maximizes the following metric: $m_{i,k}^{LOG} = b_k \log(c + a_k HOL_k) \cdot r_{i,k}$

where b_k, c, a_k are adjustable parameters, HOL_k is the current delay of user's k head of line packet and $r_{i,k}$ is the maximum achievable rate that user k can achieve transmitting on Scheduling Block i based on the channel quality.

According to the exponential rule, scheduling is performed in a similar way based on the metric:

$$m_{i,k}^{EXP} = b_k \exp\left(\frac{a_k HOL_k}{c + \sqrt{\frac{1}{N}\sum_j^N HOL_j}}\right) \cdot r_{i,k}$$

where N is the number of users and all the other parameters are as defined for the exponential rule case.

As can be seen from the formulas above, both algorithms are channel and delay aware. Specific values for the adjustable parameters where optimal performance is attained are proposed in [42]. Simulations results in the same paper suggest that EXP rule has a more robust performance compared to LOG rule.

4.3.3.4 Delay Prioritized Scheduling – DPS

A very simple QoS aware algorithm, applicable to LTE system, is proposed in [43]. It is called Delay Prioritized Scheduling and its application favors users whose HOL packets are approaching the violation of their maximum delay limit. This makes the algorithm suitable for delay sensitive services such as real time video. One of its most significant advantages is that is has very low complexity which is one of the critical issues in LTE, given that the period of the scheduling decisions is very short (TTI = 1msec).

The algorithm is completed in three very simple steps that are executed before every scheduling decision. In the first step, the remaining time for the HOL packet of every existing user i is calculated by simply subtracting the current HOL delay value from its maximum allowable limit: $m_i = HOL_{max,i} - HOL_i$ where $HOL_{max,i}$ is the maximum allowable delay of the Head Of Line packet of user i, while HOL_i is its current delay value. In the second step the user with the lowest value m_i is selected as the user that is closer to violate the maximum delay bound. Then in the third step, this user is assigned the SBs with the highest instantaneous SNR according to its transmitted uplink reports. These SBs are removed from the set of available SBs. Then steps 2 and 3 are repeated until all SBs have been allocated or all users have been served. Simulation results in [43] verify its improved performance in serving real time traffic.

4.3.3.5 Configurable Dual Mode Algorithm

Most of the algorithms proposed in the literature adopt a certain scheduling pattern concerning the way resources are allocated. The allocation is performed either user-by-user or SB-by-SB. The former pattern means that the users are prioritized according to certain metrics and then starting from the user with the highest metric the SBs that

serve its traffic efficiently are allocated. In the latter pattern, a SB is assigned at each iteration to the user that maximizes a certain metric until all SBs have been allocated. In [44], those two different approaches are combined appropriately so that benefits from using either of the scheduling patterns are gained.

In particular, two different modes of scheduling are defined, the Emergency (E-mode) and the Normal (N-mode) mode. Every time, one of the two modes is used depending on the status of the transmission queues as far as QoS is concerned. At each TTI the scheduler checks the status of the transmission queues and determines if there is a certain queue where the head of line (HOL) packet approaches expiration. A predetermined threshold D_{th} is used to quantify this, by calculating the difference between the maximum allowed delay D_{max}, as indicated by the respective QCI, and this threshold. If the current delay of the head of line packet exceeds this difference, E-mode is enabled.

The main concern of the scheduling process in the E-mode is making sure the QoS requirements are fulfilled. In this mode, the resources are allocated to the users whose transmission queues fulfill the emergency mode prerequisite and also experience favorable channel conditions. By this we mean that channel quality reported by these users should also exceed a specific predetermined CQI_{th}. The allocation is then performed on a user-by-user basis. The metric used to prioritize the users is the priority value that corresponds to the QCI according to table 1, in Chapter 2. Priority among users with equal QCI is chosen randomly. Each user is allocated the SBs with the best channel quality from the set of the SBs that have not been already allocated.

If there is no traffic that enables the E-mode or all the existing emergency traffic has been served during E-mode phase, the remaining SBs are allocated according to the N-mode. This mode allocates resources on SB-by-SB basis, and the main concern in the scheduling process is not guaranteeing QoS but ensuring resource utilization and fairness. For this reason, SB i is allocated to user k that maximizes the following PF metric: $P^{i,k} = \frac{CQI_{i,k}^{\alpha}}{\bar{R}_k}$ where the denominator is the average rate metric that is usually calculated in PF algorithms and α is a configurable parameter that is used to control the trade-off between fairness and throughput maximization.

Simulation results presented in [44] prove that the abovementioned adaptive algorithm, called Configurable Dual (CD) algorithm, outperforms considerably the popular PF algorithm in high traffic situations, while at the same time shows comparable performance under normal traffic. The impacts of delay threshold D_{th}, CQI threshold CQI_{th} and α on the CD algorithm performance are also illustrated. By extensive simulations specific values of these parameters are determined, where the ideal tradeoff between throughput maximization and fairness is achieved.

4.3.3.6 QoS Oriented Time and Frequency Domain Packet Scheduler

Several algorithms proposed in the literature separate the problem into two phases in order to provide QoS aware scheduling: the time domain and the frequency domain phase. A representative example of this approach is presented in [45] where scheduling is performed in time and frequency domain sequentially with each phase serving its own purpose. Before executing those steps, an initial schedulability check is performed which produces a schedulable set of users. This set includes all users that have pending retransmissions, due to received NACK messages in the HARQ channel. It also includes all users that have in their queues an amount of data larger than a predetermined threshold or the delay of the head of line packet is greater than a

predetermined limit. The second constraint makes sure that the users to be scheduled have an adequate amount of data to transmit in their buffers.

In the Time Domain (TD), users from the schedulable set are divided into two sets. The first set contains those users with rate below their target bit rate (TBR) and the second set contains the rest. Since the QoS requirements, as far as the TBR is concerned, are not met for users of the first set, full priority is granted to them over the users of the second set. Prioritization within each set is performed using different metrics. Users of the first set are prioritized using Blind Equal Throughput metric while users of the second set are prioritized using the common PF metric. From these two sets a specific number N_{mux} of users with the highest priorities are selected as an input to the Frequency Domain (FD) scheduler.

The Frequency Domain scheduler performs scheduling on a SB-by-SB basis. Each SB is allocated to one of the previously selected users that maximizes a certain metric. The authors suggest three different metrics. The first is the common PF metric where the numerator represents the amount of data that the user can transmit using the specific SB, depending of course on the reported channel quality (Frequency Domain-Proportionally Fair, FD-PF). A second metric is similar to the PF metric with the difference that the average rate in the denominator is an estimation of the average rate that the user would achieve as if it was scheduled in every TTI, which is not always the case (Frequency Domain- Proportionally Fair scheduled, FD-PFsch). The last metric, that results in a scheduler called Frequency Domain- Carrier over Interference to Average (FD-CoItA) is the following: $m_{k,n}^{CoItA} = \frac{CQI_{k,n}}{\sum_{j}^{N_{SB}} CQI_{j,n}}$

It is essentially the ratio of the reported CQI of user n in SB k to the sum of the CQIs of user n in the entire bandwidth. This metric quantifies the fast fading gain of user n in SB k. Extensive simulations are performed to reveal the conditions under which TD scheduler has a positive effect on the system performance and the conditions under which each FD scheduler performs better.

4.3.3.7 Token Based Scheduler

Authors in [46] use a token based scheduling technique to cope with real time traffic. In particular, their proposed scheduling algorithm is comprised of three stages. The first stage is a flow classifier that distinguishes traffic flows into real time services and non-real time services. The second stage is a buffer manager that stores and controls the data rate. The last stage is a flow scheduler that has a separate functionality for real time traffic and non-real time traffic.

For real time, scheduling is performed using a token generation technique for each of the real time flows in order to control the traffic flow. In particular, a token bucket is maintained for each traffic flow and a respective token generator is initiated for every bucket with a rate that depends on the specific service. Different token generation rates are used to differentiate the different services. The token bucket depth is updated every TTI, by adding new generated tokens and removing an amount of tokens proportional to the transmitted in the previous TTI data. It is easy to conclude that a user that has not been allocated resources in the last few TTIs will show a high token depth value.

The scheduler performs an initial allocation of scheduling blocks by assigning each one of them to the user that has reported the highest instantaneous channel condition in the specific scheduling block. Then for every user, a check is performed whether the actual amount of data that can be transmitted using the allocated resources exceeds the token

bucket depth. If this is the case scheduling blocks are removed accordingly from the user.

Once real time flows scheduling is complete and there are scheduling blocks that have not been allocated, a non-real time scheduler is initiated in order to allocate them to non-real time users. The allocation is performed using a regular PF algorithm. Simulations prove that the proposed algorithm outperforms both M-LWDF and EXP/PF algorithms in terms of throughput, delay and packet loss rate of real time flows such as VoIP and video.

4.3.3.8 Other Scheduling and Resource Allocation Solutions

A large amount of solutions to the problem of scheduling and resource allocation in LTE has been proposed. Most of those solutions may be variations of the abovementioned techniques. An efficient algorithm should take into account a lot of different and sometimes contradicting parameters. It is clear that not all algorithms address all the different aspects of the problem. Most of them focus on certain parameters and provide solutions that are very effective from a certain point of view. For instance, a QoS-aware algorithm may provide solutions that conform to the delay constraints but fail to guarantee a target bit rate. Not all solutions take into account the packer error rate, which is another important QoS feature.

In addition to this, many solutions do not specify how their proposed solution conforms to certain LTE constraints. For example many algorithms allocate Scheduling Blocks to UEs without specifying the MCS that will be used for transmission. It should be reminded that defining MCS is also part of the scheduling decision in LTE. Another issue that is not usually addressed is the control overhead implied by specific resource allocation techniques. Furthermore, the complexity factor is usually not considered, although it is a parameter that may constitute an algorithm not feasible for the LTE, given that the period of scheduling decisions is very short.

Most of those issues are addressed in [29], where a mathematical expression of the resource allocation problem in LTE is formulated. In particular scheduling in LTE is expressed as a problem of maximizing the weighted sum of achievable rates of the users in each TTI under several constraints. The weights are used to ensure fairness and differentiate the various services according to the unique QoS characteristics. The constraints require that each scheduling block is allocated to at most one UE, each UE can transmit with a single MCS and MIMO transmission mode at each TTI, the number of scheduled UEs does not exceed a maximum value that is dictated by the control overhead limitations, the queues size does not exceed a certain limit and that each transmitted codeword size is also bounded. It is proved that this maximization problem is NP-hard and a sub-optimal solution is proposed.

In [47], control overhead limitations are also considered by iterating through all possible CFI values and selecting the minimum value that serves the appropriate number of UEs dictated by their proposed scheduling algorithm. Many proposed solutions adopt the two-step approach of paragraph 4.3.3.6 by separating the scheduling process in time domain (TD) and frequency domain (FD). In [48], the NGBR UEs are prioritized in TD using as a metric the ratio of a QoS priority index to the experienced average throughput of each user. In FD, the GBR UEs are scheduled first by assigning the channels with the best reported CQIs and then the NGBR UEs are scheduled based on the TD prioritization. In [49], TD prioritization is performed using a metric that takes into account the specific service class, the common PF metric and the time remaining for a service request HOL packet to violate the time delay constraint. The FD scheduling

allocates scheduling blocks exploiting the reported channel quality and determines the proper MCS and MIMO technique (transmit diversity or spatial multiplexing) that satisfies the respective Block Error Rate (BLER) requirement. In [50], a classifier splits the traffic in RT, NRT and Background Traffic. Each class is prioritized in TD using a different metric. Metric used for RT traffic is the product of normalized HOL packet delay (actual delay divided by maximum allowable delay) and wideband channel quality. NRT traffic metric is a product of normalized HOL packet delay and common PF metric. Finally Background traffic is prioritized using PF algorithm. FD scheduling is performed using PF algorithm on each scheduling block and respecting the TD priorities.

Finally a very interesting approach in adopted in [51] and [52]. The available flows are grouped into classes depending on the QoS characteristics. The available scheduling blocks are distributed to those classes using game theory that aims to achieve the fairest possible allocation that takes into account the different bit rate requirements of each class. Then EXP rule is used to allocate resources within the RT classes and PF to allocate resources within NRT classes.

4.4 Conclusions

Closing the chapter we found it useful to comment over the conclusions we reached and the lessons learned from the survey of the scheduling problem and the studied algorithms. It was pointed out that the scheduling problem takes into consideration many variables and constraints, regarding its performance and desirable features to be achieved, resulting thus in the solution of a NP-hard problem. However due to the lack of efficient polynomial complexity algorithms all the proposed algorithms make use of a per SB metric, which aims to meet a certain objective and indicates to which UE should be assigned a particular SB. Furthermore it was illustrated that the throughput maximization and other objectives such as fairness or QoS provisioning are competing interests which leads to the idea that a good scheduling scheme is one that allows a good trade-off among the competing interests. With the term good we refer to an algorithm that has low complexity and is realistically implementable.

Studying the proposed strategies led us to classify them in three different classes: channel unaware, channel aware QoS unaware and channel aware QoS aware. The algorithms that belong to the first category are of no particular interest since their implementation leads in situation of no fairness and throughput efficiency, making them incapable to meet efficiently any of the required objectives. On the contrary it was pointed out that channel awareness exploitation is fundamental in order to achieve high throughput in a wireless channel. In fact channel awareness which is achieved by using CQI information enables an operator to improve spectral efficiency as required in the standard, making channel aware algorithms capable in meeting the throughput requirements. Lastly in the channel aware QoS aware category are found algorithms that consider QoS requirements. In many proposals it was demonstrated that these algorithms are capable to deliver packets within the required time constrains.

The large number of proposals presented in this thesis and many others found in literature that we do not concern with, suggest that considerable efforts have been made in order to provide solutions for the scheduling problem. However as we had previously pointed out many of the proposed solutions despite the rigorous mathematical foundation, from different branches of mathematics such as linear programming and game theory, lack significant features and exhibit issues in order to be deployed in real systems. Furthermore, despite the fact that in many of the proposals

suboptimal schemes were suggested, to our surprise in most of the proposals no complexity evaluation has been made.

5. COFRTS ALGORITHM

5.1 Introduction

A concise literature review of LTE scheduling and resource allocation algorithms was presented in the previous chapter. The main conclusion drawn by this review is that even the most sophisticated proposed solutions fail to address all aspects of the scheduling problem. This chapter presents an attempt to solve efficiently the problem of scheduling and resource allocation in LTE, focusing on minimizing the complexity of the scheduler's decision process while at the same time providing a fair and QoS aware solution. The proposed algorithm is called Colored OFDMA Frame Registry Tree Scheduler – COFRTS and is analyzed in the following paragraphs.

5.2 Facts Concerning Scheduling, Offered Traffic and System Throughput

The illustration, of the most prevalent techniques used to solve the problem of packet scheduling in LTE-Advanced made in chapter 4, pointed out that the most realistic proposals solve the problem of scheduling by calculating certain metrics which determine the UEs that should be scheduled at a specific TTI and the SBs to be assigned. However despite the large number of proposals found in the literature none of them did exploit certain properties regarding scheduler's functionality, characteristics of the offered traffic and systems overall throughput. In the following, we present certain facts that we will consider in solving the scheduling problem:

- Offered traffic is not constant but varies over time, and the distribution of packets arrivals exhibits certain patterns.
- System's maximum throughput is not constant but time dependent.
- Packet transmission times are subjected to delay determined by QoS requirements.
- The scheduler decides at the beginning of every TTI, the rest of the time is idle regarding scheduling.

In order to further comprehend and utilize the abovementioned facts a more in depth analysis for each one is considered necessary.

***Time dependence of the offered traffic*:** There are various factors related with the packet arrival that need to be taken under consideration such as the peak traffic time, the users density, the users' profile regarding the services etc. All these parameters are time dependent so that packet arrival distribution varies over time.

The most common and widely accepted mathematical model used for simulating packet arrival times is the Poisson distribution. In Poisson distribution interarrival times show significant variation between low and high values, which sufficiently models packet arrivals at an eNB. A typical traffic variation in an eNB over a small period of time is depicted in figure 5.1.

The black line in the graph represents the instantaneous offered traffic load in bytes, while the orange, blue and magenta lines represent the maximum traffic that the system can carry at different time intervals.

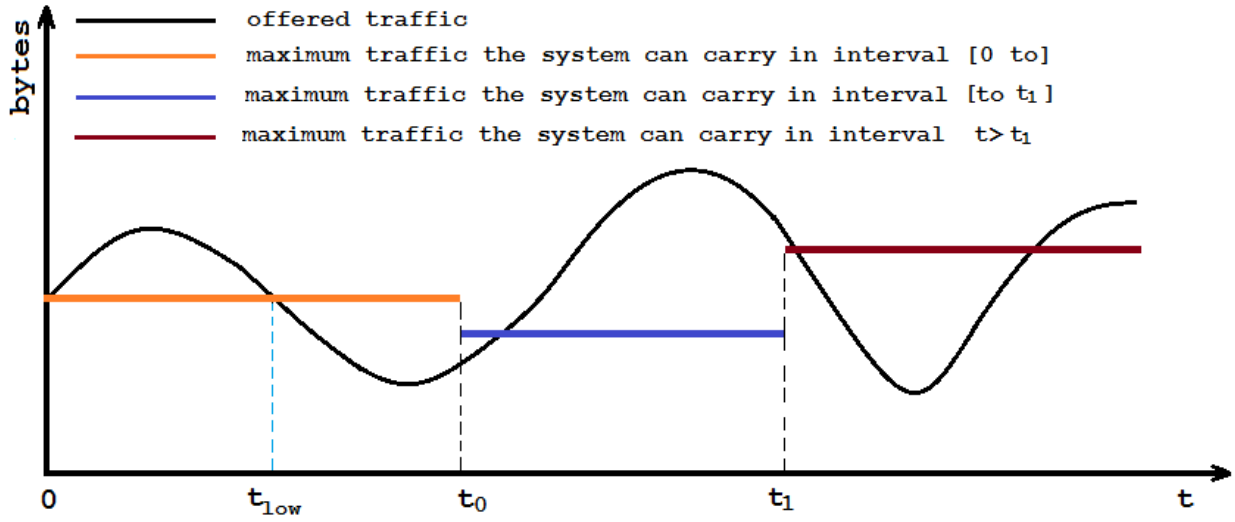

Figure 5.1: Snapshot of offered and maximum carried traffic distribution

Time dependence of system's throughput: The maximum traffic that the system can carry is defined as the system's maximum throughput, and it is achieved when optimal time scheduling and resource allocation is performed with regards to certain criteria. The maximum throughput changes over time due to the time-varying nature of the wireless channel. User's mobility and distinctive characteristics of the wireless channel, such as the multipath propagation, result in a time varying channel impulse response:

$$h(\tau,t)=\sum_{n=0}^{N(t)} a_n(t)\cdot e^{-j\varphi_n(t)}\cdot\delta\big(\tau-\tau_n(t)\big)$$

where $N(t)$ is the number of multipath components, $a_n(t)$ and $\varphi_n(t)$ are the gain and phase shift for each component respectively.

Hence adaptive scheduling should be employed in order to cope with these variations. However channel state can be assumed constant for short periods of time without inducing considerable error. As a result, the scheduling decisions may be taken periodically and considered valid for these time periods.

An interesting fact that is deduced from figure 5.1 is that there are time intervals during which the system cannot serve the offered traffic load while there are other time intervals where the system's serving capability is higher than the offered load. For example, scheduling packets in the interval $[0, t_{low}]$ would result in considerable losses.

Key conclusion regarding the previous two facts: if it is possible the scheduling of the packets to be distributed appropriately over time, then packet losses can be reduced.

Allowable delay of packet transmission: It should be reminded that the various traffic classes in LTE have different characteristics which can be utilized by the scheduler. As it was mentioned in previous chapters, LTE supports several traffic classes, both real time and non real time, which are distinguished by the following QoS features: priority, packet error rate, guaranteed bit rate and maximum allowable delay. The maximum allowable delay for the various QoS classes ranges from 50msec to 300msec. The existence of this variable acceptable delay can be exploited by the eNB scheduler to achieve the desirable traffic time scheduling, by arranging the transmission of packets to appropriately selected time moments.

Key conclusion: even distribution of the scheduled traffic over time can be achieved.

Scheduler's busy cycle: From the system's specifications the scheduling decisions are taken at the beginning of every TTI. As a result decisions regarding scheduling are taken in an extremely small time interval compared to the time interval that the scheduler is idle. Given the complexity pertaining scheduling, as it was illustrated in chapter 4, an efficient approach ideally should utilize this idle time. This can be achieved by organizing the information in a useful way i.e. preprocessing it, so as when actual decisions are taken the time complexity is reduced.

Since the information will be preprocessed, it needs to be stored temporarily. This generates the idea of storing it in data structures that allow the retrieval and process of information efficiently and with low complexity (much lower than the one that would be required without preprocessing).

5.3 The Proposed Scheduling Algorithm

The proposed scheduling algorithm aims at solving the scheduling problem by utilizing all the aforementioned facts, as well as considering other QoS characteristics and constraints as those described in chapter 4. The concept of the algorithm was firstly proposed in [53] and [54]. The main differences of the implementation presented below include the adaptation of the idea in LTE networks and the introduction of a new Frequency Domain Scheduling stage that takes into account the constraints that the LTE standard imposes. We should also highlight the fact that in our work we distinct the time domain scheduler in two parts, the PREPROCESSING and TIME DOMAIN steps.

The main innovative concept of the scheme proposed in this thesis is the preprocessing of information which utilizes scheduler's significant portions of idle time. The steps employed by the proposed algorithm are depicted in figure 5.2. The preprocessing step involves the processing of information received by packets arriving at the UEs queues and storing it in appropriate structures. Namely, metadata for each packet are being captured from the queues and then useful information is extracted and stored in a tree structure. This tree structure is described in paragraph 5.3.1 whereas all the policies regarding the construction of the tree as well as all the operations regarding it are described in paragraph 5.3.2.

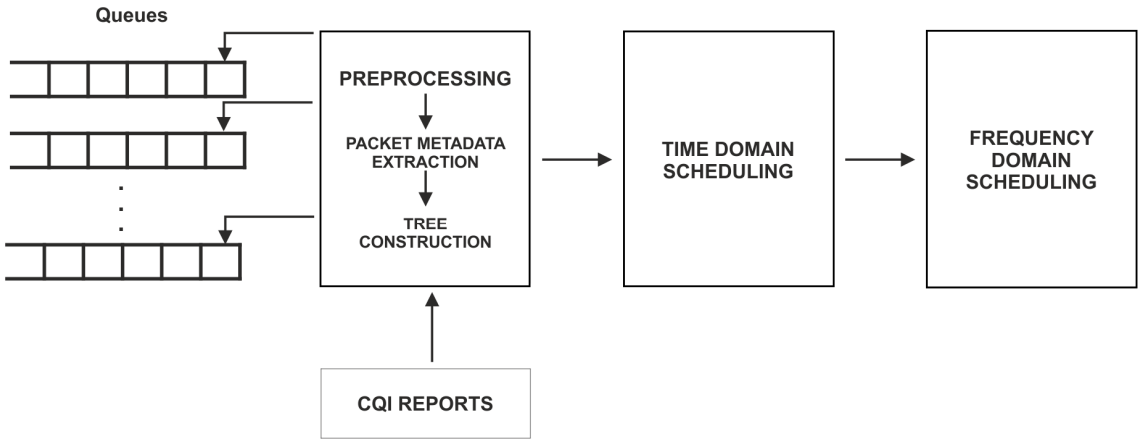

Figure 5.2: Steps of the proposed scheduling algorithm.

The time domain scheduling involves steps where information is retrieved from the tree structure and qualification of the UEs that are capable to receive at each TTI is performed. Basically for each UE, the scheduler chooses the number of packets to be scheduled and the corresponding bearer. The frequency domain scheduler also makes

use of data structures from the preprocessing step, in order to reduce required calculations done in resource allocation i.e. SB assignment. Both TD and FD scheduling algorithms take into consideration the preceded preprocessing. It is clarified that the preprocessing step includes decisions that in a traditional scheduling scheme are made in the TD and FD scheduling therefore it constitutes part of the overall scheduling.

5.3.1 Colored OFDMA Frame Registry Tree Structure

The preprocessing step employs a tree structure with multiple levels, complementary to the connections queues. The tree structure enables simultaneous processing and efficient storage of useful information from the arriving packets in the queues. The processing and the efficient storage contribute in making scheduling decisions with low complexity. It should be clarified that only packet metadata are stored in this structure and not the packets themselves.

The proposed tree is called Colored OFDMA Frame Registry Tree (CO-FRTree) and has the structure depicted in figure 5.3. Basic operations and the adopted policies related to the construction of this structure are described in paragraph 5.3.2.

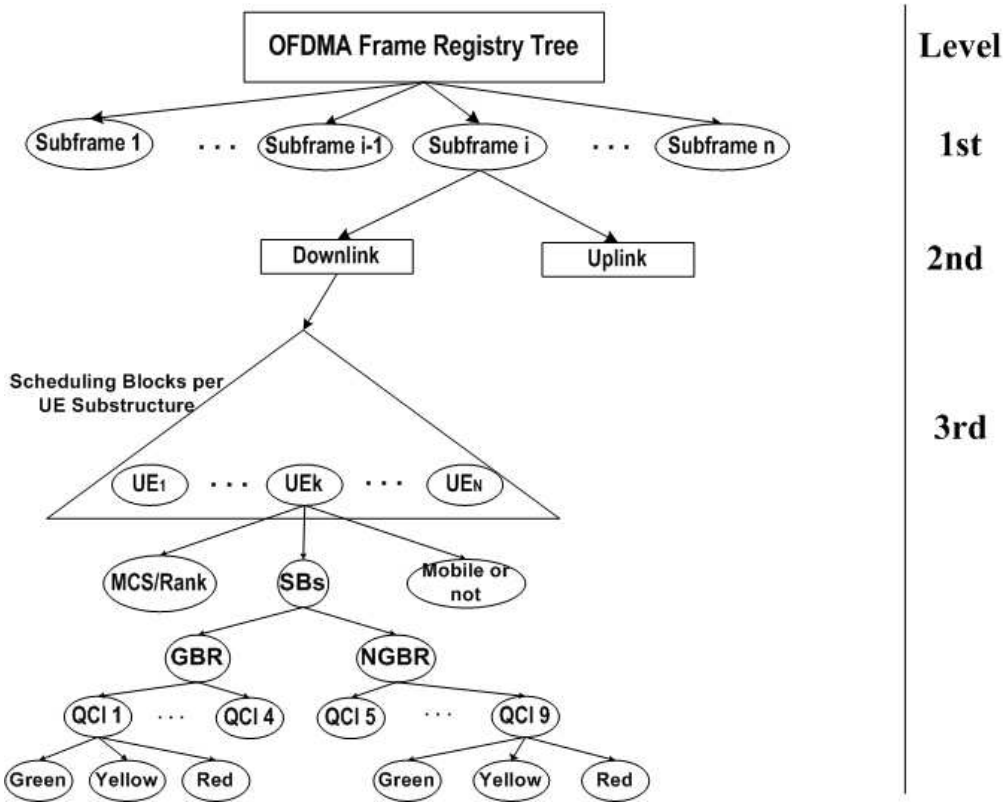

Figure 5.3: Colored Frame Registry Tree Structure

The proposed tree structure consists of nodes belonging to three distinct levels:

First Level: The nodes of this level represent consecutive subframes that a particular packet is designated to be scheduled. Subframe 1 represents the current subframe to be scheduled. Each node of this level points to two different nodes of the next level that can be considered as the roots of two subtrees where additional information is stored.

Second Level: This level contains two nodes, under every node of the first level, each one indicating whether stored information is designated for the uplink or the downlink direction.

Third Level: In this level lies the basic version of the COFRTS structure. It is a subtree which contains the majority of the stored information, under the downlink or the uplink nodes.

We refer to this subtree as "*Scheduling Blocks per User Equipment*".

The information stored in this subtree, includes but is not limited to: the size of the packets that are to be scheduled in the specific subframe, UE identification, the highest Modulation and Coding Scheme with which it can transmit or receive reliably, the Rank Indicator value, the size of the packets that are scheduled in the specific subframe and the number of Scheduling Blocks required for transmission. The Scheduling Blocks are classified according to the type and the color of the packets providing the scheduler with a convenient means of making decisions that take into account the prioritization defined by coloring and QCI.

The efficiency of this concept lies in the significant complexity improvement that the "*Scheduling Blocks per UE*" structure offers. This structure is implemented as an Interpolation Search tree [55] or a Van Emde Boas Tree [56] and contains all UEs along with their related transmission parameters. The depth of this tree structure is $O(\log\log N)$, the required space is $O(N)$ and the related cost for a search or update procedure is $O(\log\log N)$ with high probability, where N is the number of UEs**.**

5.3.2 Adopted Policies and Operations on COFRTS

In the previous paragraph we presented the tree structure which is the backbone of the preprocessing step. We did not provide any justification regarding specific choices of its structure nor did we determine the operations and the followed policies regarding its functionality. In this paragraph we make a full description of all these.

5.3.2.1 Adopted Preprocessing Policies and Features of COFRTS

Subframe Scheduling Policy: The first level of the tree has dual utility: a) determining the subframe in which a packet is to be scheduled b) providing quick access on packets scheduled in different subframes.

The initial selected scheduling policy of the newly arrived packets is to transmit them in the subframe immediately preceding their deadline time, which is calculated as:

$$t_{deadline}\left(P_{ij}(k)\right) = t_{arrival}\left(P_{ij}(k)\right) + t_{latency}\left(P_{ij}(k)\right) \qquad \mathbf{5.1}$$

where $P_{ij}(k)$ is the k_{th} packet of radio bearer B_j of UE U_i.

$$t_{latency}\left(P_{ij}(k)\right) = t_{latency}(B_j) - t_{delay\ in\ network}\left(P_{ij}(k)\right)$$

It should be reminded that in LTE all different QoS classes have a maximum latency value defined by the standard.

The selected policy along with the capabilities of the tree structure i.e. quick access enables, if possible given the system's maximum throughput pattern, the accomplishment of even traffic distribution that conforms to the key observation made in paragraph 5.2. This policy is deployed in the TD scheduling part.

Having determined the policy concerning the subframe that each arriving packet is initially scheduled, it is straightforward to determine the number of subframe nodes of the tree. The number of subframes at any moment is bounded by the maximum allowable delay of a packet in LTE divided by the subframe's duration, hence it is finite. Scheduling a packet in a subframe that exceeds this limit is pointless since it violates the maximum delay limit.

Traffic Policy and Coloring: LTE specifications require that the average bit rate of a bearer belonging to a GBR class is within certain limits. This led us to impose a certain policy which guarantees that the received rates conform to this requirement.

The proposed policy is based on ideas analyzed in [57], [58]. Specifically packets are classified into three categories, also referred as "colors", depending on whether the average throughput that results from serving a specific packet is compliant with the predetermined bit rates or not. The basic idea behind this technique is that packets that would result in excessive throughput are marked as "red", packets that would result in a compliant throughput value are marked as "yellow" and those that would result in a throughput value below the agreed limit are marked as "green".

In particular, let $P_{ij}(k)$ denote the k_{th} packet of the traffic flow of UE U_i and radio bearer B_j that arrives at eNB in between subframe $n-1$ and n (n counts from the initiation of B_j for U_i), and let $T_{ij}(n)$ denote the long term average throughput of B_j until subframe n. The eNB scheduler calculates what the resulting average throughput $T_{ij}(n)$ will be if packet $P_{ij}(k)$ is served in subframe n using the following formula:

$$T_{ij}(n) = (1-\tfrac{1}{n}) \cdot T_{ij}(n-1) + \tfrac{1}{n} \cdot R_{ij}(k) \text{ where } R_{ij}(k) = \frac{Size(P_{ij}(k))}{TTI}.$$

The painting policy for GBR services compares the value $T_{ij}(n)$ with the agreed guaranteed bit rate $R_{j,min}$ and with the maximum allowable bit rate $R_{j,max}$ of B_j.

if	$T_{ij}(n) < R_{j,min}$	$Select\ packet\ color\ Green$
else if	$R_{j,min} \leq T_{ij}(n) < R_{j,max}$	$Select\ packet\ color\ Yellow$
else if	$R_{j,max} \leq T_{ij}(n)$	$Select\ packet\ color\ Red$

In our approach the same coloring policy is also applied to non-GBR services despite the fact that no predetermined guaranteed and maximum bit rate values are defined for these bearers. However, defining maximum and minimum bit rate limits and applying such a coloring policy to non-GBR bearers as well provides us with an additional tool of ensuring some kind of fairness among users.

The coloring technique provides the scheduler with a means of prioritizing packets with respect to the bit rate QoS characteristics. Therefore green packets of a certain class are served with higher priority than yellow packets and yellow packets with higher priority than red packets of the same class. So, in addition to priorities between classes of traffic, specified by the respective QCI, the scheduler can also utilize the prioritization of packets within a class as indicated by the color, before taking a scheduling decision.

5.3.2.2 Operations on COFRTS

In this subsection we describe the operations and the functionalities of the COFRTS.

Channel Update: An update event on the tree structure can be triggered by a channel state report update (reported CQI value from a UE) received from the eNB in the uplink

control channel. Different channel states result in different maximum Modulation and Coding Scheme – MCS selections for reliable transmission.

The update steps are as follows: for each subframe of COFRTS that contains information for a specific UE all its transmission parameters are being updated by changing the values of the corresponding nodes of the tree.

Packet Insertion: This operation is triggered when a packet arrives at eNB's queues. Then the tree structure needs to be updated by processing and inserting the metadata of the actual packet.

Let a packet $P_{ij}(k)$ arrives at the U_i queue at time $t_{arrival}$, then the intervening time from $t_{arrival}$ to the beginning of the next subframe is defined as t_{gap}. Then the $t_{deadline}$ for this packet is calculated by equation 5.1. Also let MCS_i, RI_i be the MCS and the rank value for U_i respectively.

Packet insertion Algorithm in COFRTS

Input: Metadata of Packet $P_{ij}(k)$, COFRTS tree.

1. Calculate $Subframe\ ID = \left\lfloor \frac{t_{deadline} - t_{gap}}{TTI} \right\rfloor$.
2. Coloring the Packet.
3. Read $MCS_i, RI_i\ from$ "Scheduling Blocks per UE" subtree for UE U_i.
4. Calculate $S_{ij} = S_B(MCS_i, RI_i, size(P_{ij}))$.
5. **if** $Subframe\ ID$ does not exist create it and proceed **else** proceed.
6. Create or Select Proper Direction.
7. Create or Select entry of UE U_i on "*Scheduling Blocks per User Equipment*".
8. Insertion in "*Scheduling Blocks per User Equipment*".
   a. Add $size(P_{ij})$ to total packets size and S_{ij} to total required SBs.
   b. Update node whether it belongs to GBR or non-GBR bearer.
   c. Create or Select and update the node of the appropriate QCI.
   d. Create or Select and update the node of the appropriate color.

S_B: Is a function that calculates the required SBs given the MCS, the packet size and the rank value.

The search of UE U_i and the update of its underlying nodes is executed in $\Theta(\log\log N)$ expected time with high probability and at most $O(\log\log N)$, where N is the number of UEs.

5.3.3 Time Domain Scheduling

The functionalities of time domain scheduling, in our proposal, are closely related to and exploit the organization of the information that takes place in the preprocessing step. So the task of the TD scheduler, by utilizing the COFRTS structure, is to define the set of UEs that will be assigned resources in the current TTI along with their related packets.

5.3.3.1 Determining the Amount of Traffic to be Scheduled

The TD scheduling process deals with a fundamental problem: whether the traffic in the current subframe of COFRTS tree that was prescheduled during the preprocessing step is sufficient in order to approach as much as possible the maximum throughput the system can carry. Otherwise traffic from subsequent subframes should be employed so that waste of resources is avoided.

The solution of the problem above is given by defining a variable which serves as a threshold that is compared to the amount of traffic being present in the current subframe. The selection of the threshold's value is of great importance since it affects the overall performance of the algorithm. For this the following variables and objectives must be taken into consideration:

1. The achieved throughput in the previous TTI.
2. Ensure that sufficient amount of traffic will be qualified for scheduling.
3. Ensure that no excessive amount of traffic is selected from subsequent COFRTS subframes, a fact that affects the prioritizing of the packets.

The above targets are all concentrated in the following formula: let $Thr(c)$ be the threshold value for the current subframe, $Th(p)$ be the achieved throughput in the previous TTI and $Tr(c)$ be the estimated maximum throughout for the current subframe, then:

$$Thr(c) = a \cdot Th(p) + (1 - a) \cdot Tr(c)\,, \qquad a \in (0,1) \qquad \mathbf{5.2}$$

5.3.3.2 TD Scheduling Algorithm

The TD scheduling algorithm starts comparing the amount of traffic found in the current subframe to the threshold. If excessive amount of traffic is found, appropriate kind and number of packets is removed. Otherwise if sufficient amount of traffic does not exist in the current subframe, appropriate kind and amount of traffic is transferred from subsequent subframes.

In the following algorithm SF_i denotes a subframe and $size(SF_i)$ represents the amount of traffic in that subframe. The current subframe considered is SF_1, without loss of generality.

Algorithm Input: COFRTS

1. $Calculate\ Thr(c)$

2. $\boldsymbol{if}\ size(SF_1) < Thr(c)$

3. $i = 2$

4. $\boldsymbol{while}\ size(SF_1) < Thr(c)$

   a. $Select\ a\ packet\ of\ the\ highest\ priority\ according\ to\ 5.3\ from\ SF_i$

   b. $Transfer\ it\ to\ SF_1$

   c. $\boldsymbol{if}\ size(SF_i) == 0\ \boldsymbol{then}\ i = i + 1$

5. $\boldsymbol{end\ while}$

6. ***else***
7. ***while*** $size(SF_1) \geq Thr(c)$
   a. $Select\ a\ packet\ of\ the\ lowest\ priority\ according\ to\ 5.3\ from\ SF_1$
   b. $Check\ the\ packet\ for\ delay\ expiration$
   c. $Drop\ the\ packet\ if\ it\ expires\ or\ transfer\ it\ to\ SF_2\ if\ not$
8. ***end while***
9. ***end if***

$$P_{green,QCI_1} > \cdots > P_{green,QCI_9} > P_{yellow,QCI_1} > \cdots > P_{yellow,QCI_9} > P_{red,QCI_1} > \cdots > P_{red,QCI_9} \quad \mathbf{5.3}$$

After prioritizing the UEs and the packets, the resource allocation algorithm takes over, aiming to serve the packets that are present in the current subframe in the most efficient way. In the next paragraph two resource allocation techniques are described. The proposed algorithms make use of an extended version of the COFRTS in order to reduce the complexity.

The COFRTS tree structure can be used in a similar manner for both uplink and downlink, but we focus on the downlink case only. In fact, the algorithm steps of time domain scheduling, described in this paragraph, can also be applied in the uplink case. However frequency domain scheduling described in the next paragraph is only valid for downlink, given that LTE imposes some additional constraints for the uplink that this proposal does not take into consideration.

5.3.4 Frequency Domain Scheduling

The frequency domain scheduling algorithm assigns and distributes resources i.e. SBs efficiently among the UEs qualified by the TD scheduler. Efficiency is referred in the sense that packets selected for each UE from the TD scheduler will be served maximizing the throughput and ensuring fairness among users. Common resource allocation techniques such as Proportionally Fair and Maximum Throughput can be used for this purpose.

In the following paragraphs, we propose two algorithms that focus on minimizing the complexity by using some additional sophisticated tree substructures. Those tree structures are added to the already presented COFRTS tree structure enhancing its functionality so that efficient resource allocation can be run with low complexity.

5.3.4.1 Extended COFRTS for Supporting FD Scheduling

Operations that take place in the FD algorithms that are described in the following paragraphs require further information. This information should be acquired from the preprocessing step. This fact incites us to extend the concept of the COFRTS structure by adding two additional substructures. These two subtrees the "*SBs Interval Tree*" and the "*Bytes per User Equipment*" tree are inserted in the third level of the COFRTS structure, in parallel with the "*Scheduling Blocks per User Equipment*" subtree. The final version of the COFRTS structure with these additional substructures is depicted in figure 5.4. The three substructures are appropriately interlinked, so that queries concerning combined information from all these subtrees can be run recurrently with low computational cost.

5.3.4.1.1 SBs Interval Tree

This subtree is a structure in which information regarding the SBs, where a UE is able to receive reliably using a specific MCS, is efficiently stored and retrieved. These particular SBs will be referred to as active SBs. The main properties that this tree utilizes are the facts that: a) the active SBs overlap among different UEs b) active SBs tend to be consecutive i.e. in continues intervals for a specific UE.

An active interval $(a, b]$ is essentially a set of active and continuous SBs which contains the SBs indexed as $[a + 1 < \ldots . < b]$.

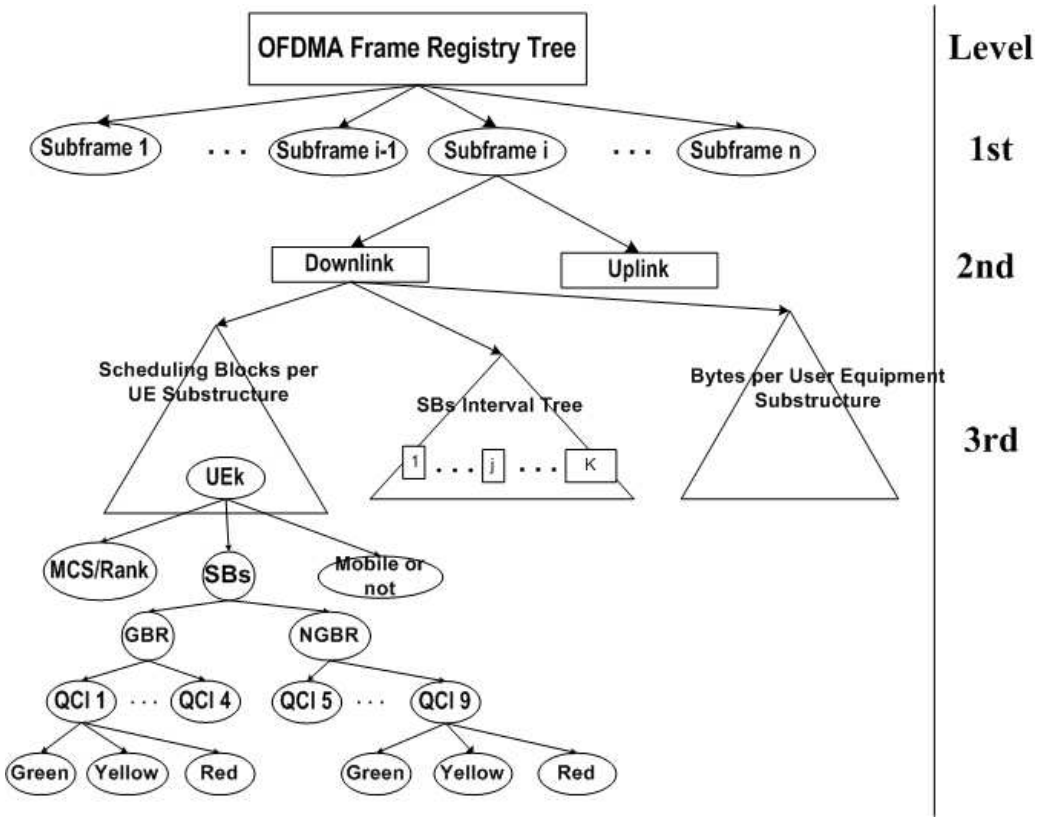

Figure 5.4: Full version of the COFRTS structure

The interval tree provides an efficient way of accessing the active intervals of a specific UE or the UEs that are able to receive in certain active SBs. For this, all the active intervals of different UEs are expressed as union of special intervals called *elementary intervals* which have the following properties:

- They are of the form $(a, b]$.
- Let $(c_i,\ c_j]$, $(c_m,\ c_n]$ be elementary intervals, then $(c_i,\ c_j] \cap (c_m,\ c_n] = \emptyset$.
- Let $(c_i,\ c_j]$ be an elementary interval and $(c, d]$ an active interval then $(c_i,\ c_j]$,$\cap$ $(c,\ d] = (c_i,\ c_j]\ \ or\ \emptyset$.
- The active intervals of each UE can be expressed as union of elementary intervals.

The concept of the elementary intervals contributes to the following facts:

a. Let K be the total number of SBs and K_E the number of elementary intervals. Due to the fact that each elementary interval contains one or more SBs then $K_E \leq K$ is always true. It is obvious that there will be a reduction in the complexity of the resource allocation algorithms since the search space contains K_E intervals rather than K SBs.

b. It can be organized in a tree structure providing efficient access.

The interval tree is a binary tree in which each of the internal nodes represents an active interval of SBs and points to two child nodes, whereas the leaves of the tree represent the elementary intervals.

The key of every internal tree node is an interval edge c_{med} within the examined interval $(c_i,\ c_j]$ so that the interval $(c_i,\ c_{med}]$ contains the largest possible number of elementary intervals that is power of 2. The root of the tree is a node that represents the entire bandwidth $(c_0,\ c_K]$, where $c_0 = 0$ and K is the total number of SBs.

An internal node of the interval tree is said to span the union of intervals in its subtree, meaning that the interval represented by the node can be derived as the union of the intervals represented by all the nodes of its subtree. The ID of a UE that has an active interval of SBs $(a, b]$, is stored in a node x if it has the following properties:

- $\mathrm{span}(\mathrm{x})$ is completely contained in $(a, b]$.
- $\mathrm{span}(\mathrm{parent}(\mathrm{x}))$ is not completely contained in $(a, b]$.

UE ID	UE 01	UE 03	UE 04	UE 05	UE 06	UE 07	UE 09	UE 11	UE 12	UE 14	UE 15	UE 17	UE 18
Active Intervals of SBs	(0,3], (13,15]	[1,9]	(3,5], (9,10]	(0,1], (9,10], (12,13], (14,15]	(13,14]	(5,12]	(22,25]	(12,17], (19,22]	(0,5], (15,17]	[13,21]	(21,22], (23,25]	(17,10], (22,23]	(12,13], (17,21]

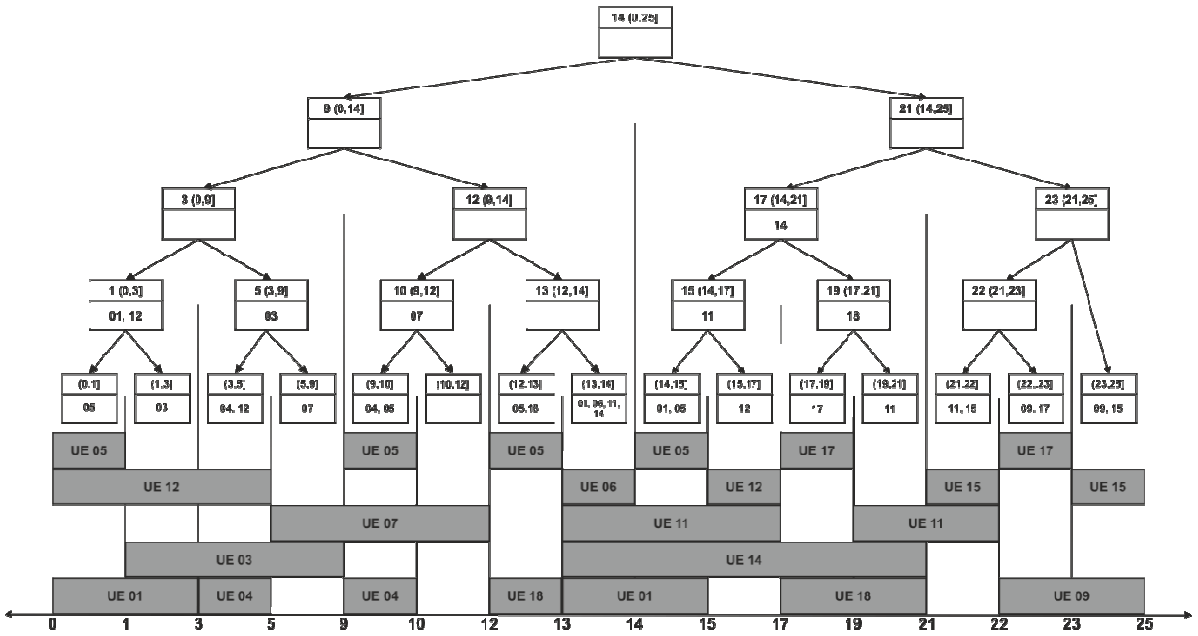

Figure 5.5: Example of an Interval Tree with 25 SBs (5MHz).

The construction of the interval tree requires the active intervals for each UE. This means that MCS has to be chosen prior to constructing the interval tree. In our

approach we chose to match the concept of reliable transmission for every UE with the transmission using the best reported MCS. This is a choice that aims to exploit the multiuser diversity in the sense that different UEs may show favorable channel conditions in different parts of the spectrum. However it is a choice that limits the options for a single UE.

In figure 5.5 is depicted an example of the interval tree, which is constructed by the active intervals of the UEs using the best MCS each UE can achieve. As it is depicted in the figure the number of elementary interval is 15 whereas the number of SBs is 25.

The construction process, the insertion as well as further details regarding the interval tree are presented in Appendix 1.

The complexity of the interval tree construction procedure is $O(K_E \cdot \log log K_E)$, the depth of the tree is $O(\log K_E)$, the storage required is $O(K_E \cdot \log K_E)$. Insertion or deletion of a node has $O(\log K_E)$ complexity.

Information regarding which UEs have favorable channel conditions in a certain interval costs $O(\log K_E)$ time, whereas information in which elementary interval a UE can transmit reliably requires also $O(\log K_E)$ time.

5.3.4.1.2 Bytes per User Equipment

The "*Bytes per UE*" subtree is essentially a priority queue that contains all UEs sorted according to the size in bytes of the packets stored in the current subframe. It is implemented as an Interpolation Search Tree which means that the access to the stored information costs $\Theta(\log \log N)$ time with high probability and at most $O(\log \log N)$ time, where N is the number of UEs. Insertions and deletions can be executed in $\Theta(\log \log N)$ average time. The required storage space is $O(N)$.

Hence, this structure offers a means of accessing in minimum time information regarding the UE that has the highest or the lowest data size requirements in the current subframe.

5.3.4.2 Resource Allocation Algorithm 1 – RA1

This is a very simple resource allocation algorithm which utilizes the tree structures introduced in the previous paragraph.

The main steps of the algorithm can be summarized in the following: for each elementary interval, the UEs that can reliably receive in it, are obtained from the interval tree. The specific interval is allocated to the UE, from the obtained group of UEs, which has the greatest data size requirements as indicated by the "Bytes per UE" subtree. Then the amount of data that the selected UE can receive in the elementary interval is calculated based on the MCS and the number of SBs contained in the elementary interval. This value is subtracted from the initial data requirements and the "Bytes per UE" subtree is updated.

The simplicity and the low complexity of the RA1 algorithm lie on the following properties:

- The algorithm scans the spectrum by examining each elementary interval separately, exploiting the already created in the preprocessing step interval tree substructure. This is beneficial from a time consumption perspective compared to

a solution that would examine each SB individually, given that the number of elementary intervals can be considerably lower than the number of SBs.

- The algorithm introduces no additional complexity by making choices or calculations that take into account the individual QoS characteristics of the packets. The overall solution is QoS-aware only due to the choices performed in the TD step.

A more formal description of the algorithm is given in the pseudocode presented below. In order to comprehend the algorithm an illustration of the notation used in the pseudocode is considered necessary.

With $\mathbf{T_I}$, $\mathbf{T_B}$ we denoted the *Interval Tree* and the *Bytes per UE* respectively. Also let $\mathbf{I_E} = (\mathbf{c_0}, \mathbf{c_1}] \cup \ldots\ldots \cup (\mathbf{c_{K-1}}, \mathbf{c_K}]$ be the set of elementary intervals of $\mathbf{T_I}$ and S_i be the set of SBs assigned to UE U_{i} at each TTI.

We define $B_S: (SBs, MCS) \rightarrow Bytes$ the function that calculates the number of bytes given a set of SBs and the MCS. The $\mathrm{j_{th}}$ MCS of the UE U_i is denoted as $\mathrm{m_{U_i},_j}$ with $\mathrm{j} = 1$ representing the best MCS with which U_i can receive reliably.

Furthermore two functions which perform operations over the presented data structure are defined:

$\boldsymbol{get}_P(\cdot)$: is a function which obtains from a data structure members that have a certain property defined by proposition P.

$\boldsymbol{update}_P(\cdot)$: function which updates the members of a data structure determined by P.

Input: $\mathbf{T_I}$, $\mathbf{T_B}$

1. $\mathbf{I_E} = \boldsymbol{get}_{Elementary\ Intervals}(\mathbf{T_I})$
2. **for** $\boldsymbol{c} \in \mathbf{I_E}$
3. $\mathbf{U_c} = \boldsymbol{get}_{UEs\ send\ reliably\ in\ \boldsymbol{c}}(\mathbf{T_I})$
4. $(U_k, b_k) = \boldsymbol{get}_{(U_k,b_k) \in \boldsymbol{U_c}:\ b_k = max(b_i)}(\mathbf{T_B})$
5. $S_k = S_k \cup \mathbf{c}$
6. $b_{served} = B_S(\boldsymbol{c}, \mathrm{m}_{U,1})$
7. $b_k = b_k - b_{served}$
8. $\boldsymbol{update}_{(U_k,b_k)}(\mathbf{T_B})$
9. **end**

The choice of assigning the resources to the UE with the greatest data requirements provides a mechanism of ensuring some kind of fairness. The update of the "Bytes per UE" subtree ensures that UEs already assigned elementary intervals in previous iterations of the current subframe are less likely to be given more intervals. The same mechanism also prevents the waste of resources, since spectrum is allocated to UEs that most likely have high data requirements.

The exact assessment of RA1 complexity is not straightforward as it depends on random measures such as the distribution of the data and UEs in the subframe. An estimation of the mean case may only be achieved, if further hypothesis, regarding the distribution of these measures, is assumed. Instead, the worst case scenario is not affected by this randomness and its complexity is: $O(K_E \cdot \log K_E + K_E \cdot N)$. The term $K_E log K_E$ corresponds to the complexity involved in acquiring the UEs from the interval tree, and the term $K_E N$ to the degenerated case of scanning the whole "Bytes per UE" subtree in order to select the UE with the most data for a specific elementary interval.

5.3.4.3 Resource Allocation Algorithm 2 – RA2

Despite the low complexity, RA1 may result in poor performance under certain conditions. For example, the constraint for using only the best MCS may be beneficial from a complexity point of view, but it may limit the available spectrum resources for a UE with high data requirements. In addition to this, parts of the spectrum where none of the UEs reports its best MCS may be wasted, which is very likely when the desirable multiuser diversity is not achieved. Also it is beneficial from a QoS perspective that the scheduler shows some preference in assigning to a UE with QoS demanding services parts of the spectrum where other UEs do not have packets requiring the same QoS treatment. As previously stated the RA1 algorithm does not apply such a policy.

The RA2 algorithm presented in this paragraph attempts to address the aforementioned issues with the cost of introducing some additional complexity.

In RA2, the assignment of the spectrum is performed on a UE-by-UE basis starting from the UE with the highest data size requirements as a result of the TD scheduler. For every UE the amount of required SBs is compared to the number of SBs in its set of active intervals. If the number of required SBs is less than its active SBs, a selection process is executed to determine which of the available SBs will be assigned. At this point, a QoS-aware decision is made. This is achieved by designating a metric, which will be analyzed in detail below, for every active interval. The metric determines the intervals assignment sequence starting from the one that exhibits the highest value. If the number of required SBs is larger than the number of SBs contained in the active intervals, a second MCS is examined. In particular, the next highest MCS for which the active intervals contain more SBs is determined. It is noted that the number of the active SBs is a decreasing function of the MCS. This is valid because if a UE reliably receives in a SB with a specific MCS, it certainly receives reliably with a lower MCS. However the rate per SB decreases. The same check is performed again to estimate if the active intervals that correspond to the new MCS are adequate. If this is the case then the same QoS-aware decision is made using the particular metric. In the case that none of the two examined MCS values result in a bandwidth value that satisfies the data requirements of the UE, the one that supports higher data rate is selected and all the SBs are assigned to the particular UE.

Let $M_{(a,b]}$ be the aforementioned metric where $(a, b]$ is an interval. Then, $M_{(a,b]}$ is defined as the ratio of the weighted SBs that a UE can receive in this interval to the weighted SBs that all the non-served UEs can receive in the same interval. Formally:

$$M_{(a,b]} = \frac{W_{(a,b],U}}{\sum_{u \in \boldsymbol{U}} W_{(a,b],u}} \qquad \mathbf{5.4}$$

$W_{(a,b],U}$ is the weighted SBs of interval $(a, b]$, $\boldsymbol{U}_{(\boldsymbol{a},\boldsymbol{b}]}$ is the set of UEs that can reliably receive in $(a, b]$ and have not been already served in the current TTI.

The weighted SBs of UE U on $(a, b]$ are calculated as $W_{(a,b],U} = \frac{\sum_i w_{QCI,c} \cdot size(P_{i,j}^{QCI,c})}{C_{(a,b],U}}$

where $P_{i,j}^{QCI,c}$ is the i_{th} packet of the UE U and bearer B_j in the current TTI.

With $w_{QCI,c}$ are denoted the weights applied on the packets size. They are a function of the color and QCI priority i.e. $QCI \in \{1, \ldots .9\}$ and $c \in \{green, red, yellow\}$. Their values are selected according to the priorities indicated by the expression 5.3.

$C_{(a,b],U}$ is the capacity of a SB for the particular UE and for the selected MCS.

The absolute value of the metric is of no importance. However, comparing the metric values of the various examined intervals assists in determining those where the least amount of high priority packets are likely to be scheduled.

The exact steps of the algorithm are shown in the pseudocode below. The notation followed is the same as in the case of RA1, so only new symbols are explained thereafter.

With N we denote the total number of UEs, $\mathbf{MCSI}_{\mathrm{U_i}}$ is a data structure which contains all the active intervals for each MCS that are supported for UE $\mathrm{U_i}$. Let $\boldsymbol{\mu(I)} := \sum_{\mathbf{i=1}}^{\mathbf{L}} (\mathbf{a_i} - \mathbf{a_{i-1}})$ be the cardinality of the set $\mathbf{I} = (\mathbf{a_0}, \mathbf{a_1}] \cup \ldots \ldots \cup (\mathbf{a_{L-1}}, \mathbf{a_L}]$.

Also function S_B is defined as $S_B: (Bytes, MCS) \rightarrow Number\ of\ SBs$ which calculates the number of required SBs given the number of bytes and a MCS.

Lastly with $\mathbf{T_B}$ we denote the "Bytes per UE" structure.

Input: $\mathbf{T_I}$, $\mathbf{T_B}$, $\mathbf{MCSI}_{\mathrm{U_i}}, i = 1, \ldots, N$.

while $\mathbf{T_B} \neq \emptyset$ and $\mathbf{T_I} \neq \emptyset$

1. $(\mathrm{U}, \mathrm{b}) = \boldsymbol{get}_{(\mathrm{U,b}):\ \mathrm{b} \geq \mathrm{b_i}, \mathrm{i}=1 \ldots \mathrm{N}} (\mathbf{T_B})$
2. $(\mathrm{I_{U,1}}, \mathrm{m_{U,1}}) = \boldsymbol{get}_{(\mathrm{I_{U,1}}, \mathrm{m_{U,1}}):\ \mathrm{m_{U,1}} > \mathrm{m_{U,k}}} (\mathbf{MCSI}_{\mathrm{U}})$
3. **if** $S_B(\mathrm{b}, \mathrm{m_{U,1}}) \leq \mu(\mathrm{I_{U,1}})$
    a. Construct metric from $\mathbf{T_I}$ for each interval of $\mathrm{I_{U,1}}$ according to 5.4.
    b. Assign required SBs sequentially starting from metric's highest value.
4. **else**
    a. $(\mathrm{I_{U,j}}, \mathrm{m_{U,j}}) = \boldsymbol{get}_{(\mathrm{I_{U,j}}, \mathrm{m_{U,j}}):\ \mu(\mathrm{I_{U,j}}) > \mu(\mathrm{I_{U,1}}),\ \mathrm{m_{U,1}} > \mathrm{m_{U,j}} \geq \mathrm{m_{U,k}}} (\mathrm{MCSI_U})$
    b. **if** $S_B(\mathrm{b}, \mathrm{m_{U,j}}) \leq \mu(\mathrm{I_{U,j}})$
    c. $\boldsymbol{reconstruct}_{\text{using } (\mathrm{I_{U,j}}, \mathrm{m_{U,j}}) \text{ as active intervals and MCS}} (\mathbf{T_I})$
    d. Construct metric from $\mathbf{T_I}$ for each interval of $\mathrm{I_{U,j}}$ according to 5.4.
    e. Assign required SBs sequentially starting from metric's highest value.
    f. **else**

g. $Assign\ all\ SBs\ of I\colon I = \arg_{I \in \mathrm{I_{U,1}}, \mathrm{I_{U,j}}} \max\left(B_S(\mathrm{I_{U,j}}, \mathrm{m_{U,j}}), B_S(\mathrm{I_{U,1}}, \mathrm{m_{U,1}})\right).$

5. $\boldsymbol{update}_{\text{ i=1,…,N remove intervals used in step 3 or 4}}\left(\mathbf{MCSI}_{\mathrm{U_i}}\right)$

6. $\boldsymbol{update}_{\text{remove (U,b)}}\left(\mathbf{T_B}\right)$

7. $\boldsymbol{reconstruct}_{\text{remove intervals used in step 4 or 5}}\left(\mathbf{T_I}\right)$

end

The complexity of the algorithm as in the case of RA1 is evaluated only for worst case, since there are steps in it varies for different distribution of the active intervals and the bytes each UE requests to receive.

The worst case scenario occurs when step 4.c is executed for all the steps. In this case the complexity is evaluated as $O(N^2 \cdot \log\log N + K_E \cdot N)$.

5.3.4.4 TD and FD Post-Processing

After the resource allocation step packets, qualified in TD scheduling step for each UE, are being served based on priorities defined in relation 5.3.
Packets that are not served in this TTI are checked for time delay violation. If they expire they are simply dropped otherwise they are transferred to the next subframe. The transfer involves the insertion of their metadata in the next subframe in a way similar to the packet insertion process that is performed upon the packet arrival.

5.4 Conclusions

An analytical description of the proposed scheduling algorithm was presented in this chapter. The algorithm constitutes a complete solution of the scheduling problem in LTE that attempts to address all of its various aspects with minimum complexity. The introduction of the COFRTS tree enables the efficient storage of all the useful parameters, providing the eNB scheduler with immediate access. Appropriate preprocessing results in an even distribution of the traffic as well as in the construction of additional tree structures that simplify the final resource allocation process, taking advantage of the scheduler's idle time. A time domain scheduling part, utilizing the COFRTS tree, distinguishes the most significant from QoS perspective packets that require immediate serving, thus improving the QoS awareness properties of the algorithm. Finally two solutions for the frequency domain scheduling part are proposed. Both solutions allocate the resources efficiently utilizing the additional tree structures and have different levels of complexity and as a consequence different expected performance.

6. SIMULATION SCENARIO RESULTS AND DISCUSSIONS

6.1 Introduction

Mathematic analysis of every proposed scheduling scheme is perhaps the most rigorous tool to analyze and assess its performance. However given the nature of the scheduling problem and the available mathematical tools an attempt to achieve this is quite difficult if not impossible. Despite the existing difficulties, performance assessment is necessary. In order to achieve this, simulations are employed which aim to model the system as closer as possible to real world ones. The validity of the analysis based on simulations relies on three factors: a) the capabilities of the simulation tool in order to receive trustworthy results regarding scheduler's performance b) the adopted simulation scenario, which include hypotheses about the network, the channel model, the parameters for traffic generation etc, and c) the metrics that are being used in order to assess the results obtained from the simulations.

In the beginning of the chapter we present the system components for the simulation scenario: networks parameters, channel model, traffic model, simulation scenario parameters. Thereafter we describe all the measures we use in order to assess the performance of the algorithms as well as to compare them. Lastly different scenarios are considered in order to assess the performance of the algorithms proposed in chapter 5.

6.2 System and Simulation Scenario

6.2.1 Network and System Parameters

Network and system parameters include choices and general assumptions, required for the simulator configurations, pertaining the network itself as well as the characteristics for some of the system's components. The values or features for the particular components follow the specifications of the system's scenario described in 3GPP's technical report [59] for LTE.

Table 9: **Network and system parameters**

Cells Scheme	Hexagonal, eNBs in Middle
Sectors per Cell	3
Cell Radius (m)	700
Number of Cells	1
Number of eNBs per Cell	1
eNB's Features	
1. eNB's antenna height (m)	30
2. Transmission Schemes	SISO
3. Antennas Types	
• Max Transmission Power	30 dBm
• Antennas Pattern	$A(\theta) = -\min\left[12\left(\frac{\theta}{\theta_{3dB}}\right)^2, A_m\right]$ $A_m = 20\ dB\ max\ attenuation$
UE features	
1. UE height (m)	1
2. Max Number of Antennas	1
3. Antennas Types	Omni-directional

6.2.2 Channel Model

The adopted channel model is one of the most important components of the simulation process. An accurate real world model is difficult to be developed due to the fact that such models involve parameters that are dependent on the geometry of the space where the signal propagates and the materials that objects are made of. Such good channel models that take into consideration the geometry of the space exist but we do not consider any of them. Instead the model that we adopt in this thesis belongs in the class of the so called statistical or empirical models in which the received power in decibels can be calculated as: $P_r = P_t - L$ where P_r is the received power, P_t is the transmitted power and L are the losses.

The important component that is calculated in the aforementioned channel models is the total losses. In statistical models losses are distinguished in two types of fading: a) large scale fading which depends on parameters such as distance between the transmitter and the receiver, carrier frequency, shadowing effects etc, and b) fast fading which are a consequence of the multiple paths that the signal propagates prior to arriving at the receiver.

In our scenario the large fading losses will be calculated by the model proposed in [59], which applies in an urban environment, where the path losses are expressed by the following relation:

$L_s = 40(1 - 4 \cdot 10^{-3} \cdot h_b) \cdot \log_{10}(d) - 18 \cdot \log_{10}(h_b) + 21 \cdot \log_{10}(f) + 80\ dB$ where h_b is the eNB's antenna height in meters, d is the distance between the eNB and the UE in kilometers and f is the carrier frequency in MHz.

The fast fading losses L_f are derived as a flat Rayleigh fading taking into consideration the Doppler effect for each subcarrier.

Thereafter the overall losses are: $L = L_s + L_f.$ (6.1)

6.2.3 Traffic Modeler

Another important component related to the validity of the simulation is the modeling of the traffic that arrives at the eNB's queues. Generated traffic should correspond as much as possible to real one for all the different traffic classes supported by the system.

In our scenario we consider for every UE five different radio bearers to be established, each one producing a traffic flow with different QCI. Three of these radio bearers are GBR, while the other two are non-GBR. A different service type, representative of each QoS class, is used and the traffic produced in each service is simulated using specific stochastic models. Table 10 lists all the service types that are simulated along with the bit rate limits that are used to determine the colors of the related packets while table 11 lists the specific parameters of the produced traffic.

Table 10: Simulation QoS characteristics

Service Type	GBR/NGBR	QCI	Minimum Bit Rate (kbps)	Maximum Bit Rate (kbps)
VoIP	GBR	1	29	35
Uncompressed Video	GBR	2	200	210
Compressed Video	GBR	4	384	512
ftp	NGBR	6	128	640
http	NGBR	8	256	256

Table 11: Simulation traffic parameters

Service Type	Min Generated Rate (kbps)	Max Generated Rate (kbps)	Latency (ms)	Jitter (ms)
VoIP	29	35	30	30
Uncompressed Video		240	50	15
Compressed Video	128	768	40	-
ftp	160	1280	-	-
http	92	512	-	-

In particular, uncompressed video traffic is simulated by a constant bit rate flow of packets with constant size of 64 bytes.

For VoIP traffic, an AMR codec output is modeled as an ON-OFF traffic source with 50% duty cycle and ON-OFF periods exponentially distributed with mean value 3 sec, according to [60]. During an ON period the bit rate is selected randomly from 6 different values according to [61], while during OFF periods (silence) the bit rate is constant and equal to 1.95 kbps [62]. The VoIP packets are encapsulated in RTP/UDP/IP packets, and the appended headers increase their size accordingly.

The compressed video traffic is modeled by a normal distribution producing packet flows with rate values between the minimum and the maximum generated rate. The packet size is selected randomly between 7 possible values 54, 66, 70, 80, 92, 102 and 114 bytes. The time instances in which the video traffic packets are created depend on whether the frame rate is 25fps or 30fps.

Finally, the traffic produced for http and ftp requests is simulated using the same model without setting a maximum generated rate.

6.2.4 Simulation Scenario Parameters

In this topic we quote the parameters and their values used in our simulation scenario.

Table 12: System's and simulation scenario parameters

Simulation Time Duration (ms)	10000
System's Bandwidth (MHz)	10
▪ Effective Bandwidth (MHz)	9
▪ Number of Subcarriers	600
▪ Number of Scheduling Blocks	50
▪ Carriers per SB	12
▪ Subcarrier Spacing (KHz)	15
▪ OFDM symbols per slot	7
▪ OFDM Control symbols per RB	3
Carrier Frequency (GHz)	2
Frame Structure	FDD
Admission Control	No
HARQ retransmissions	Off
Relay Nodes	No
UEs Mobility	No
Min distance between UEs and eNB	300
UEs Distance form eNB	Distributed as in **Figure 6.1**
SNR to CQI mapping	Table 6.5 (Mode -1)

Target BLER	10%
Packet Size	Variable for different Services
CQI feedback	Error free, Delay Free
Set of MCSs	Table 3

From the channel model we calculate the received power, and estimate the SNR as follows: $SNR = P_r - N$ where $N = N_0 + N_f + 10 \cdot \log_{10} B$, N_0 is noise power spectral density in dBm/Hz, N_f is noise figure and B is channel bandwidth in Hz.

The SNR to CQI mapping is done as in table 13 which was derived in [63] for LTE downlink, Rayleigh flat fading channel and different transmission modes. Values that correspond to mode 1 are used in our simulation scenario, as SISO transmission mode is assumed.

Table 13: SNR to CQI mapping for flat Rayleigh fading channel for BLER≤ 10%.

CQI	SNR			
	Mode - 1	Mode - 2	Mode - 3	Mode - 4
1	1,95	-7	-3,1	-4,8
2	4	-5	-1,15	-2,6
3	6.00	-3.15	1.50	0.00
4	8.00	-1.00	4.00	2.60
5	10.00	1.00	6.00	4.95
6	11.95	3.00	8.90	7.60
7	14.05	5.00	12.70	10.60
8	16.00	6.90	14.90	12.95
9	17.90	8.90	17.50	15.40
10	19.90	10.85	20.50	18.10
11	21.50	12.60	22.45	20.05
12	23.45	14.35	23.20	22.00
13	25.00	16.15	24.90	24.55
14	27.30	18.15	27.00	26.80
15	29.00	20.00	29.10	29.60

Mode - 1: SISO Single eNB antenna, $1\ T_x$ and $1\ R_x$.

Mode - 2: MIMO Transmit Diversity, $2\ T_x$ and $2\ R_x$.

Mode - 3: MIMO Spatial Multiplexing, $2\ T_x$ and $2\ R_x$.

Mode - 4: MIMO Spatial Multiplexing, $4\ T_x$ and $2\ R_x$.

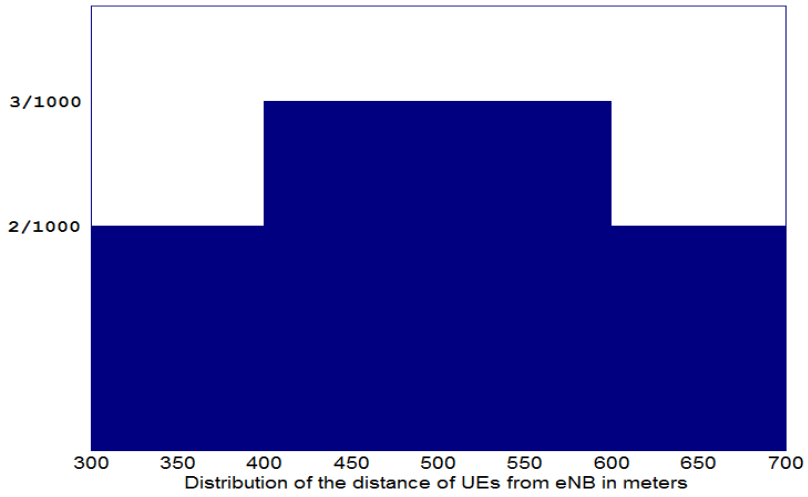

Figure 6.1: pdf of the distances of UEs from eNB's antennas.

6.2.5 Simulation environment

A simulation environment created in C++ was used to evaluate the performance of the algorithms presented in Chapter 5. The simulator, that was initially developed for the purposes of [53] and [64], is properly designed and adapted to accurately model the link layer functionality of an LTE network cell. It mainly focuses on the scheduling functions of the eNB and enables the development and evaluation of scheduling and resource allocation algorithms. Its overall design concept is depicted in the following block diagram.

The different blocks in Figure 6.2 represent specific entities of the simulator along with their interactions. Instances of those entities are implemented as objects in C++.

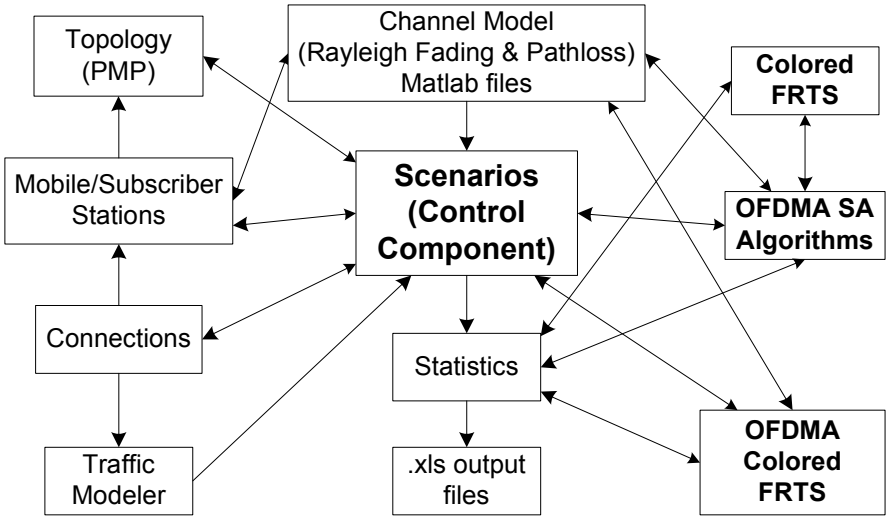

Figure 6.2: Simulation model implemented in C++.

Each entity serves specific purposes. The "Scenario" entity is used to coordinate the overall functionality of the simulator platform. It is the entity that enables the introduction of all the different variables of a simulation scenario such as the traffic kinds, the number and the type of connections per UE, the lowest number of UEs that will be simulated in the scenario as well as the step and the highest number, the duration of the

simulation as a number of TTIs etc. All instances that represent the fundamental components of a network simulation scenario are initiated inside the “Scenario” entity.

The “Connections” entity facilitates the creation and management of all the radio bearers that every UE establishes according to the simulation scenario. The entity allows the user to configure all the established bearers by adjusting various related parameters. A similar entity that is linked to the “Connections” entity is the “Mobile/Subscriber Stations” entity. It contains all the objects that represent the UEs of the simulated scenario according to the number set in the scenario entity.

The “Channel Model” enables the introduction of channel state information for every UE at any time instant and subchannel. In our simulation scenarios the channel model used is described in paragraph 6.2.2. Similarly the “Traffic Modeler” entity provides the other network’s input by producing traffic flows that accurately model real time and non real time services. The traffic models used in our scenarios are analyzed in paragraph 6.3.2. It should be noted that both “Channel Model” and “Traffic Modeler” could be expanded to include more channel and traffic models than those used for the scope of this thesis.

“Colored FRTS” and “OFDMA colored FRTS” entities are designed so that the tree structures, presented in Chapter 4, can be implemented. The “Colored FRTS” entity is used for the construction of the basic tree structure used for the time domain scheduling, while the “OFDMA colored FRTS” entity is extended to include the complementary subtrees that support the resource allocation decisions. The “OFDMA SA Algorithms” entity provides the means for the development of the resource allocation algorithms such as the Proportionally Fair, Maximum Throughput and Resource Allocation 1 and 2 described in Chapter 4.

The “Statistics” entity provides the means for monitoring the performance of the network as far as throughput and other QoS metrics are concerned. It is also used to update predetermined Excel files with the final results. The results that are presented in paragraph 6.4 have been extracted from the content of these excel files.

Additional entities such as the “Topology” entity are also included in the simulator, providing more capabilities. However they have not been used for the scope of this thesis.

6.3 Performance Measures – Metrics

Performance metrics are of great importance since the assessment of the performance of the proposed schemes is based on the results of their comparison. Thus it is essential that the performance metrics have specific properties which at least accurately measure a certain target that each algorithm aims to meet in order to highlight the differences or the similarities among the compared schemes. Some of the most widely used performance metrics found in the literature that we will be using in our assessment of the proposed algorithms are the following: *System’s throughput, Fairness, Packet Loss Ratio, Packet Delay.*

- *Average Throughput per UE and System’s Average Throughput*.

  This metric is a good indicator of the average transmission rate that the system can carry. A formal definition is derived as follows: let $b_{k,t}$ be the number of bytes UE U_k receives at TTI t, then the *Average Throughput* for UE U_k for time period of $T\ TTIs$ is defined as: $Th_{U_k} = \frac{1}{T}\sum_{t=1}^{T} b_{k,t}$ and *System’s Average Throughput* for time period of $T\ TTIs$ is defined as: $S_{Th} = \sum_{k=1}^{N} Th_{U_k}$

- *Fairness.*

  The concept of fairness among UEs is defined in terms of the proportion of throughput that each UE should achieve from the overall throughput given that any UE should achieve the maximum bit rate for all of its bearers. There is a variety of fairness measures and each one has certain properties and is applied differently.

  Jain's Fairness Index proposed in [65] is defined as: $JFI = \frac{\left(\sum_{k=1}^{N} x_k\right)^2}{N \cdot \sum_{k=1}^{N} (x_k)^2}$ where $x_k = \frac{Th_{U_k}}{\sum_{s \in \boldsymbol{Bearers\ of\ U_k}} MBRate(s)}$. Jain's Fairness Index is a commonly used fairness metric that quantifies fairness among UEs that belong in different traffic classes. The metric's values range continuously in the interval [0 1] with fairness increasing when its value increases and absolute fairness achieved when it equals one.

  Another fairness measure proposed in [66] is defined as the normalized deviation between the bit rate for UE U_k to average bit rate, formally:

  $F = \frac{1}{N} \sum_{k=1}^{N} \frac{\left| Th_{U_k} - \overline{Th} \right|}{\overline{Th}}$ where $\overline{Th} = \frac{1}{N} \sum_{k=1}^{N} Th_{U_k}$. This metric is a decreasing function of fairness i.e. the greater the fairness the smallest the metric.

- *Packet Loss Ratio.*

  Packet Loss Ratio is a measurement of how good does a scheduling algorithm behaves in scheduling the packet within time constraints determined by QCI. It is defined as the ratio of the total size of the discarded packets to the size of the total arrived packet size. Formally let $P_{k,t}$ be all the received packets and $P_{discard,k,t}$ be the discarded packets of UE U_k at TTI t, then the loss ratio is:

  $$LR = \frac{\sum_{k=1}^{N} \sum_{t=1}^{T} Size(P_{discard,k,t})}{\sum_{k=1}^{N} \sum_{t=1}^{T} Size(P_{k,t})}$$

- *Delay.*

  Packet Delay is a QoS measure which in combination with the Packet Loss metric provides an accurate indication whether the scheduling algorithm respects the QoS requirements or not. It is calculated as follows: let $P_{ij}(k)$ be a packet, that has been scheduled successfully, which belongs to bearer B_j of UE U_i and k be the arrival sequence of the successfully scheduled packets. Then the delay time D_{ij} for that packed is defined as:

  $D_{ij}(k) = t_{scheduled}\left(P_{ij}(k)\right) - t_{arrival}\left(P_{ij}(k)\right)$. Then the *Mean Delay per Bearer* is defined as: $D_{B_j} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{M_{ij}} D_{ij}(k)}{\sum_{i=1}^{N} M_{ij}}$ and the *System's Packet Delay* as: $S_{Delay} = \frac{1}{L} \sum_{j=1}^{L} D_{B_j}$ where L is the number of bearers.

It is noted that all the performance metrics, even though some of them measure bytes or bits, are packet oriented or are related somehow to packets in accordance to the designed features of LTE system as a packet optimization network.

6.4 Simulation Results and Discussions

The new algorithms presented in Chapter 5 have been compared to the Proportionally Fair and Maximum Throughput algorithms using the simulation assumptions and the simulation environment described in the previous paragraphs. Due to the large number of the PF algorithm variations found in the literature, the selected version is the one described and analyzed in paragraph 4.3.2.2. It was selected as an appropriate one for the LTE that takes into account all its restrictions. This algorithm is properly adapted so that it can be executed in combination with the TD scheduling part that was described in Chapter 5. The analytical steps of this algorithm can be found in Appendix II. The Maximum Throughput algorithm was implemented following exactly the same steps with the difference of using the Maximum Throughput metric for sorting the UEs instead of the PF metric. By no means does this algorithm achieve a throughput maximization of the system. Maximizing the throughput of a multicarrier system with the constraints of LTE would be an NP-hard problem. However it involves choices and decisions that favor the total throughput. The name of the algorithm is due to the metric used which is the same as the one used in the typical single carrier Maximum Throughput algorithm.

The simulations aim to reveal the advantages and drawbacks of each algorithm by evaluating them in different aspects of a network's performance.

Throughput: As already mentioned one of the most common performance metrics is the total achieved throughput which represents how efficiently the radio resources are utilized. The following diagram depicts the total average throughput of all algorithms as a function of the number of UEs connected to the eNB.

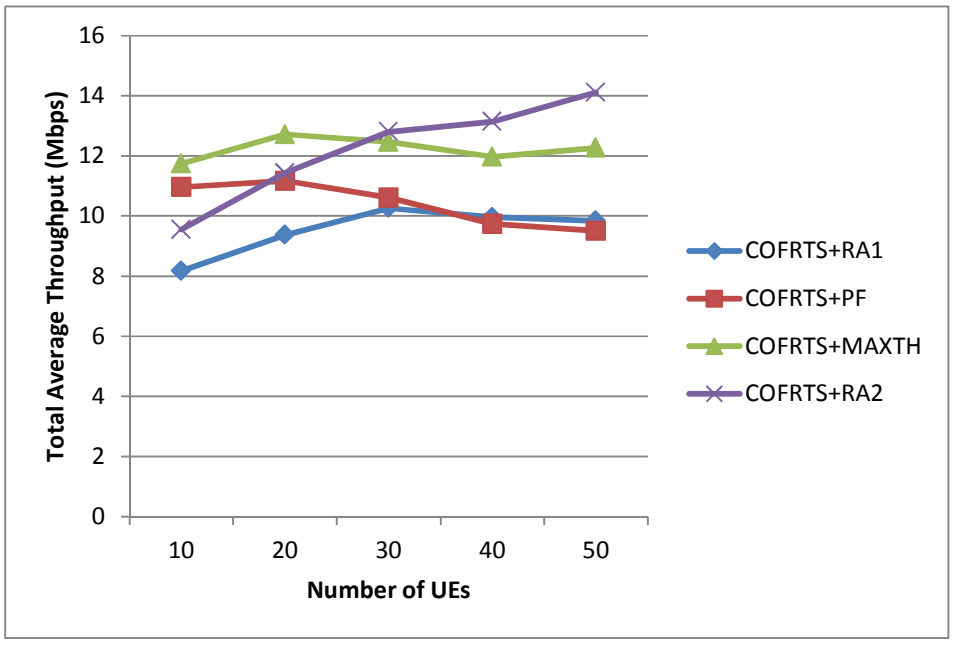

Figure 6.3: Total average throughput as a function of the number of UEs

As it was expected, both RA1 and RA2 are favored by the multiuser diversity since they exhibit considerable throughput increase as the number of UEs increases. RA1 shows a poor performance for low number of UEs due to the fact that only the best MCS is examined. This increases the likelihood of having SBs where no UE exhibits favorable channel conditions, resulting in a potential waste of resources. This situation is improved in RA2 where the option of a second MCS examination is allowed. However in terms of throughput, conventional methods like PF and Maximum Throughput outperform RA1 and RA2 for low number of UEs. As the number of UEs increases, both RA1 and RA2 exhibit improved performance, with RA2 outperforming even Maximum

Throughput. RA1 eventually achieves throughput values at the level of PF for high number of UEs.

Fairness: In order to evaluate fairness the two metrics referenced in paragraph 6.3 are calculated for all different numbers of UEs. The results are depicted in figures 6.4 and 6.5.

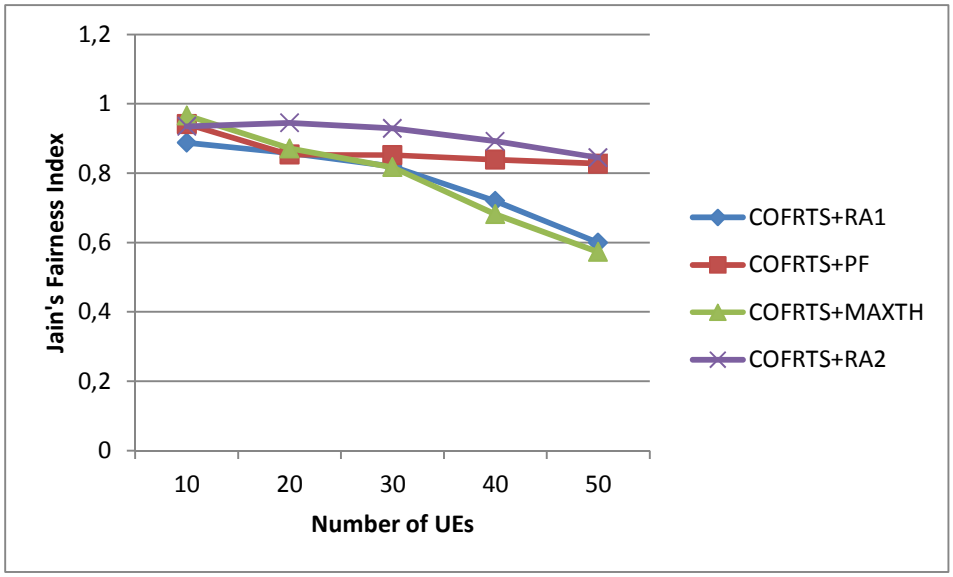

Figure 6.4: Jain's Fairness Index as a function of the number of UEs

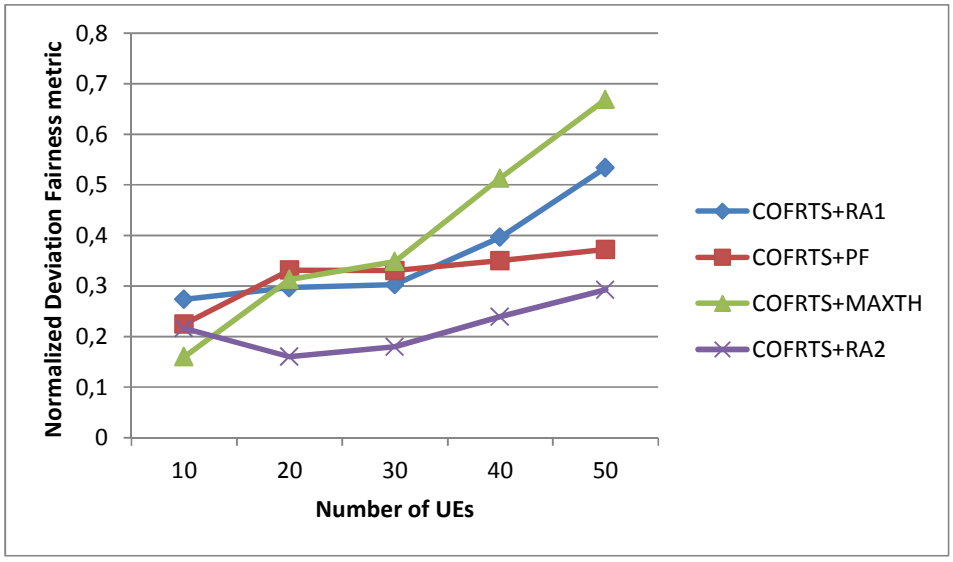

Figure 6.5: Normalized Deviation Fairness metric as a function of the number of UEs

It is reminded that JFI metric is an increasing function of an algorithm fairness performance, while the normalized deviation fairness metric is inversely proportional to the fairness performance. With this in mind, it can be easily concluded that RA2 algorithm shows exceptional fairness properties outperforming even PF. The performance of RA1 lies between the performance of PF and Maximum Throughput. It should be noted that even though PF has well known theoretically founded fairness properties, the preceded time domain scheduling may alter them to a certain extent. The same stands for the performance of Maximum Throughput as far as throughput is concerned. However, for a system such as the LTE with strict QoS requirements, some form of QoS-aware packet distinction, as the one executed in the Time Domain, is mandatory. Another important observation from the diagrams above is that the differences in the fairness performance become clear as the number of UEs increases.

This is a consequence of the fact that in lower amount of traffic situations all algorithms succeed in serving all UEs (no matter how far from the eNB they are) in a satisfactory degree.

It is also interesting to examine how the algorithms perform in the combined throughput fairness space. This is illustrated in figure 6.6 where each point corresponds to a specific algorithm and a specific number of UEs.

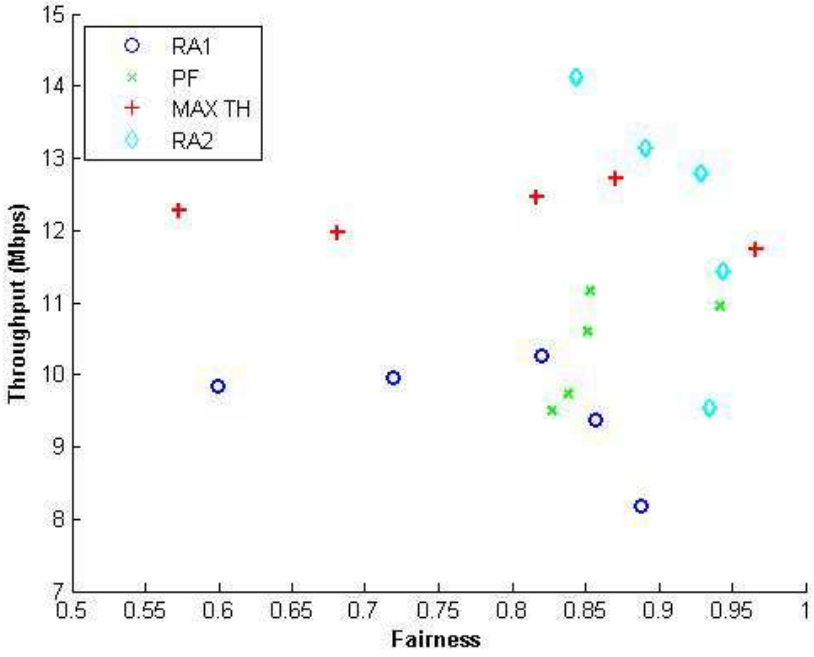

Figure 6.6: Combined fairness-throughput performance

Delay: Regarding the total mean delay the performance is shown in figure 6.7. Both RA1 and RA2 exhibit better properties regarding the total mean delay of the served packets.

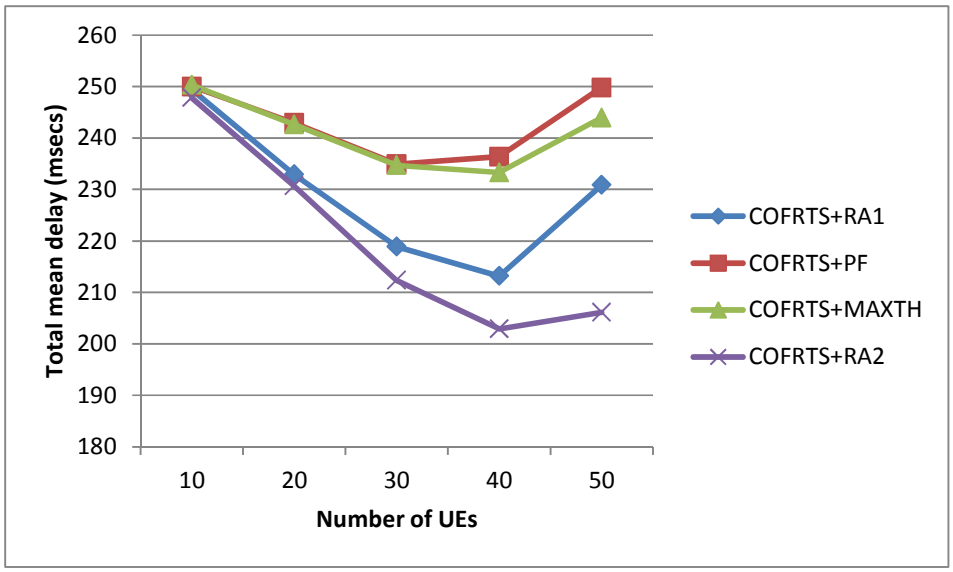

Figure 6.7: Total mean delay as a function of the number of UEs

The exceptional delay properties of RA1 and especially RA2 are justified by the fact that those algorithms show a beneficial treatment to high priority services, as will be depicted below. The packets of those services have lower maximum allowable delay

limits as has already been mentioned. As a result of this, conforming to the maximum allowable delay of such traffic flows leads to a lower total mean delay.

The mean delay encountered in each of the high priority services is shown in figures 6.8 and 6.9. All algorithms show similar VoIP delay with RA2 having a slightly better performance. Some differences are observed in the uncompressed video mean delay with RA2 having improved performance compared to PF and Maximum Throughput as the number of UEs increases. The mean delay performance for the lower priority classes is comparable for all the algorithms and is approaching the maximum allowable delay value, as the algorithms are tested under high traffic load. It should also be mentioned that the achieved delay is mostly affected by the time domain scheduling that is common in all the examined techniques.

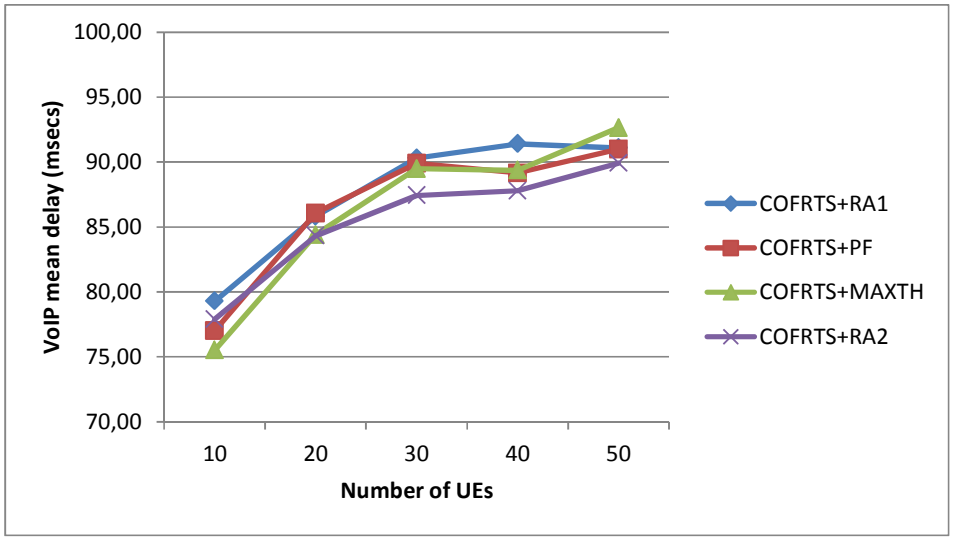

Figure 6.8: VoIP mean delay (QCI=1)

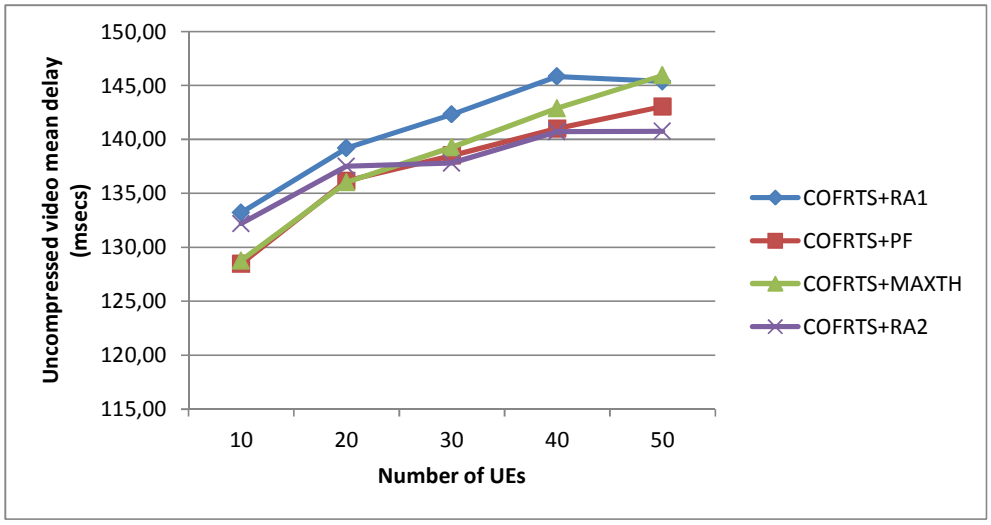

Figure 6.9: Uncompressed video mean delay (QCI=2)

QoS classes: As far as serving the various QoS classes is concerned, figures 6.10 to 6.12 show the achieved throughput per UE for each one of GBR classes.

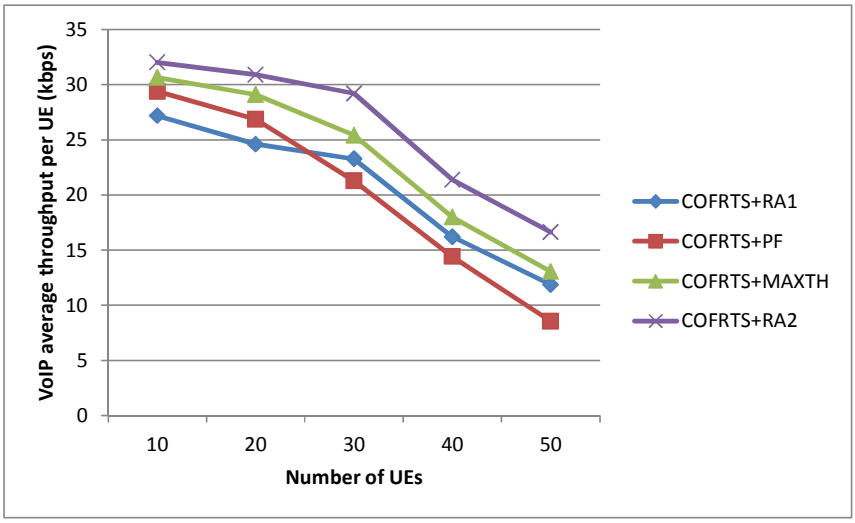

Figure 6.10: VoIP average throughput (QCI=1)

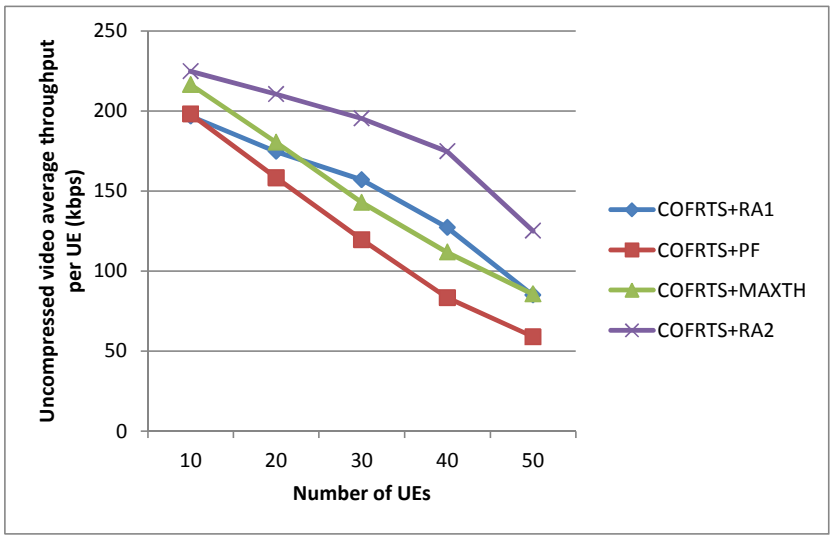

Figure 6.11: Uncompressed video average throughput (QCI=2)

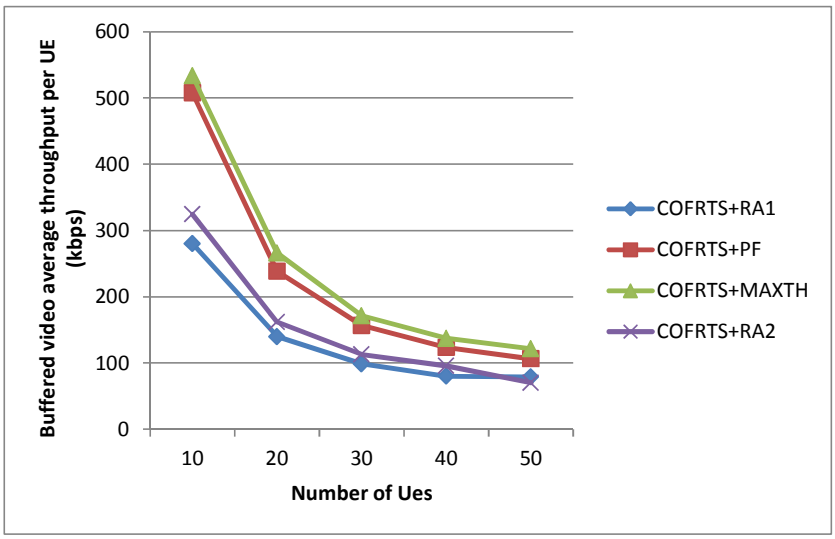

Figure 6.12: Buffered video average throughput (QCI=4)

RA2 clearly shows higher throughput for classes with QCI of higher priority such as VoIP (QCI=1, priority 2) and uncompressed video (QCI=2, priority =4). RA1 shows

satisfactory performance as far as high priority services are concerned. However this is achieved with the cost of worse performance in lower priority services for RA1 and RA2 compared to the conventional PF and Maximum Throughput algorithms.

Similar conclusions can be attained from observing the percentage of served packets as a function of the number of UEs for each of the GBR classes. Once again RA2 exhibits an exceptional performance in the high priority services, given that the percentage of served packets for VoIP and uncompressed video are relatively high. For VoIP specifically, this percentage approaches almost 100% for low number of UEs. RA1 also has a satisfactory performance for the high priority services, considering its low complexity. As the number of UEs increases, it outperforms PF in VoIP and both PF and Maximum Throughput in uncompressed video.

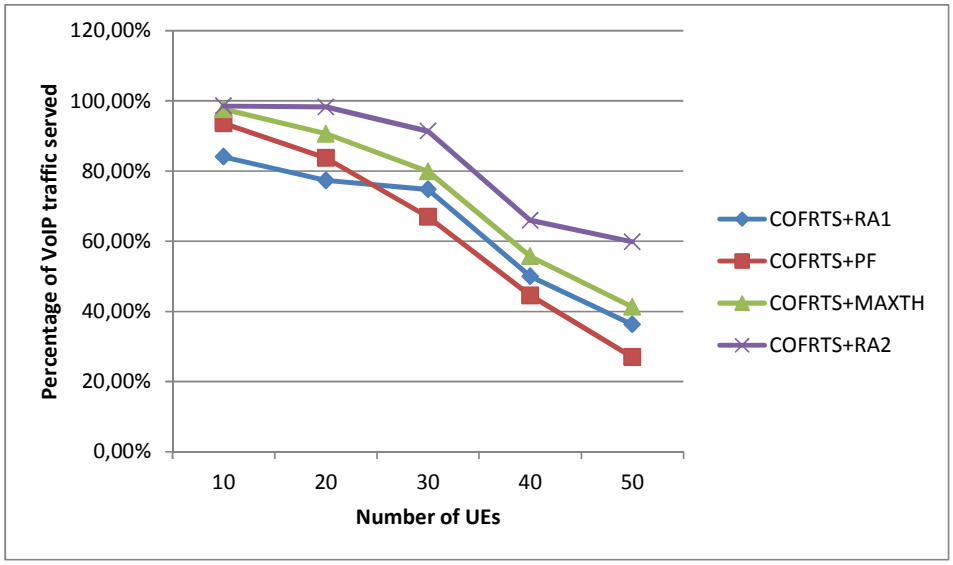

Figure 6.13: Percentage of VoIP packets served (QCI=1)

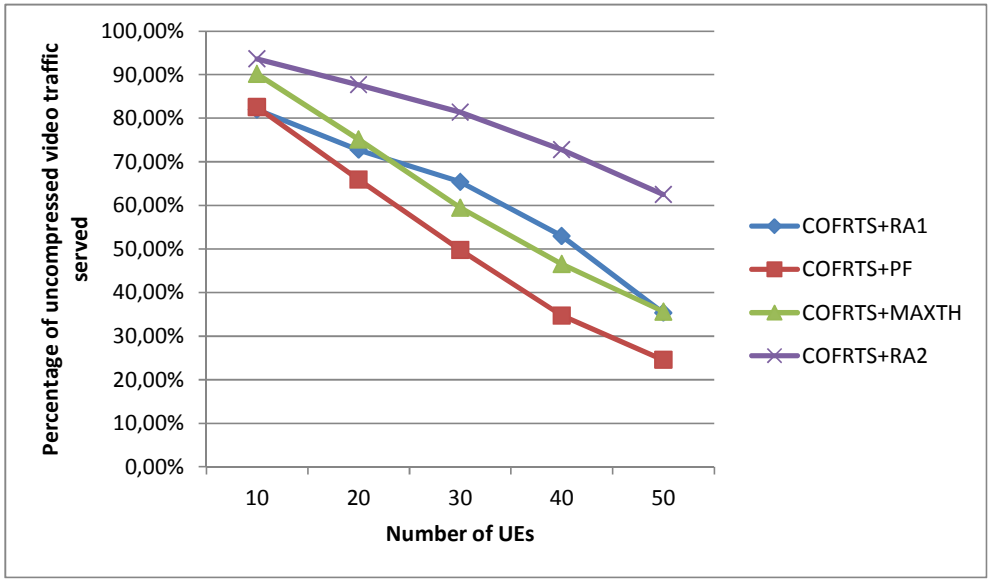

Figure 6.14: Percentage of uncompressed video packets served (QCI=2)

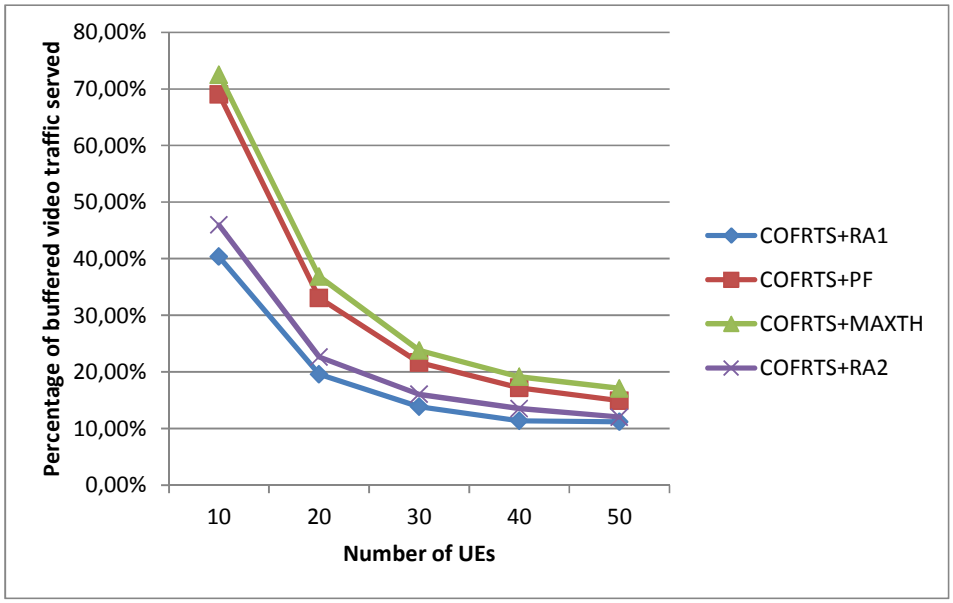

Figure 6.15: Percentage of buffered video packets served (QCI=4)

Summarizing, the following conclusions can be drawn from the previous simulation results:

- RA2 exhibits the best performance in terms of throughput and fairness achieving high values for both metrics.
- Both RA1 and RA2 are favored by multiuser diversity as their performance improves considerably with the increase in the number of UEs.
- RA1 exhibits satisfactory performance, outperforming PF in terms of throughput and Maximum Throughput in terms of fairness, as the number of UEs increases, making it an appropriate solution in a multiuser environment due to its low complexity.
- RA2 shows a beneficial treatment to services with high QoS demands by exploiting the prioritization imposed by LTE standard with the cost of worse performance in lower priority services.
- RA1 also achieves higher throughput, compared to PF and in some cases to Maximum Throughput, in high priority services.
- RA1 and RA2 are better solutions for delay sensitive services even though the delay performance is mainly influenced by the TD scheduling part of the algorithm.

7. CONCLUSIONS

In this thesis, a complete scheduling and resource allocation solution for the LTE downlink is presented, analyzed and evaluated, after a brief summary of the LTE theory and a detailed study of the related literature. The overview of the LTE theory, presented in the first chapters, focuses mainly on the basic standardized parameters that affect resource allocation. Chapter 4 contains a concise review of the literature regarding the proposed LTE scheduling algorithms. Based on ideas that arise from the study of the literature, a new scheduling solution for the downlink was proposed called Colored OFDMA Frame Registry Tree Scheduler (COFRTS). The solution aims to provide efficient scheduling in the sense that the strict LTE QoS restrictions are met with minimum complexity cost.

The functionality of the scheduler is based on a sophisticated tree structure and is divided in two parts, the TD scheduling and the FD scheduling. In TD scheduling, the arriving packets are scheduled appropriately in time so that the maximum delay limits are satisfied. This is reflected by the insertion of their metadata under appropriately selected nodes of the COFRTS tree. The metadata include the size and the required resources, named Scheduling Blocks in LTE, along with specific UE info such as the best supported MCS and the RI.

The purpose of the FD scheduling part is to allocate the available SBs to the UEs that have been qualified in the preceded TD part. Common widely accepted techniques such as PF and Maximum Throughput may be used. However the limited time that is available for the final scheduling decision led us to the design of some newly proposed algorithms that, combined with the TD part, exhibit comparable performance with low complexity. The reduced complexity of these algorithms is based on two additional tree structures with the more interesting of them being an interval tree that stores efficiently the parts of the spectrum where each UE can reliably receive. The other complementary subtree structure is a priority queue that sorts the UEs that are scheduled to receive data in the current subframe according to the size of these data. Two algorithms were proposed that utilize these tree structures, named RA1 and RA2. The design of RA1 is based on providing a solution with minimum possible complexity, while RA2 exhibits improved performance with the cost of some additional complexity.

Extended simulation tests were conducted to verify the expected advantages of the proposed algorithms compared to the PF and Maximum Throughput algorithms. The results indicate the exceptional performance of RA2 in terms of throughput and fairness and also verify RA1’s satisfactory performance as the number of UEs increases. Both algorithms offer a beneficial treatment to high priority services as indicated from the increased throughput and percentage of served packets values of the individual QoS classes. This also justifies the observed relatively low mean delay values as those services are more delay sensitive.

Following the research results presented in this thesis, an adaptation of the proposed algorithm in the uplink case may be considered as an interesting future extension. The requirement for contiguous spectrum assignment in the uplink requires a different approach at least in the frequency domain scheduling part, so that this additional restriction is satisfied. Furthermore, the implementation and utilization of a complete and realistic MIMO channel model may also allow the examination of how proper selection of MIMO parameters (rank, number of codewords, precoding matrix) as well as employment of advanced MIMO techniques (e.g. multi-user MIMO) can be exploited to lead to efficient scheduling decisions.

ABBREVIATIONS

3GPP	3rd Generation Partnership Project
ACK	Acknowledgement
AMC	Adaptive Modulation and Coding
AMR	Adaptive Multi-Rate
APN-AMBR	Access Point Name Aggregate Maximum Bit Rate
ARP	Allocation Retention Priority
ARQ	Automatic Repeat Request
BCCH	Broadcast Control Channel
BCH	Broadcast Channel
BE	Best Effort
BET	Best Equal Throughput
BLER	Block Error Rate
BPSK	Binary Phase Shift Keying
CCCH	Common Control Channel
CCE	Control Channel Element
CD	Configurable Dual
CFI	Control Format Indicator
CO-FRTree	Colored OFDMA Frame Registry Tree
COFRTS	Colored OFDMA Frame Registry Tree Scheduler
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CRC	Cyclic Redundancy Check
DCCH	Dedicated Control Channel
DCI	Downlink Control Indicator
DFT	Discrete Fourier Transform
DL	Downlink
DL-SCH	Downlink Shared Channel
DM-RS	Demodulation Reference Signal
DPS	Delay Prioritized Scheduling
DRX	Discontinuous Reception
DTCH	Dedicated Traffic Channel
DwPTS	Downlink Pilot Time Slot

E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
EDF	Earliest Deadline First
eNB	Evolved Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
EXP/PF	Exponential/Proportional Fair
FD	Frequency Domain
FD-CoItA	Frequency Domain- Carrier over Interference to Average
FD-PF	Frequency Domain- Proportionally Fair
FD-PFsch	Frequency Domain- Proportionally Fair scheduled
FDD	Frequency Division Duplexing
FFT	Fast Fourier Transform
FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GP	Guard Period
GSM	Global System for Mobile
HARQ	Hybrid Automatic Repeat Request
HOL	Head of Line
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
ICI	Intercarrier Interference
IFFT	Inverse Fast Fourier Transform
IMS	IP Multimedia Subsystem
ISI	Intersymbol Interference
JFI	Jain's Fairness Index
LI	Length Indicator
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
LWDF	Largest Weighted Delay First
M-LWDF	Modified Largest Weighted Delay First
MAC	Medium Access Control
MAC-I	Message Authentication Code for Integrity
MBR	Maximum Bit Rate

MBSFN	Multicast-Broadcast Single Frequency Network
MCCH	Multicast Control Channel
MCH	Multicast Channel
MCS	Modulation and Coding Scheme
MIB	Master Information Block
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MTCH	Multicast Traffic Channel
NACK	Negative Acknowledgement
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OSI	Open System Interconnection
P-GW	Packet Data Network Gateway
PAPR	Peak to Average Power Ratio
PBCH	Physical Broadcast Channel
PCCH	Paging Control Channel
PCFICH	Physical Control Format Indicator Channel
PCH	Paging Channel
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PF	Proportionally Fair
PHICH	Physical Hybrid ARQ Indicator Channel
PMCH	Physical Multicast Channel
PMI	Precoding Matrix Indicator
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying

RACH	Random Access Channel
RB	Resource Block
RBG	Resource Block Group
RI	Rank Indicator
RIV	Resource Indicator Value
RLC	Radio Link Control
RNC	Radio Network Controller
RNTI	Radio Network Temporary Identifier
ROHC	Robust Overhead Compression
RR	Round Robin
RRC	Radio Resource Control
RTP	Real Time Protocol
S-GW	Serving Gateway
S-RS	Sounding Reference Signal
SB	Scheduling Block
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDU	Service Data Unit
SINR	Signal to Interference Ratio
SISO	Single Input Single Output
SN	Sequence Number
SNDR	Signal to Noise Distortion Ratio
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
TB	Transport Block
TBR	Target Bit Rate
TCP/IP	Transmission Control Protocol/ Internet Protocol
TD	Time Domain
TDD	Time Division Duplexing
TTA	Throughput to Average
TTI	Transmission Time Interval
UE	User Equipment
UE-AMBR	User Equipment Aggregate Maximum Bit Rate
UDP	User Datagram Protocol
UL	Uplink

UL-SCH	Uplink Shared Channel
UMTS	Universal Mobile Telecommunications System
UpPTS	Uplink Pilot Time Slot
VoIP	Voice over Internet Protocol
VRB	Virtual Resource Block
W-CDMA	Wideband Code Division Multiple Access

APPENDIX I

The construction of the interval tree is quite straightforward. At first, the intervals where transmission with the best MCS possible is supported are determined for every UE. These are called active intervals of SBs. All the interval edges of those active intervals are then specified and sorted in ascending order. The elementary intervals are formed by all possible pairs of consecutive interval edges. Every node of the tree represents an interval of SBs, with the leaves of the tree representing the elementary intervals. The key of each node is a properly selected interval edge within the node's interval. Then, starting from the root of the tree that represents the entire bandwidth, an internal interval edge is selected as a key using a certain rule (the rule applied in our simulations is to select a SB so that the interval formed by the lower edge and this SB contains an amount of elementary intervals equal to the highest possible power of 2). This key, referenced as median SB, divides the entire bandwidth in two intervals that form the child nodes. The same procedure is repeated for every child node until only nodes representing the elementary intervals are left.

Interval Tree Construction Algorithm

K: number of SBs
$S \coloneqq \{c_1, c_2, \dots, c_K\}$: set of SBs
E: set of interval edges
I_E: set of elementary intervals

$I = (0, c_K]$
$root \coloneqq new\ IntTree(I)$

$IntTreeNode\ \ IntTree(Interval\ I)$
{
  $if\ (I \in I_E)$
    $return\ null$
  $else$
    $e_{med} = pickMed(I)$ //picks an interval edge from E that is contained in I
    $L = (I.low, e_{med}]$ //left child interval
    $R = (e_{med}, I.high]$ //right child interval
    $left \coloneqq new\ IntTree(L)$ //left child node
    $right \coloneqq new\ IntTree(R)$ //right child node
}

APPENDIX II

The pseudocode of the PF algorithm that was used in the simulation is presented below, after listing the used notation:

Notation

- N: Total number of User Equipments – UE.
- $\mathbf{U}$: The set of UEs, i.e. $\mathbf{U} = \{\mathrm{U}_1, \mathrm{U}_2, \dots \dots \mathrm{U}_\mathrm{N}\}$.
- $\mathbf{M} := \{\mathrm{m}_1, \mathrm{m}_2, \dots \dots \mathrm{m}_\mathrm{W}\}$ the ordered set of all available MCSs, W number of available MCSs.
- $\mathbf{b} := \{\mathrm{b}_1, \dots \dots \mathrm{b}_\mathrm{N}\}$, where b_i is the number of available bytes for UE U_i to send.
- $\mathbf{MCSI}_{\mathbf{U_i}}$ is a data structure which contains all the active intervals for each MCS that are supported for UE U_i.

PF Algorithm

1. $\boldsymbol{for}\ i = 1\ to\ N$
2. $(\mathrm{I}_{\mathrm{U_i}}, \mathrm{m}_{\mathrm{U_i}}) = \mathrm{argmax}_{(\mathrm{I_j} m_j) \in \mathbf{MCSI}_{\mathbf{U_i}}} S_B(m_j, \mathrm{I_j})$ // get SBs that achieve the maximum possible rate using as MCS $\mathrm{m}_{\mathrm{U_i}}$.
3. $\bar{R}_i(t) = (1-a)\bar{R}_i(t-1) + a\, R_i(t-1)$ //$a \in [0\ 1]$
4. $\varphi_i = \frac{r_{\mathrm{m}_{\mathrm{U_i}}}}{\bar{R}_i(t)}$ // $r_{\mathrm{m}_{\mathrm{U_i}}}$ is the maximum rate $\mathrm{U_i}$ can achieve
5. $\bar{R}_i(t-1) = \bar{R}_i(t)$ // set mean bit rate for the next frame
6. ***end***
7. $[\varphi_{\theta(1)}, \varphi_{\theta(2),\dots\dots} \varphi_{\theta(U)}] = \boldsymbol{Ord}([\varphi_1, \varphi_{2,\dots\dots} \varphi_U])$ // order set of φ_i so $\varphi_{\theta(k)} \geq \varphi_{\theta(k+1)}$ $\theta(k)\ is\ a\ function\ that\ maps\ k\ to\ a\ single\ i$

// resource allocation is done in sequential fashion, start with highest ranked user $\theta(1)$.

8. $\mathcal{N} = S$ // $\mathcal{N}$ is the current set of available SBs
9. $\boldsymbol{for}\ k = 1\ to\ N$ // for each UE
10. $R_{\theta(k)} = \mathcal{N} \cap K_{\theta(k)}$ // $R_{\theta(k)}$ contains resources for UE $\theta(k)$
11. $\mathcal{N} = \mathcal{N} - R_{\theta(k)}$ // remove allocated SBs from current available set
12. ***end***

REFERENCES

[1] T. S. Rappaport. Wireless Communications: Principles and Practice, 2nd ed. Prentice Hall, 2002.
[2] http://www.budde.com.au/Research/2013-Global-Mobile-Communications-Statistics-Trends-and-Regional-Insights.html
[3] http://www.itu.int/ITU-R/index.asp?category=study-groups&rlink=rsg5-imt-advanced&lang=en
[4] 3GPP Technical Report-TR 36.912
[5] 3GPP TS 36.211 V11.5.0 Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation
[6] 3GPP TS 36.300 V11.9.0, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Overall Description.
[7] 3GPP TS 23.401 V11.9.0 General Radio Packet Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access.
[8] 3GPP TS 23.002 V11.6.0 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Network architecture.
[9] A. Larmo "The LTE Link-Layer Design", IEEE Commun. Mag., 2009.
[10] 3GPP TS 36.323 V11.2.0 Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification.
[11] 3GPP TS 36.322 V10.0.0 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification.
[12] 3GPP TS 36.321 V11.4.0 Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification.
[13] Najah A. Abu Ali, Abd-Elhamid M. Taha, Hossam S. Hassanein, "Quality of service in 3GPP R12 LTE-advanced", IEEE Communications Magazine, 2013.
[14] 3GPP TS 23.002 V11.6.0 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Policy and charging control architecture.
[15] 3GPP TS 23.107 V11.0.0 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Quality of Service (QoS) concept and architecture.
[16] 3GPP TS 23.207 V11.0.0 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; End-to-end Quality of Service (QoS) concept and architecture.
[17] Mehdi Alasti, Behnam Neekzad, Jie Hua and Rath Vannithambi, "Quality of Service in WiMAX and LTE Networks", IEEE Communications Magazine, pp 104-111, May 2010.
[18] C. Gessner, "UMTS long term evolution (LTE) technology introduction," Rohde & Schwarz Application Note 1MA11, 09.2008.
[19] 3GPP TS 36.213 V11.5.0 Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures
[20] Q. Li, G. Li, W Lee, M. il Lee, D. Mazzarese, B. Clerckx, and Z. Li, "MIMO Techniques in WiMAX and LTE: A Feature Overview," Communications Magazine, IEEE, vol. 48, no. 5, pp. 86-92, May 2010.
[21] Lee, J., Han, J. K. and Zhan, J. (2009) "MIMO technologies in 3GPP LTE and LTE-Advanced", EURASIP Journal on Wireless Communications and Networking, 2009, article ID 302092.
[22] Alamouti, S. (1998), "A simple transmit diversity technique for wireless communications", IEEE Journal on Selected Areas in Communications, 16, 1451–1458.
[23] 3G Americas (June 2009) MIMO Transmission Schemes for LTE and HSPA Networks.
[24] 3G Americas (May 2010) MIMO and Smart Antennas for 3G and 4G Wireless Systems: Practical Aspects and Deployment Considerations.
[25] 3GPP TS 36.212 V11.4.0 Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding.
[26] Erik Dahlman, Stefan Parkvall, Johan Sköld and Per Beming, "3G Evolution: HSPA and LTE for Mobile Broadband", Second Edition.
[27] NEC, 'R1072119:DLUnicastResourceAllocationSignalling', www.3gpp.org, 3GPPTSGRAN WG1, meeting 49, Kobe, Japan, May 2007.
[28] I. Griva, S.G. Nash and A. Sofer, "*Linear and Nonlinear Optimization"*, Society for Industrial and Applied Mathematics, 2009.
[29] Honghai Zhang, Narayan Prasad and Sampath Rangarajan, "MIMO Downlink Scheduling in LTE systems" in 31st Annual *IEEE* International Conference on Computer Communications, Orlando FL, 2012.
[30] Hossam Fattah and Hussein Alnuweiri, "A Cross-Layer Design for Dynamic Resource Block Allocation in 3G Long Term Evolution System" in 6th *IEEE* International Conference on Mobile Adhoc and Sensor Syatems, Macau, 2009.
[31] Yehuda Ben-Shimol, Itzik Kitroser and Yefim Dinitz, "Two-Dimensional Mapping for Wireless OFDMA Systems", *IEEE* Transactions on Broadcasting, vol. 52, no. 3, 2006.

[32] F. Capozzi, G. Piro, L.A. Grieco, G. Boggia and P.Camarda, “Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey”, IEEE Communications Survey & Tutorials, vol. 15, no. 2, 2013.
[33] A. S. Tanenbaum, *Modern Operating Systems*, 3rd ed, Prentice Hall Press, 2007.
[34] A. Stolyar and K. Ramanan, “Largest Weighted Delay First Scheduling: Large Deviations and Optimality,” *Annals of Aplied Probability*, vol. 11, pp. 1–48, 2001.
[35] D. Liu and Y.-H. Lee, “An efficient scheduling discipline for packet switching networks using Earliest Deadline First Round Robin,” in *Proc. IEEE Int. Conf. on Computer Commun. and Net., ICCCN*, Oct. 2003, pp. 5 – 10.
[36] Raymond Kwan, Cyril Leung and Jie Zhang, “Proportional Fair Multiuser Scheduling in LTE” IEEE Signal Processing Letters, Vol. 16, N.6, June 2009.
[37] Ioan Sorin Comsa, Sijing Zhang, Mehmet Aydin, Pierre Kuonen and Jean-Frederic Wagen “A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-Advanced Technologies Using Neural Networks,” in 37th Annual *IEEE* Conference on Local Computer Networks*,* Clearwater, Florida, 2012.
[38] 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.
[39] 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.
[40] J.-H. Rhee, J. M. Holtzman, and D. K. Kim, "*Performance Analysis of the Adaptive EXP/PF Channel Scheduler in an AMC/TDM System*," IEEE Communications Letters, vol. 8, pp. 4978-4980, Aug. 2004.
[41] J.-H. Rhee, J. M. Holtzman, and D. K. Kim, "*Scheduling of Real/Non- real Time Services: Adaptive EXP/PF Algorithm,*" in The 57th IEEE Semiannual Vehicular Technology Conference. vol. 1, 2003, pp. 462- 466.
[42] 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.
[43] Sandrasegaran, Kumbesan, M. Ramli, H. Adibah, Basukala and Riyaj, “Delay-prioritized scheduling (DPS) for real time traffic in 3GPP LTE system,” IEEE Wireless Communications and Networking Conference (WCNC), Syd. Australia, pp. 1-6, Apr. 2010.
[44] S. Sun, Q. Yu, W. Meng; C. Li, “A configurable dual-mode algorithm on delay-aware low-computation scheduling and resource allocation in LTE downlink”, Wireless Communications and Networking Conference (WCNC), 2012 IEEE, vol., no., pp.1444-1449, 1-4 Apr. 2012.
[45] G.Monghal, K.I.Pedersen, I.Z.Kovacs, and P.E.Mogensen, “*QoS oriented time and frequency domain packets schedulers for the UTRAN long term evolution*,” in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Marina Bay, Singapore, May 2008.
[46] L. Q. Zhao, Y. Qin, M. Ma, X. X. Zhong and L. Li, “QoS Guaranteed Resource Block Allocation Algorithm in LTE Downlink,” Proceedings of the 7th International ICST Conference on CHINACOM, Kun Ming, Aug. 2012, pp. 425-429.
[47] Tran S.V., and Eltawil A.M., “*Optimized scheduling algorithm for LTE downlink system*”, Proc. IEEE Wireless Communications and Networking Conference (WCNC), April 2012.
[48] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-QoS-Aware Fair Scheduling for LTE", Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd, pp. 1-5, 15-18 May 2011.
[49] Yun Li, Xin Chen, Wheliang Zao, and Bin Cao, “*Packet Scheduling with QoS support in LTE Downlink MIMO System*”, 1st IEEE International Conference on Communications in China, 2012.
[50] Kausar R, Chen Y, Chai KK (2012), “Qos aware packet scheduling with adaptive resource allocation for OFDMA based LTE-advanced networks”, IET Conference Publications vol. 2011, (586 CP) 207-212.
[51] 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.
[52] 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.
[53] S. Xergias, N. Passas, S. Sioutas, and L. Merakos, “CO-FRTS: An integrated solution for efficient Scheduling and OFDMA Slot Allocation in mobile WiMAX Networks”, International Journal of Communications Systems, WILEY (to appear).
[54] S. Xergias, N. Passas, A. Lygizou and A.K. Salkintzis, “A Multimedia Traffic Scheduler for IEEE 802.16 Ppoint-to-Multipoint Networks” accepted in the IEEE Internaional Conference on Communications (ICC) 2008, Beijing, China, May 2008.
[55] Kaporis, A., Makris, Ch., Sioutas, S., Tsakalidis, A., Tsichlas, K. and Zaroliagis, Ch. (2006) “Dynamic Interpolation Search Revisited”, ICALP 2006, Part I, LNCS 4051, 382-394.
[56] Peter van Emde Boas, R. Kaas, and E. Zijlstra: “Design and Implementation of an Efficient Priority Queue”, Mathematical Systems Theory 10: 99-127, 1977.

[57] J. Heinanen, R. Guerin, “A Single Rate Three Color Marker”, RFC2697, September 1999.
[58] Park, Kun I., “*QoS in Packet Networks*”, Boston: Springer Science + Bussiness Media, Inc., 2005.
[59] 3GPP TR 36.942 version 8.4.0 Release 8, “Evolved Universal Terrestrial Radio Access: Radio Frequency system Scenarios”, July 2012.
[60] Yacoub, M. D., Foundations of Mobile Radio Engineering, CRC Press, Boca Raton, FL, USA, 1993.
[61] 3GPP TS 26.236 v5.7.0, “Transparent end-to-end Packet-switched Streaming Service (PSS), Protocols and codecs (Release 5)”, June 2005.
[62] 3GPP TS 26.071 v5.0.0, “AMR speech Codec; General description (Release 5)”, Dec. 2002.
[63] Mohammad T. Kawser, Nafiz Imtiaz Bin Hamid, Md. Nayeemul Hasan, M. Shah Alam, and M. Musfiqur Rahman, “Downlink SNR to CQI Mapping for Different Multiple Antenna Techniques in LTE”, in International Journal of Information and Electronic Engineering, vol. 2, no. 5, September 2012.
[64] S. Xergias, PhD thesis: "Multi access protocol study and design, for sensor networks", National & Kapodistrian University of Athens, Greece, March 2010.
[65] R. Jain, W. Hawe, D. Chiu, “A Quantitative measure of fairness and discrimination for resource allocation in Shared Computer Systems”, DEC-TR-301, September 26, 1984.
[66] H. Ayoub and M. Assaad, "Scheduling in OFDMA Systems with Outdated Channel Knowledge," in *Communications (ICC), 2010 IEEE International Conference on*, 2010, pp. 1-5.