UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
Ingegneria dell'Informazione

Tesi di Dottorato di Ricerca

Performance Evaluation of LTE and LTE advanced standards for next generation mobile networks

Autore:

Matteo Maria Andreozzi *Firma*_____

Relatori:

Prof. Luciano Lenzini *Firma*_____

Ing. Giovanni Stea *Firma*_____

Anno 2012

1. ABSTRACT

The 3GPP standards LTE and LTE-Advanced for next generation mobile cellular networks are analysed.
The OptiMOS algorithm, which can be employed by the Base Station to efficiently serve VoIP connections, is described in Chapter [8].
The Relay Link scheduling algorithm, aimed to optimize LTE Advanced networks in presence of relay nodes is described in Chapter [9].
This work has been submitted in partial fulfilment of the requirements for
the Degree of Doctor of Philosophy in Information Engineering at the Information Engineering Department of the University of Pisa, Italy.

1. SOMMARIO

Nel corso della trattazione sono analizzati gli standard 3GPP LTE e LTE-Advanced per la prossima generazione delle reti mobili cellulari. L'algoritmo OptiMOS, che può essere impiegato dalla Stazione Base per servire in modo efficiente connessioni VoIP, è descritto nel capitolo [8].
L'algoritmo di link scheduling Relay, finalizzato a ottimizzare le reti LTE avanzate in presenza di nodi relay è descritto nel capitolo [9]. Questo lavoro è stato presentato in adempimento parziale dei requisiti per la Laurea di Dottore di Ricerca in Ingegneria dell'Informazione presso l'ufficio informazioni Dipartimento di Ingegneria dell'Università degli Studi di Pisa, Italia.

Index

2. INTRODUCTION

Yeah, you'll be the coolest person in the room when you pull one out and show it around, but that gets old fast when three other people have them and one person somehow has one that glows in the dark.

John C. Dvorak

The American columnist and broadcaster in article 'Rethinking the iPhone' in PC Magazine.

The early days of mobile telephony started in 1946 in St. Louis, where the Mobile Telephone Service was first introduced. Call set-up required manual operation by an operator and there were only three radio channels available for use, therefore the service was limited by having only a few voice channels per district.

In 1964 additional channels were added to the service (IMTS) and handling of calls to the public switched telephone network (PSTN) was more automated.

Later on, other technologies were introduced in so-called **pre-cellular** systems (or zero G), such as the Push to Talk (PTT) and Advanced Mobile Telephone System (AMTS).

These early mobile telephone systems could be only distinguished from closed radiotelephone systems in that they were available as a commercial service that was part of PSTN, with their own telephone numbers, rather than part of a closed network such as a police radio or taxi dispatch system.

Even tough this early services became very popular and commercial useful, they were limited to that phones could only be installed in cars and other vehicles.

It was April 3, 1973. Martin Cooper, now CEO and co-founder of ArrayCom Inc, was Motorola general manager of Communication Systems division.

He made that call. He called Cooper, his rival at AT&T Bell Labs from the streets of New York City, and the world of communications made a giant leap forward towards the nowadays worldwide interconnected network of mobile nodes.

The handheld telephony was born.

The first commercial first generation network (1G) was launched by NTT in Japan, in 1979. There were 23 radio sites (base stations) covering the whole Tokyo area, serving more than 20 million people, and base stations already supported the hand-over between sites, i.e. the ability to transfer calls between one radio site and another.

Five years later, NTT became the first operator in the World to cover a whole nation with a mobile cellular network.

In the early 80s many other nations saw the mobile communications dawn with the implementation of nation-wide 1G networks: UK, Canada, Mexico were the first ones.

Among the first international mobile communication systems started in the early 1980s, the best-known ones were NMT that was started up in the Nordic countries, AMPS in the USA, TACS in Europe, and J-TACS in Japan.
Equipment was still bulky, mainly car-borne, and voice quality was often inconsistent, with “cross-talk” between users being a common problem.

With NMT came the concept of “roaming”, providing a service for users traveling outside the area of their “home” operator.
This opened a larger market for mobile phones, attracting more companies into the mobile-communication business.

Mobile technology evolved rapidly, and in early 90s the second generation of mobile cellular networks came to life (GSM or 2G).

In Europe, the GSM (originally Groupe Spécial Mobile, later Global System for Mobile communications) was deployed, based on a project for pan-European mobile-telephony system, which was initiated in the mid 1980s by the telecommunication administrations in CEPT1 and later continued within the new European Telecommunication Standards Institute (ETSI).

The GSM standard was based on Time-Division Multiple Access (TDMA), as were the US-TDMA standard and the Japanese PDC standard that were introduced in the same time frame.

Some years later development of a Code-Division Multiple Access (CDMA) standard called IS-95 was completed in the USA in 1993.

It was the first full digital mobile cellular technology, which featured also out-of-band signaling and – most important thing – introduced the short messaging service (SMS), which had a great success among customers, and is still supported by current cellular mobile technologies. Along with that, circuit-switched data services were also introduced, enabling e-mail access and other narrowband data applications, at initial peak data rate of 9.6 kbit/s. Higher data rates were introduced later in evolved 2G systems by assigning multiple time slots to a user and through modified coding schemes.

Later on, in 1999, Docomo, Japan, was introducing mobile Internet access service for the first time in the history, with the Japanese PDC standard.
General Packet Radio Services (GPRS) was introduced at the same time in GSM for supporting packet data transmission.
These technologies are often referred to as 2.5G.

Starting from 2000, daily use of mobile phones became a worldwide fact, and demand for evolved services, such as Internet access, incessantly grew.
Along with that, users also were demanding for always-higher data speeds, having experienced the same kind of evolution with respect to fixed broadband access technology.

However, during same years, having the 2G technologies clearly reached its limits, the mobile operators and devices industry players begun to work on the third generation (3G) technology. With the advent of 3G and the higher-bandwidth radio interface of UTRA (Universal Terrestrial Radio Access) came possibilities for a range of new services that were only hinted at with 2G and 2.5G.

The 3GPP – The Third Generation Partnership Project – was established in 1998, as a collaboration among international groups of telecommunication associations, aiming to define globally applicable standards for next generation mobile networks.

Figure 1 - The 3GPP consortium

As a result of the 3GPP consortium, the 3G standard was defined (commercially known as UMTS), merging the winning WCDMA (Wideband CDMA) concepts from a European research project (FRAMES) and from the ARIB standardization in Japan.

Before 3GPP, standardization of WCDMA was continuing in parallel among several standards groups. This solved the problem of trying to maintain parallel development of aligned specifications in multiple regions.

The present organizational partners of 3GPP are ARIB (Japan), CCSA (China), ETSI (Europe), ATIS (USA), TTA (South Korea), and TTC (Japan).

The first release of WCDMA Radio Access was called release 99. This release was characterized by circuit-switched voice and video services, and data services over both packet-switched and circuit switched bearers. The first major addition of radio access features to WCDMA was HSPA, which was added in release 5 with High Speed Downlink Packet Access (HSDPA) and release 6 with High Speed Uplink Packet Access (HSUPA). These two are together referred to as HSPA.

HSPA introduced new basic functions and was aimed to achieve peak data rates of 14.0 Mbit/s. The main introduced technologies within the HSPA were the adaptive modulation QPSK and 16QAM and the High Speed Medium Access protocol (MAC-hs) in base station, which enabled it to fast scheduling execution.

Soon after the second phase of HSPA was specified in the 3GPP release 7, named HSPA Evolved. It was designed to achieve data rates of up to 42 Mbit/s, by introducing multiple antenna techniques (MIMO and Beam-Forming).

In last years, however, the number of mobile subscribers has increased tremendously and voice communication has become mobile in a massive way.
At the same time data usage has grown fast and mobile devices are now used for a wide range of other applications like web browsing, video streaming and online gaming. Beside smart phones, also notebooks, tablets and other hand-held devices are now part of the mobile environment, leading to a heterogeneous population of User Equipment (UEs) with different needs.
Within such a scenario, the demand for an ubiquitous and wide-band Internet access is constantly increasing, and mobile communication systems have to provide evolved technologies to improve support for bandwidth intensive and delay sensitive applications.
As direct evolution of HSPA Evolved, the 3GPP roadmap leaded to E-UTRA (Evolved Universal Terrestrial Radio Access), the technology specified in 3GPP Release 8.

This project is called the Long Term Evolution initiative (LTE). The first release of LTE offers data rates of over 320 Mbit/s for downlink and over 170 Mbit/s for uplink using OFDMA modulation.

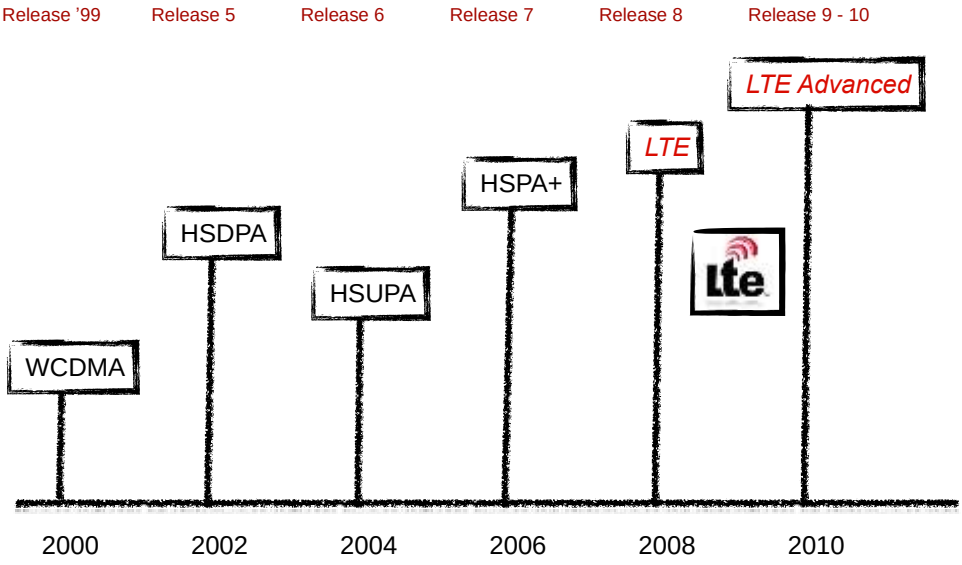

Figure 2 - 3GPP releases

3. LTE TECHNOLOGY

Long Term Evolution (LTE) is a service-optimized 4th-generation cellular technology, aiming to speed improvements up to 10-fold over existing 3G technologies was planned starting from 2004, when a workshop was organized to initiate work on the 3GPP Long-Term Evolution (LTE) radio interface.

The result of the LTE workshop was that a study item in 3GPP TSG RAN[1] was created in December 2004. The first 6 months were spent on defining the requirements, or design targets, for LTE, which were approved in June 2005.

Most notable were the requirements on high data rate at the cell edge and the importance of low delay, in addition to the normal capacity and peak data rate requirements.
Furthermore, spectrum flexibility and maximum commonality between FDD and TDD solutions were pronounced.

Being able to support the same Internet Protocol (IP)-based services in mobile devices that people use at home with a fixed broadband connection was a major challenge and a prime driver for the evolution of LTE, thus one of the main ways in which 4G differed technologically from 3G was in its elimination of circuit switching, instead employing an all-IP network.

4G ushered in a treatment of voice calls just like any other type of streaming audio media, utilizing packet switching over Internet, LAN or WAN networks via VoIP.

---

[1] *TSG Radio Access Network (TSG RAN) is responsible for the definition of the functions, requirements and interfaces of the UTRA/E-UTRA network in its two modes, FDD & TDD. More precisely: radio performance, physical layer, layer 2 and layer 3 RR specification in UTRAN/E-UTRAN; specification of the access network interfaces (Iu, Iub, Iur, S1 and X2); definition of the O&M requirements in UTRAN/E-UTRAN and conformance testing for User Equipment and Base Stations.*

LTE has been designed also in order to ensure the competitiveness of 3GPP technologies for the next years.

These improvements rely on some enabling technologies which were not considered in the previous releases of the UMTS technologies, i.e. the adoption of Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier Orthogonal Frequency Division Multiple Access (SC-OFDMA) as downlink and uplink access scheme are an essential part of the LTE standard, as well as the inclusion of complex Multiple Input Multiple Output (MIMO) antenna schemes.

The major improvements LTE accomplished with respect to previous 3GPP releases are:

- **Data rate**: theoretical achievable peak data rate (measured at physical layer) of 300 Mbps in downlink and 75 Mbps in uplink, given the standard configuration of full spectrum (20 MHz) bandwidth.

- **Spectrum flexibility**: spectrum bandwidth scalable from 1.25 up to 20 MHz in order to support different deployment requirements.

- **Architecture simplification**: the number of elements composing the UMTS Terrestrial Radio Access Network (UTRAN) structure has been reduced, converging towards a flat architecture.

- **Enhanced support for mobility**:
  - System performances optimized for slow-moving users (0 - 15 km/h)
  - High performances still achieved for users moving at speeds between 15 - 150 km/h
  - Minimum quality of experienced services still guaranteed at very high speed (up to 350 km/h).

– **Reduced latency**: the one-way transit time between a packet being available at the IP layer in either the UE or radio access network and the availability of this packet on the counterpart shall be 5ms in normal operating conditions.
Also Control Plane latency is reduced, being attested at less than 100ms.

– **Enhanced support for end-to-end Quality of Service (QoS)**: an all-IP paradigm is used to deal with different traffic flows. Voice traffic is served as VoIP (Voice over IP) and the call quality should be at least the same as in UMTS circuit switched networks, measured as ITU-MOS score ([17],[18]).

Figure 3 - LTE will fully support VoIP technology with the highest QoS w.r.t. current technologies

– **Inter-working**: inter-working with existing legacy 3GPP systems and non-3GPP systems is ensured, with handover added delay time between 300 and 500ms.

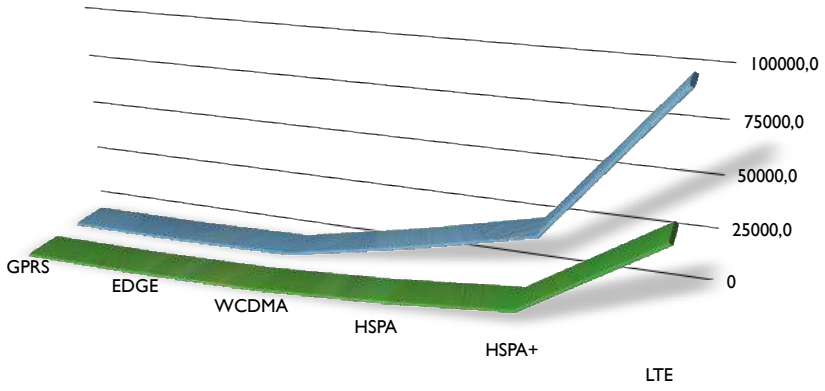

Figure 4 - LTE data rate improvement

4. LTE ARCHITECTURE

The LTE has been designed to be a full packet-switched services technology (thus dropping circuit-switched services of the previous cellular generations).
Full IP connectivity is provided between the UE and the Packet Data Network (PDN).

Four high-level logical domains compose the LTE network architecture [Figure 5]:

- **Services domain**
- **EPC** (Evolved Packet Core)
- **E-UTRAN** (Evolved UMTS Terrestrial Radio Access Network)
- **User Equipment domain**

Figure 5 - LTE E-UTRAN and EPC

The Evolved Packet Core (EPC) is a part of the so-called System Architecture Evolution (SAE), which comprehends all the non-radio access part of the LTE system (opposed to the E-UTRAN level). EPC is shown in detail in Figure 6.

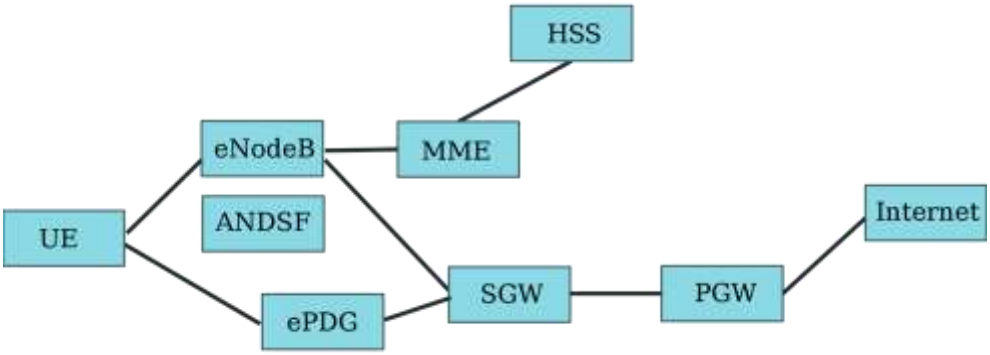

Figure 6 - Evolved Packet Core

The EPC evolves from previous core network structure showing a full separation of *Control Plane* and *User Plane* functionalities, embodied by two different network components: the Mobility Management Entity (MME) and the Serving Gateway (SGW) respectively. These two entities communicate using the S11 interface as shown in the above picture. Full domain separation provides the system with a dedicated node (S-GW) for high bandwidth packet processing and a control node (MME), which is responsible of all signalling transactions.

The MME is responsible for idle mode UE tracking and paging procedure including retransmissions. It is involved in the bearer activation/deactivation process and is also responsible for choosing the SGW for a UE at the initial attach and at time of intra-LTE handover involving Core Network node relocation. It is responsible for authenticating the user. The Non Access Stratum (NAS) signalling terminates at the MME and it is also responsible for generation and allocation of temporary identities to UEs. It checks the authorization of the UE to camp on the service provider's Public Land Mobile Network (PLMN) and enforces UE roaming restrictions. The MME is the termination point in the network for ciphering/integrity protection for NAS signalling and handles the security key management. Lawful interception of signalling is also supported by the MME. The MME finally provides

the control plane function for mobility between LTE and 2G/3G access networks with the S3 interface terminating at the MME from the SGSN. Here follows a list of all MME features:

- Idle mode UE tracking and paging procedure including retransmissions;
- Bearer activation/deactivation process;
- SGW selection at the initial attach and at time of intra-LTE handover;
- Mobility between LTE and 2G/3G access networks;
- UE authentication;
- Generation and allocation of temporary identities to UEs;
- Authorization checks of the UE to camp on the service provider's PLMN;
- UE roaming restrictions enforcing.
- Ciphering and integrity protection for NAS signalling and security key management;
- Lawful interception.

The SGW routes and forwards user data packets, while also acting as the mobility anchor for the user plane during inter-eNodeB handovers and as the anchor for mobility between LTE and other 3GPP technologies (terminating S4 interface and relaying the traffic between 2G/3G systems and PDN Gateway **-** PGW). For idle state UEs, the SGW terminates the downlink data path and triggers paging when downlink data arrives for the UE. It manages and stores UE contexts, e.g. parameters of the IP bearer service, network internal routing information. It also performs replication of the user traffic in case of lawful interception.
Its main functions are listed below:

- The anchor role for inter-eNodeB handovers and for inter-3GPP mobility;
- Routing and forwarding of user data packets;
- Traffic relaying among PGWs;
- Replication of the user traffic in case of lawful interception;

- Downlink packets buffering;
- Paging triggering upon UE downlink data arrival;
- Handling UE context (e.g. parameters of the IP bearer service, network internal routing information, etc.)
- Downlink packet marking (e.g. by marking DiffServ field in IP packets making use of QCI field).

The PDN Gateway (PGW) is also part of the EPC and provides connectivity from the UE to external packet data networks by being the point of exit and entry of traffic for the UE. A UE may have simultaneous connectivity with more than one PGW for accessing multiple Packet Data Networks (PDNs). The PGW performs policy enforcement, packet filtering for each user, charging support, lawful interception and packet screening. Another key role of the PGW is to act as the anchor for mobility between 3GPP and non-3GPP technologies such as WiMAX and 3GPP2 (CDMA 1X and EvDO). Its features, detailed point by point, are:

- The anchor role for mobility between 3GPP and non-3GPP technologies such as WiMAX and 3GPP2 (CDMA 1X EvDO);
- Policy enforcement;
- Packet filtering for each user;
- Charging support;
- Lawful interception;
- Packet screening.

Other three nodes complete the EPC structure, each one providing specific control functions:

1. HSS (Home Subscriber Server): The HSS is a central database that contains user-related and subscription-related information. The functions of the HSS include functionalities such as mobility management, call and session establishment support, user authentication and access authorization. The HSS

is based on pre-Release-4 Home Location Register (HLR) and Authentication Centre (AuC).

2. ANDSF (Access Network Discovery and Selection Function): The ANDSF provides information to the UE about connectivity to 3GPP and non-3GPP access networks (such as Wi-Fi). The purpose of the ANDSF is to assist the UE to discover the access networks in their vicinity and to provide rules (policies) to prioritize and manage connections to these networks.
3. EPDG (Evolved Packet Data Gateway): The main function of the ePDG is to secure the data transmission with a UE connected to the EPC over an untrusted non-3GPP access. For this purpose, the ePDG acts as a termination node of IPSec tunnels established with the UE.

The EPC and E UTRAN together form the Evolved Packet System (EPS), shown in [Figure 7]

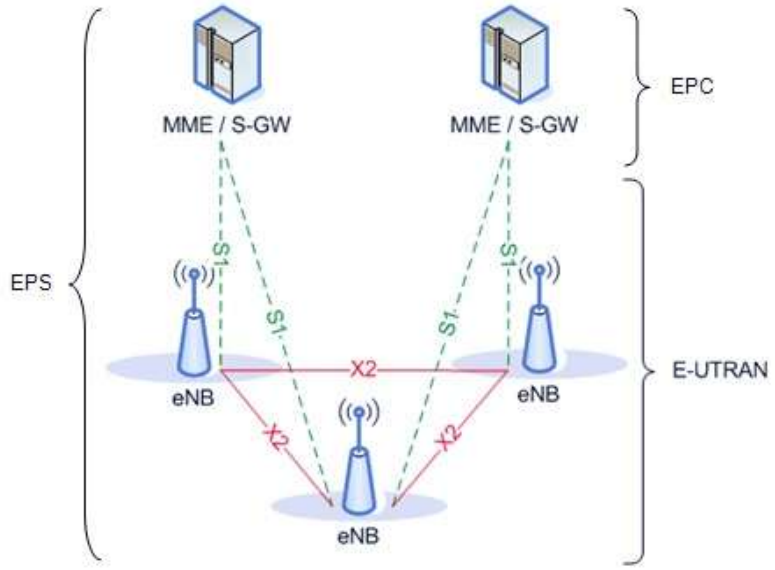

Figure 7 - Evolved Packet System

The E-UTRAN is the radio access network of LTE (air interface) and consists of eNodeB, which are an evolution of the UTRAN NodeB, which are interconnected by X2 interfaces.

An eNodeB provides the U-Plane (PDCP/RLC/MAC/PHY) and C-Plane (RRC) and acts as termination towards the UE.
The eNodeB is responsible for Radio Resource Management (RRM), i.e. it controls and coordinates the Radio Bearer Control (RBC), Radio Admission Control (RAC), Connection Mobility Control (CMC), and dynamic allocation of resources to UEs in uplink and downlink directions.
The eNodeB is also the node responsible of establishing C-Plane connectivity towards a specific MME (MME selection function), and of routing of U-Plane data towards SGW.

4.1.1. E-UTRAN PROTOCOL STACK

In figure [Figure 8] are shown the protocol stacks of U-Plane and C-Plane.

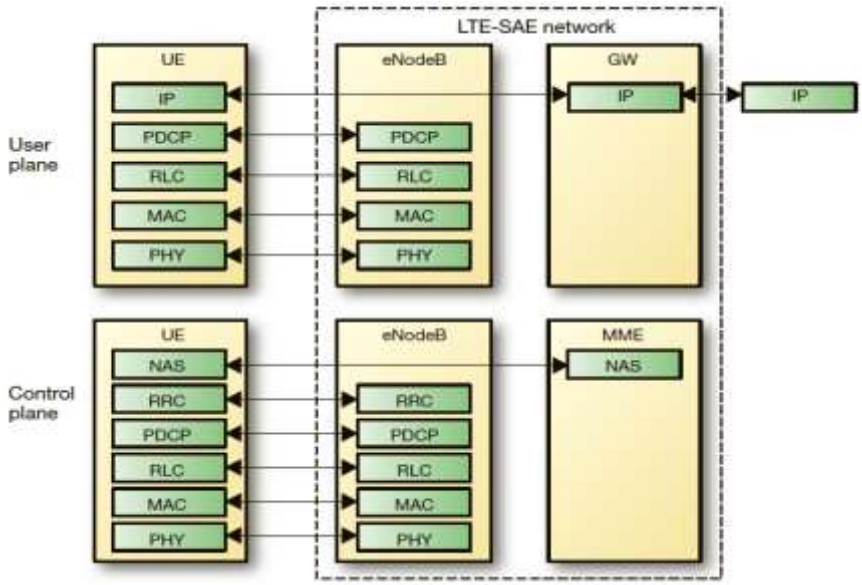

Figure 8 - User Plane and Control Plane

4.1.2. U-Plane

The protocol stack for the U-Plane consists of:

Packet Data Convergence Protocol (PDCP): provides transport of the IP packets, with ROHC header compression, ciphering, and depending on the RLC mode in-sequence delivery, duplicate detection and retransmission of its own SDUs during handover**.** Its features are:

- Header compression and decompression (i.e. ROHC[2]);
- Transfer of user data;
- In-sequence delivery of upper layer PDUs (i.e. IP datagrams);
- Duplicate detection of lower layer SDUs (i.e. RLC SDUs);
- Retransmission of PDCP SDUs at handover;
- Ciphering and deciphering;
- Timer-based SDUs discarding in uplink.

[Figure 9] represents the PDCP PDU Structure.

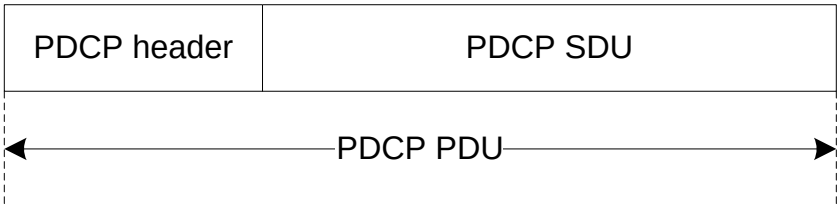

Figure 9 - PDCP PDU

[2] *When header is as big as the data being transmitted, it generates overhead wasting the precious air resource. For that, header compression is generally performed. A compressor is used before the data is sent and a de-compressor at the receiving end adds the uncompressed headers back to the received packets. Both the ends must use the same protocol (i.e. ROHC).*

PDCP header can be either 1 or 2 bytes long and both PDCP PDU and PDCP header are octet-aligned.

Radio Link Control (RLC): it transports the PDCP's PDUs. It can work in 3 different modes depending on the reliability provided: RLC Acknowledged Mode, Unacknowledged Mode, and Transparent Mode. Depending on this mode it can provide: ARQ error correction (AM only), segmentation/concatenation of PDUs, reordering for in-sequence delivery, duplicate detection (AM and UM). Its functions are, in detail:

- Transfer of upper layer PDUs (i.e. PDCP PDUs);
- Error correction through ARQ[3] (for AM data transfer);
- Concatenation, segmentation and reassembly of RLC SDUs (for both UM and AM data transfer);
- Re-segmentation of RLC data PDUs for AM data transfer;
- In-sequence delivery of upper layer PDUs (i.e. PDCP PDUs) (for both UM and AM data transfer);
- Duplicate detection (for both UM and AM data transfer);
- Protocol error detection and recovery;
- RLC SDUs discarding (for both UM and AM data transfer);
- RLC re-establishment.

[Figure 10] shows the RLC PDU Structure.

[3] *The ARQ functionality provides error correction by implementing a retransmission protocol in RLC AM at layer 2. Reception of a negative RLC status message will trigger a retransmission of the corresponding RLC PDU(s).*

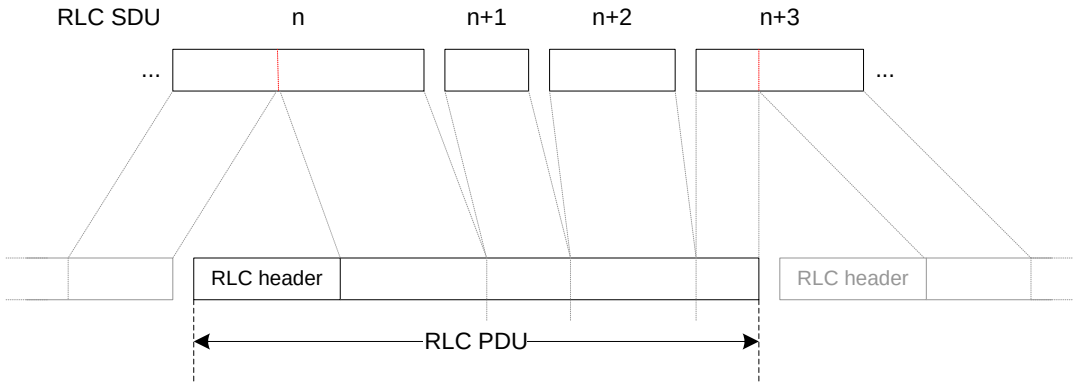

Figure 10 - RLC PDU

Red lines indicate the occurrence of segmentation. Segmentation only occurs when needed and concatenation is done in sequence. PDU sequence number carried by the RLC header is independent of the SDU sequence number (i.e. PDCP sequence number).

Media Access Control (MAC): the MAC layer offers to the RLC layer a set of logical channels, which are multiplexed in physical layer transport channels. It manages the HARQ error correction, handles the prioritization of the logical channels for the same UE and the dynamic scheduling between UEs.

Its main roles are:

- Channel mapping (between logical channels and transport channels);
- Multiplexing/de-multiplexing of MAC SDUs (belonging to one or different logical channels into/from transport channels);
- Delivering of TBs[4] to/from the PHY layer (on transport channels);
- Scheduling information reporting;
- Error correction through HARQ[5];

---

[4] Each MAC PDU is mapped 1 to 1 onto physical transport block (TB).

[5] The HARQ functionality ensures delivery between peer entities at layer 1. It is an N-channel stop-and-wait protocol with asynchronous downlink retransmissions and synchronous uplink retransmissions. If RLC AM is enabled, the two-layer ARQ design (ARQ+HARQ) achieves low latency and low overhead without sacrificing reliability. Most of the errors are however captured and corrected by the HARQ

- Priority handling (between logical channels of one UE);
- Dynamic scheduling;
- MBMS service identification;
- Transport format selection;
- Padding.

All data flowing to and from the MAC layer are structured in MAC Packet Data Units (PDU). Figure 11 illustrates the MAC PDU Structure.

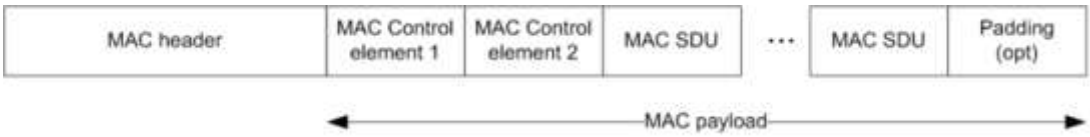

Figure 11 - MAC PDU

A MAC PDU consists of:

- *MAC header.* It is composed of one or more MAC PDU sub-headers; each sub-header refers to a MAC Service Data Unit (SDU), a MAC Control Element (CE) or indicates presence of padding. MAC header has therefore a variable size.
- *MAC CEs.* There are different kinds of MAC CEs:
    - *Buffer Status Report (BSR).* BSR are used by UEs to advertise their transmission buffer sizes.
    - *C-RNTI.* As the C-RNTI may be used on common transport channels for optimization purposes, this MAC CE is used to specify it.
    - *DRX Command.* As UEs enter IDLE mode when there is no activity, they use DRX in IDLE mode in order to wake up periodically and check for paging messages.

---

protocol and only residual errors are detected and resolved by the more expensive (in terms of latency and overhead) ARQ retransmissions.

- *UE Contention Resolution Identity.* Contention Resolution procedure is based on either C-RNTI on PDCCH or UE Contention Resolution Identity on DL-SCH. In order to specify the UE Contention Resolution Identity this MAC CE is used.
- *Timing Advance Command.* This MAC CE specify a value used to control the amount of timing adjustment that UE has to apply in order to adjust its uplink transmission timing for PUCCH/PUSCH/SRS.

- *MAC SDUs:* MAC SDUs are always placed after any MAC CE and have variable sizes.
- *Optional padding.* Data that UE ignores used for formatting the PDU in precise standardized size formats (TBS – Transport Block Size).

Physical layer (PHY): PHY offers data transport services to higher layers. Carries all information from the MAC transport channels over the air interface. Takes care of the link adaptation (AMC), power control, cell search (for initial synchronization and handover purposes) and other measurements (inside the LTE system and between systems) for the RRC layer. The access to these services is through the use of "transport channels" via the MAC sub-layer. It performs:

- Error detection (on the transport channel and indication to higher layers);
- FEC encoding/decoding (of the transport channel);
- HARQ soft-combining;
- Rate matching (of the coded transport channel to physical channels);
- Channel mapping (mapping of coded transport channel onto physical channels);
- Power weighting (of physical channels);
- Modulation and demodulation (of physical channels);
- Frequency and time synchronization;
- Radio characteristics measurements (and indication to higher layers);
- MIMO antenna processing;

- Transmit Diversity;
- Beam-forming;
- RF processing.

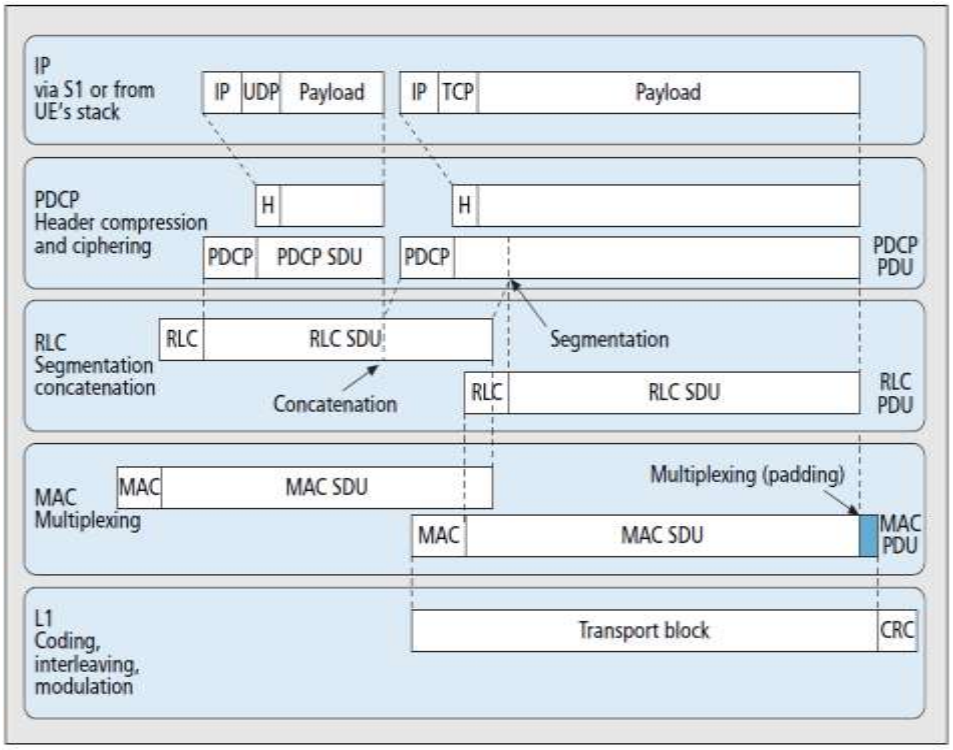

Figure 12 - Data flow for U-Plane protocol stack

PHY layer packets are the Transport Blocks (TB). A maximum of one MAC PDU can be transmitted per TB per UE[6]. Figure 12] shows the data flow through the whole U-Plane protocol stack.

[6] *In both uplink and downlink, only one TB is generally created in a TTI. However, up to 2 TBs may be generated in a TTI if MIMO is enabled.*

4.1.3. C-Plane

The protocol stack for the C-Plane consists of:

- ***PDCP.*** For the RRC layer it provides transport of control plane and performs ciphering and integrity protection.
- ***RLC, MAC and PHY.*** These layers are responsible for providing transport support of RRC control data.
- ***Radio Resource Control (RRC).*** It takes care of: the broadcasted system information related to the access stratum and transport of the non-access stratum (NAS) messages, paging, establishment and release of the RRC connection, security key management, handover, UE measurements related to inter-system (inter-RAT) mobility, QoS establishment. Its features are listed below:
    - Broadcast of system information related to the NAS and AS;
    - Paging;
    - Establishment, maintenance and release of an RRC connection between the UE and E-UTRAN including allocation of temporary identifiers between UE and E-UTRAN and configuration of signaling radio bearers for RRC connection;
    - Security functions including key management;
    - Establishment, configuration, maintenance and release of point-to-point Radio Bearers;
    - Mobility functions including UE measurement reporting and control of the reporting for inter-cell and inter-RAT mobility, handover, UE cell selection and reselection and control of cell selection and reselection, context transfer at handover;
    - Notification for MBMS services;
    - Establishment, configuration, maintenance and release of RBs for MBMS services;
    - QoS management functions;
    - UE measurement reporting and control of the reporting;

    - NAS direct message transfer to/from NAS from/to UE.

- ***Non-Access Stratum (NAS).*** NAS performs additional control functions:
    - EPS bearer management;
    - Authentication;
    - Mobility handling;
    - Paging origination;
    - Security control.

RRC and NAS layers top each UE C-Plane. All NAS control messages are exchanged between UE and MME: eNB silently hands over all NAS messages to MME.

5. LTE CHANNELS MAPPING

5.1.1. Layer 2 channel mapping

Layer 2 is split into previously discussed PDCP, RLC and MAC layers. Figure 13 illustrates the downlink structure of layer 2, while Figure 14 shows the uplink one.

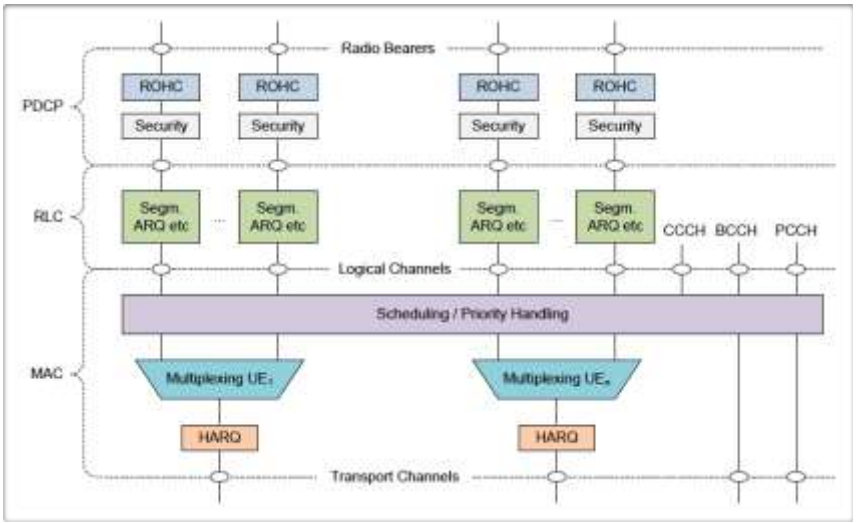

Figure 13 - Downlink Layer 2 structure

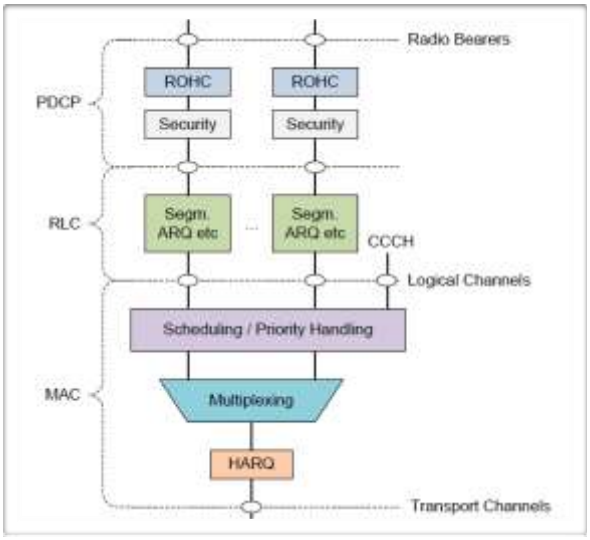

Figure 14 - Uplink Layer 2 structure

SAPs are marked with circles at the interface between sub-layers:

- SAPs between the PDCP sub-layer and the upper layer offer connection to the radio bearers;
- SAPs between the MAC sub-layer and the RLC sub-layer are the connection point for logical channels;
- SAPs between the physical layer and the MAC sub-layer are dedicated to transport channels.

Radio Bearers (RBs) are logical channel over the radio interface and are established using RRC protocol. There are two main types of radio bearers:

- ***Signalling Radio Bearer (SRB).*** SRBs are defined as RBs dedicated to transmission of RRC and NAS messages.
- ***Data Radio Bearer (DRB).*** The DRBs are used to carry user data (i.e. IP datagrams). Each DRB is associated with an EPS Bearer[7].

Part of the bearer establishment procedure is authentication and activation of encryption. The required data for this process is retrieved by the eNodeB from the MME. The MME also delivers all necessary information that is required to configure the data radio bearer, like minimum and maximum required bandwidth, maximum tolerable delay and other QoS parameters.
MAC layer offers different service modes for data transfer. Each logical channel type defines the type of information transferred on it. A general classification of logical channels is listed below:

- ***Control channels.*** Control channels are employed for transferring C-Plane information only. Control channels available at MAC layer are:

[7] *Each EPS Bearer is associated with a Traffic Flow Template (TFT). A TFT is a set of packet filters. Packet filtering uses IP header information like source address, destination address, ToS field, TCP port, etc. Because of that each TFT is mapped on a given QoS and all the data that mapped to this TFT receives the same QoS treatment.*

- ***Broadcast Control Channel (BCCH)***, a downlink channel used to transmit broadcast system control information;
- ***Paging Control Channel (PCCH)***, a downlink channel used to transmit paging information and system information change notifications
- ***Common Control Channel (CCCH)***, a channel used to transmit control information between UEs and network (used for UEs having no RRC connection with the network – RRC idle mode);
- ***Multicast Control Channel (MCCH)***, a point-to-multipoint downlink channel used to transmit MBMS control information from the network to the UE, for one or several MTCHs (this channel is only used by UEs that receive MBMS);
- ***Dedicated Control Channel (DCCH)***, a point-to-point bi-directional channel that transmits dedicated control information between a UE and the network (this channel is used by UEs having an RRC connection – RRC connected mode).

- ***Traffic channels.*** Traffic channels are dedicated to transmission of U-Plane information. Traffic channels available at MAC layer are:

  - ***Dedicated Traffic Channel (DTCH)***, point-to-point channel, dedicated to single UE, for transmission of user information (a DTCH can exist in both uplink and downlink).
  - ***Multicast Traffic Channel (MTCH)***, a point-to-multipoint downlink channel carrying data from the network to the UE (this channel is only used by UEs that receive MBMS).

MAC layer is transparent for BCCH, CCCH and PCCH (i.e. it forwards their content without any processing).

In order reduce complexity of LTE protocol architecture; the number of transport channels has been reduced with respect to previous 3GPP releases. Transport channels abstract at PHY layer different kinds of data transfer services: each

transport channel type defines the type of information transferred on it. Downlink transport channels are:

- ***Broadcast Channel (BCH)***;
- ***Multicast Channel (MCH)***;
- ***Downlink Shared Channel (DL-SCH)***;
- ***Paging Channel (PCH)***.

Uplink transport channels are:

- ***Uplink Shared Channel (UL-SCH)***;
- ***Random Access Channel (RACH)***.

Radio bearers are mapped 1 to 1 on logical channels, while logical channels are mapped on transport channels following the rules shown in [Figure 15] and [Figure 16].

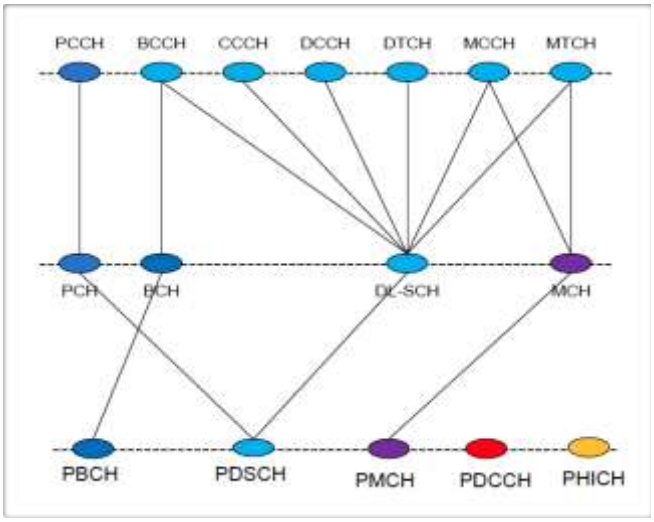

Figure 15 - Downlink logical channels to transport channels mapping

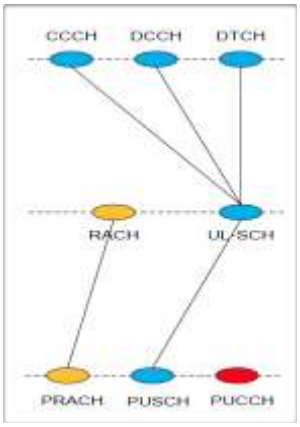

Figure 16 - Uplink logical channels to transport channels mapping

5.1.2. Layer 1 channel mapping

PHY layer provides as well channels enabling different data services to be transferred on separated channels. Transport channels are mapped on physical channels by PHY layer, which carry information originating from higher layers (i.e. MAC layer). A physical layer is identified also by a specific set of Resource Elements (REs)[8]. Downlink physical channels are:

- ***Physical Downlink Shared Channel (PDSCH).*** PDSCH purpose is enabling user data transport. It has been designed in order to deliver very high data rates, by supporting all LTE available modulations - QPSK, 16QAM and 64QAM.
- ***Physical Broadcast Channel (PBCH).*** PBCH carries system information for accessing the network. Only QPSK modulation is available on this channel.
- ***Physical Multicast Channel (PMCH).*** PMCH carries system information for multicast. Uses QPSK modulation.

[8] *A Resource Element (RE) is a unit of allocation in the time and frequency domains. One RE corresponds to one subcarrier on frequency domain and one OFDM symbol on time domain.*

- ***Physical Downlink Control Channel (PDCCH).*** PDCCH conveys UE-specific control information (e.g. scheduling information). Encodes its data by using QPSK modulation only.
- ***Physical HARQ Indicator Channel (PHICH).*** PHICH is dedicated to HARQ signalling.
- ***Physical Control Format Indicator Channel (PCFICH).*** PCFICH carry required information for decoding PDSCH.

Uplink physical channels are:

- ***Physical Uplink Shared Channel (PUSCH).*** PUSCH is used for unicast transmission and paging. QPSK, 16QAM or 64QAM modulations are available.
- ***Physical Uplink Control Channel (PUCCH).*** PUCCH carries uplink control information. It is never transmitted simultaneously with PUSCH data. PUCCH conveys control information about channel quality, HARQ ACKs and NACKs, uplink-scheduling requests.
- ***Physical Random Access Channel (PRACH).*** PRACH is used for random access procedures[9].

Transport channels are mapped on physical channels following the rules shown in figures [Figure 16] and [Figure 17].

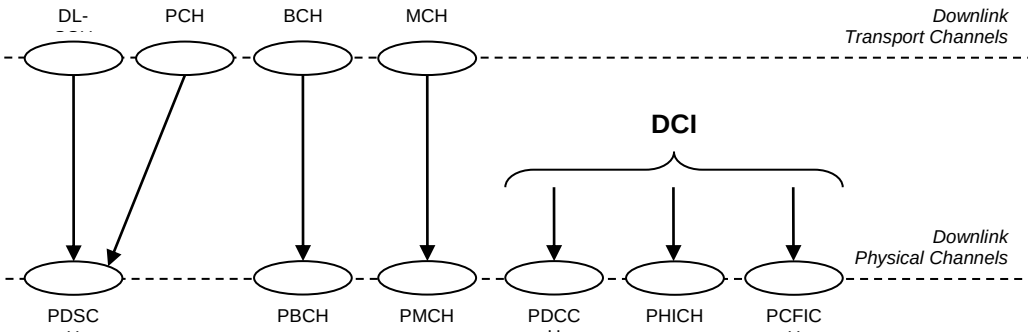

Figure 17 - Downlink transport channels to physical channels mapping

[9] *Random access is the process by which a terminal requests a connection setup.*

Channels marked in the figure as DCI are responsible for carrying downlink control information. Such information are HARQ feedback, scheduling grants for resource assignment (supporting dynamic scheduling), modulation and coding scheme indications, power control commands, feedback reporting requests.

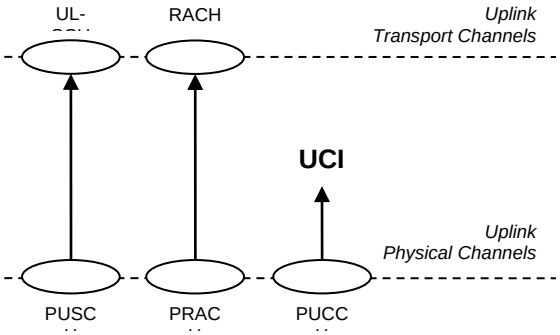

Figure 18 - Uplink transport channels to physical channels mapping

Some uplink physical channels as well are enabled to transmit Uplink Control Information (UCI). This information consist of HARQ feedback, scheduling requests, feedback reports.

6. PHYSICAL LAYER

Cellular systems previous to LTE have been characterized by the use of single carrier modulation schemes while LTE adopts the OFDMA scheme, which can be view as an array of multiple single carrier systems.

In the following the multipath fading phenomena is illustrated for single carrier systems, and it will be shown why such systems wouldn't be feasible for delivering LTE's data rate, thus leading to LTE use of ODFMA technology as radio access one.

6.1.1. Multipath Fading

Multipath fading is the propagation phenomenon consisting in two or more radio signal paths propagating between the TX antenna and the RX antenna. In a multipath environment, signals can experience different delays traveling along different paths. Because of this, the receiver gets duplicates of the original signal; each of them at a different time, with delay spreads which can reach several microseconds.

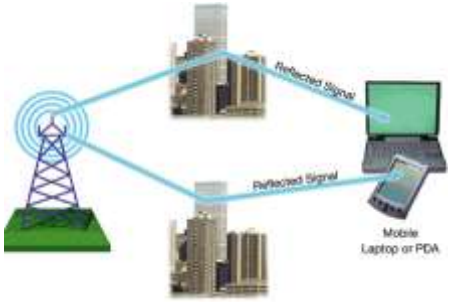

Figure 19 - Multipath example

The delay induced by multipath fading can cause a symbol, received along a delayed path, to "overlap" into a subsequent symbol arriving to the receiver via a different path. This effect is referred to as Inter-Symbol Interference (ISI), as shown in Figure 19.

In a conventional single carrier system the symbol time T decreases as data rates R increases.

$$R = \frac{1}{T}$$

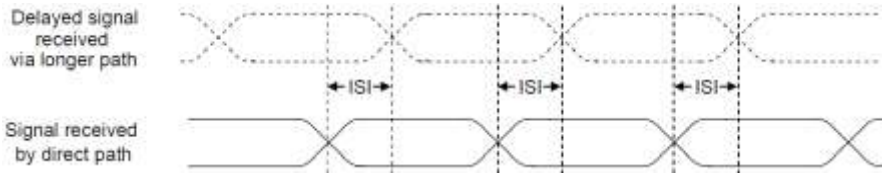

Figure 20 - Inter Symbolic Interference

Because of this effect, at very high data rates correspond shorter symbol periods and it is quite possible for ISI to exceed an entire symbol period and spill into a second or third subsequent symbol.

Multipath produces distortion in frequency domain too. Each path of different length and reflection will result in a specific phase shift. When all signal duplicates are combined at the receiver, some of those will generate a “constructive interference” (linear combination of signals in-phase), while others will lead to a “destructive interference” (linear combination of signals out-of-phase).

The composite received signal is thus distorted by frequency selective fading.

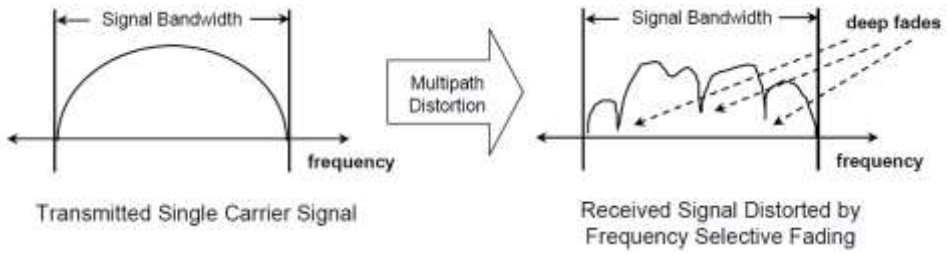

Figure 21 - Frequency selective fading

Using time domain equalization can usually compensate time-domain distortions introduced by multipath. Time domain equalizers compensate for multipath induced distortion using one of following two methods:

- **Channel inversion**. An a-priori known sequence is transmitted over the channel before sending data. Since the sequence is also known at

the receiver, the equalizer is able to infer channel response and multiply the subsequent data-bearing signal (unknown) by the inverse of the channel response matrix in order to cancel the effects of multipath.

- **Rake equalizers**. CDMA systems can employ rake equalizers. Rake equalizers are capable of separating received signal into multiple individual paths and then combine those copies of the received signal enhancing it (i.e. exploiting constructive interference).

Each of previous described channel equalizer implementations become increasingly complex as data rate increases; this is due to ISI getting much more severe and possibly spanning several symbol periods.
As symbol time becomes shorter, receiver sample clock must become correspondingly faster.
Willing to make use of a single carrier system, LTE data rates (up to 100 Mbps) and corresponding delay spreads (approaching 17 μ-sec), would make channel equalization for ISI cancellation impractical.

Due to this, OFDMA has been chosen as transmission technology for LTE and LTE advanced systems, which, as described in the following, does not suffer of ISI increasing along with data rate.

6.1.2. Orthogonal Frequency-Division Multiplexing

Orthogonal Frequency-Division Multiplexing (OFDM) systems do not rely on increased symbol rates in order to achieve higher data rates, thus complexity of managing ISI is kept low. OFDM systems split the available bandwidth into many subcarriers and transmit on each one parallel data streams. Those subcarriers are independently modulated using different levels of QAM modulation (e.g. QPSK, QAM, 64QAM or even higher orders in beyond LTE systems).

An OFDM symbol is a linear combination of the instantaneous signals transmitted on each subcarrier.

Since data is transmitted in parallel rather than serially, OFDM symbols need to last longer in time than symbols on single carrier systems of equivalent data rate. This is due to each OFDM symbol needing to be preceded by a cyclic prefix (CP), which is appended as guard interval, in order to mitigate ISI effects. Figure 21 shows a representation of an OFDM signal.

In order to understand how OFDM deals with ISI induced by multipath, consider the time domain representation of an OFDM symbol as shown in [Figure 23].

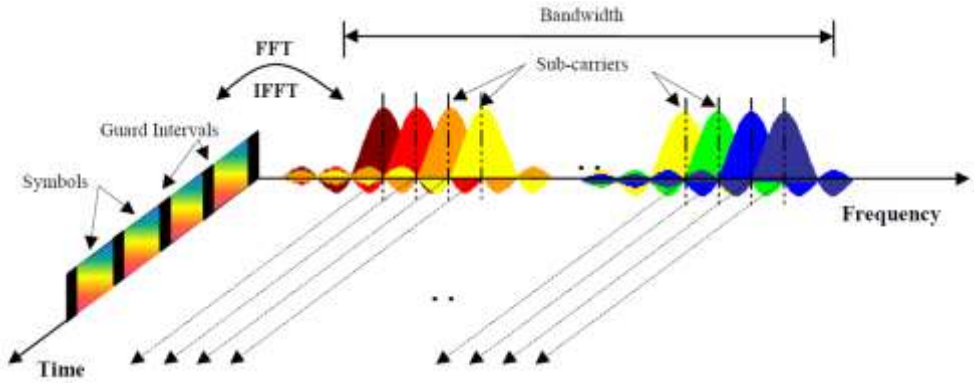

Figure 22 - OFDM signal

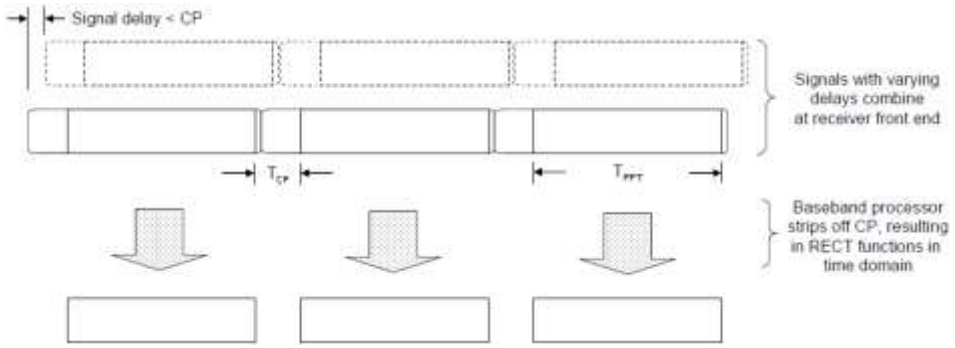

Figure 23 - Time domain representation of an OFDM symbol

One OFDM symbol is formed by two major components: Cyclic Prefix (CP) and Fast Fourier Transform (FFT) period (T_{CP} and T_{FFT} respectively). If T_{CP} is long enough, the preceding symbol does not overlap with the subsequent symbol FFT period and no ISI is produced at all. When the signal is received and digitized, the receiver can discard the CP, and recover a rectangular pulse signal for each one of the subcarriers. The recovered uniform rectangular pulse (RECT function) in the time domain corresponds to a SINC function ($sinc(t) = sin(t)/t$) in the frequency domain. This leads to the whole signal being received as a SINC pattern in the frequency domain with uniformly spaced zero-crossings at Δf intervals. By choosing opportune subcarrier spacing is possible to eliminate also interference among different sub-carriers (Inter-Carrier Interference - ICI). It is sufficient to choose as carrier spacing:

$$\Delta f = \frac{1}{T_{FFT}}$$

This choice makes efficient the use of available bandwidth and virtually eliminates ICI. This is because of zero-crossings being positioned precisely between adjacent subcarriers (thus at SINCs maximum points) as shown in [Figure 24].

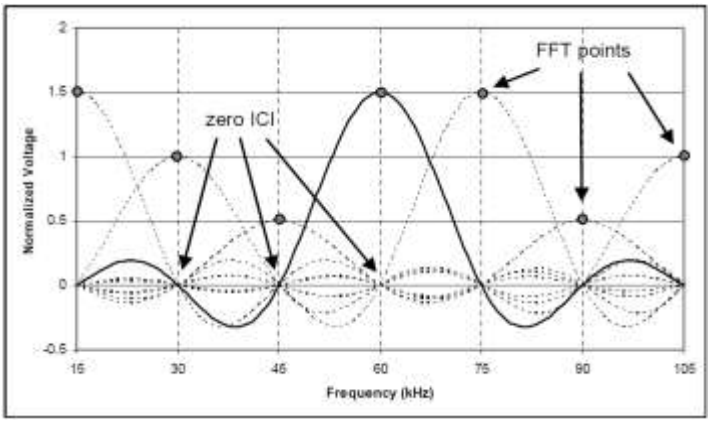

Figure 24 - Zero level ICI

By doing this, it is possible to sample at the centre frequency of each subcarrier while encountering no interference from neighbouring subcarriers (zero-ICI).

In order to achieve this result it is just sufficient that the OFDM signal has to be generated by transmitter (Figure 25) using the IFFT (Inverse Fast Fourier Transform) digital signal processing. The IFFT converts a number N of subcarrier modulated data symbols $[x_0, \dots, x_{N-1}]$ into a time domain signal $[s_0, \dots, s_{N-1}]$, resulting from the time superposition of N modulated subcarriers.

Therefore a waveform composed of N orthogonal subcarriers is obtained from a parallel stream of N sources of data. Each of those subcarriers is independently modulated, and each one shaped as a frequency-domain SINC function.

The N-point time domain blocks obtained from the IFFT are finally serialized in order to generate a unique time domain signal.

Cyclic prefix insertion is performed between each of those blocks.

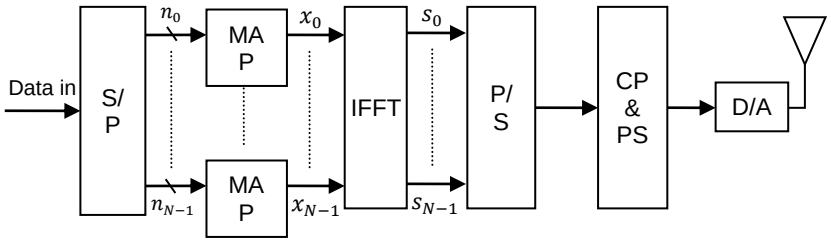

Figure 25 - OFDM transmitter

At receiver side (Figure 26), the transmitted time-sampled OFDM signal is converted back into frequency domain by using a FFT (Fast Fourier Transform) digital signal processor.

The FFT is performed at baseband frequency, so the received signal has to be down-converted from the RF carrier frequency before entering the FFT DSP.

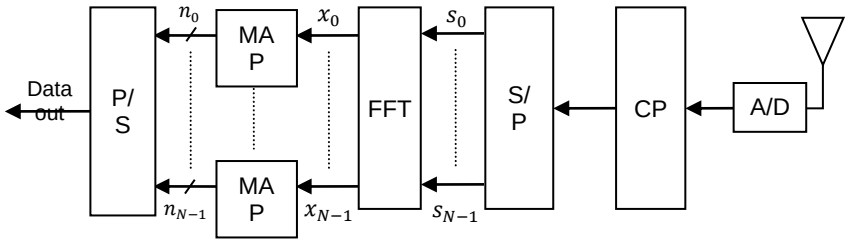

Figure 26 - OFDM receiver

Figure 2.27: OFDM signals receiver.

This down-conversion operation requires synchronization between carrier signal and receiver local oscillator frequency.

Oscillator phase noise and Doppler effect can alter this synchronization, thus resulting in a non-zero-ICI as shown in Figure 28.

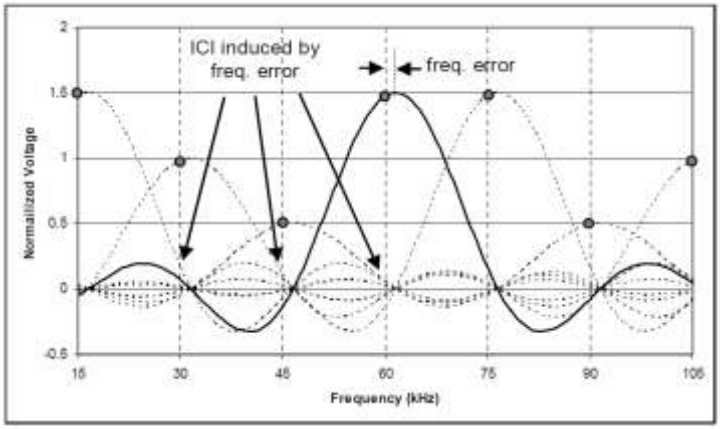

Figure 28 - Non zero ICI due to frequency errors

Due to this issue, the signal frequency must be tracked continuously and signal offsets must be tuned in order to avoid excessive ICI that might lead to extensive data loss.

Another drawback of OFDM transmission technique is its high signal Peak-to-Average Power Ratio (PAPR).

Showing an high PAPR bear some disadvantages:

- Wide dynamic range is required for A/D and D/A converters (expensive);
- Efficiency of transmitter RF Power Amplifier (RFPA) is reduced (higher power consumption).

There exist some fine-tunings that can reduce PAPR, but still OFDM power efficiency always remains lower than single-carrier constant envelope systems.

In Table 1, the main advantages and disadvantages of OFDM are summarized.

Advantages	Disadvantages
Makes efficient use of the spectrum by allowing overlap.	*The OFDM signal requires RFPAs with a high PAPR.*
By dividing the channel into flat fading subcarriers, OFDM is more resistant to frequency selective fading than single	*It is more sensitive to carrier frequency errors.*

carrier systems.	
Eliminates ISI and ICI through use of a cyclic prefix and frequency synchronization.	
Using adequate channel coding and interleaving one can recover symbols lost due to the frequency selectivity of the channel.	
Channel equalization is simpler respect to single carrier systems.	
OFDM is computationally efficient by using FFT techniques.	

Table 1 - OFDM advantages and disadvantages

6.1.3. Orthogonal Frequency-Division Multiple Access

Orthogonal Frequency-Division Multiple Access (OFDMA) allows time and frequency concurrent access of multiple users on same bandwidth: this is achieved by assigning a specific time-frequency resource to each of the OFDMA users.

OFDMA is highly efficient for broadband wireless networks, due to its major advantages with respect to other techniques including feasible scalability and application of MIMO techniques, and capacity of exploiting channel frequency selectivity.

OFDMA can be view as a combination of both frequency domain and time domain multiple access techniques, in which resources are partitioned in a time-frequency space, composed by slots which are an intersection of OFDM symbols and subcarriers.

Adaptive OFDMA slots assignment can be achieved, by exploiting feedback information about receivers channel conditions.

If subcarrier assignment is executed at reasonable short intervals, in order to track receiver channels changes, the OFDM robustness to fast fading can be improved

and co-channel interference can be reduced, thus achieving higher system spectral efficiency.

QoS can also be implemented by assigning different number of subcarriers to users thus tuning data rate and error probability for each one of them.

6.1.4. Single Carrier – Frequency Division Multiple Access

LTE uplink transmission scheme is Single Carrier-Frequency Division Multiple Access (SC-FDMA), a slightly variation of OFDMA, designed in order to meet the power consumption requirements of UEs.

[Figure 28] shows the block diagram of a SC-FDMA transmitter.

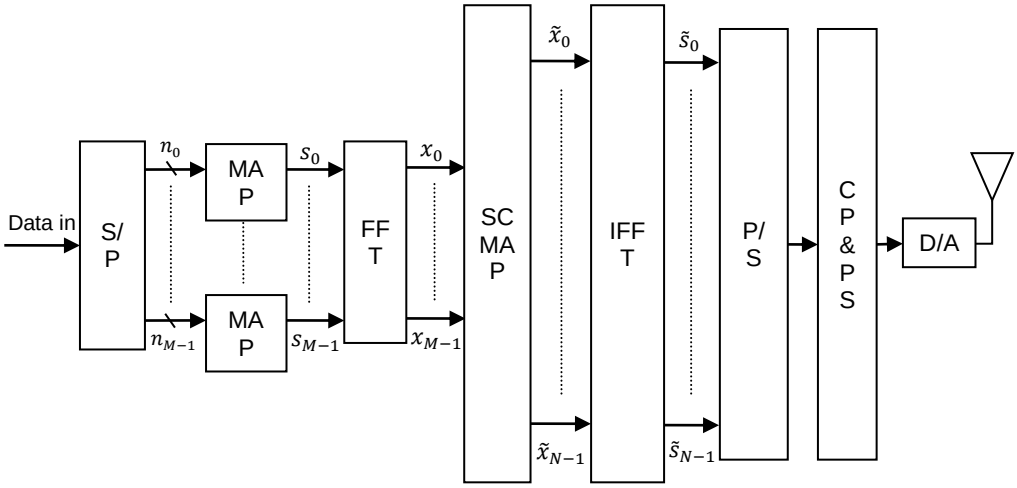

Figure 29 - SC-FDMA Transmitter

Functional blocks shown above in SC-FDMA transmit chain are:

- **Serial/Parallel converter (S/P):** multiplexes input bit stream into M bit streams which enter the constellation mapper.
- **Constellation mapper (MAP):** converts incoming bit streams to carrier symbols (QPSK, 16QAM and 64 QAM modulation are currently supported by LTE and LTE-Advanced).

- ***M*-point DFT (FFT):** converts time domain SC symbols into M discrete tones.
- **Subcarrier mapping (SC-MAP):** maps (using either a “distributed” or “localized” mode) the M discrete tones to specific transmission subcarriers.
- ***N*-point IDFT (IFFT):** converts mapped subcarriers (where $N > M$) back into time domain for transmission.
- **Parallel/Serial converter (P/S):** serializes the N parallel time domain blocks.
- **Cyclic prefix and pulse shaping (CP&PS):** adds cyclic prefix in order to increase multipath fading immunity and executes a pulse shaping (low-pass filtering).
- **Digital-to-Analogue converter (D/A):** converts digital signal to analogue.
- **TX antenna:** up-converts to RF for transmission.

The receiver side chain is characterized by an inverse of the transmission process. Figure 2.11 shows the block diagram of a SC-FDMA receiver.

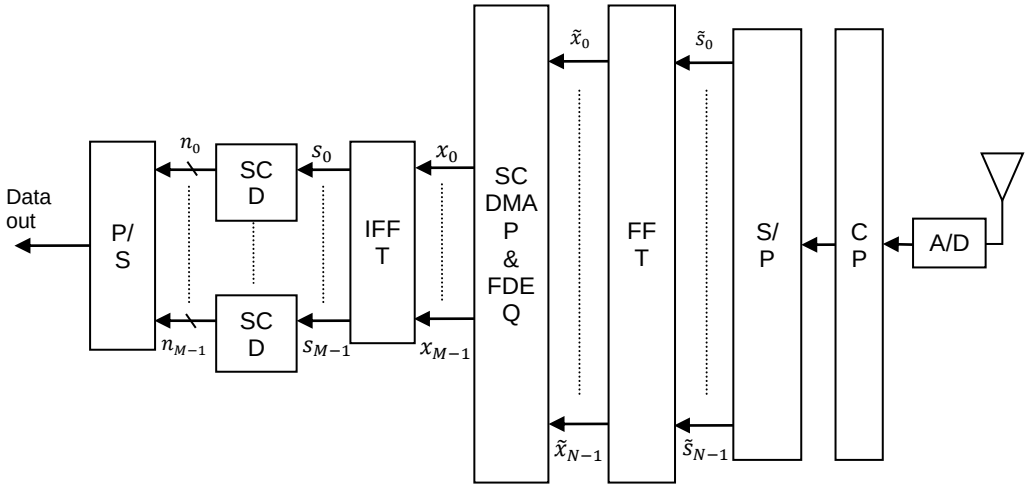

Figure 30 - SC-FDMA receiver

SC-FDMA receiver blocks are:

- **RX antenna:** down-converts the received RF signal.
- **Analogue-to-Digital converter (A/D):** converts analogue signals to digital ones.
- **Cyclic prefix (CP):** removes the cyclic prefix.
- **Serial/Parallel converter (P/S):** coverts serial time domain blocks in N parallel blocks.
- **N-Point DFT (FFT):** converts the received signal into frequency domain.
- **Subcarrier de-mapper and frequency domain equalizer (SC-DMAP & FDEQ):** subcarrier de-mapping and frequency domain equalization is performed in order to produce M-equalized symbols.
- **M-Point IDFT (IFFT):** converts the equalized symbols back to the time domain.
- **Subcarrier detector (SCD):** detects subcarriers and decodes time domain symbols.
- **Serial/Parallel converter (S/P):** produces the final output bit sequence from input M bit streams.

OFDMA and SC-FDMA signal processing techniques show many similarities:
They share the most of functional blocks of transmitter and receiver.
However some differences have to be noted: unlike OFDMA, the SC-FDMA discrete subcarriers are not independently modulated. Due to this, SC-FDMA achieved PAPR is lower than OFDMA (it is usually lower by 2 dB), which makes this scheme suitable for UE side transmissions.
As drawback, SC-FDMA introduces more complexity both at receiver and transmitter sides.
Complexity added to the transmitter is generally considered as negligible, but the increased complexity of the receiver is high if requirements for supporting multiple users in parallel are included:

- eNodeB must allocate to each uplink stream a corresponding IFFT module;

- Increased dynamic range at the transmitter IFFT input affects the dynamic range at the eNodeB receiver FFT input as well.

Even though those drawbacks, the increased eNodeB receiver complexity becomes perfectly tolerable considering achieved benefits about UE power consumption.

6.1.5. LTE downlink multiplexing scheme

E-UTRA data channels are all shared among users.

During each Transmission Time Interval (TTI) of 1ms, a scheduling decision has to be taken about which users are assigned to which time/frequency resources during next TTI.

In OFDMA scheme, users are allocated a specific number of subcarriers for a specific amount of time.

A dedicated scheduling algorithm executed at the eNodeB MAC layer handles allocation of time/frequency space to UEs.

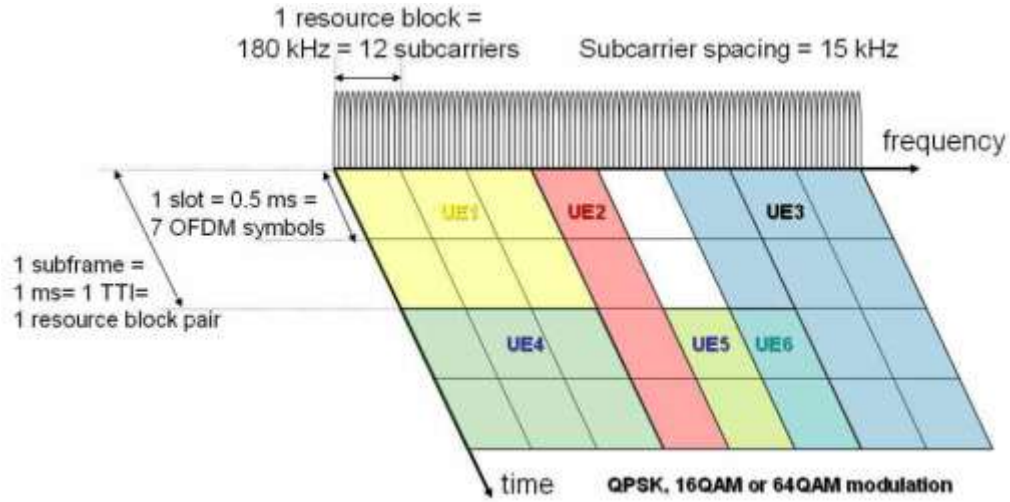

Figure 31 - Time and Frequency domain allocation

LTE signal frames are defined as being 10ms in duration. They logically are sub-divided into 10 sub-frames, each sub-frame being 1ms long (1 TTI).

Each of those sub-frames is further divided into two time slots, each of 0.5 ms duration.

Slots can consist of either 6 or 7 ODFM symbols, depending on whether the normal or extended cyclic prefix is used.

Data symbols are independently modulated and transmitted over a high number of closely spaced orthogonal subcarriers.

In E-UTRA, downlink modulation schemes QPSK, 16QAM, and 64QAM are available for subcarrier modulation.

Subcarriers in LTE are spaced by 15 kHz. The total number of available subcarriers depends on the overall transmission bandwidth of the system (which can go from 1,25 MHz up to 20 MHz).

There are 2 types of frame structure available in E-UTRA:

- ***Type 1.*** Type 1 frame structure is used for FDD. The 10 ms radio frame is divided into 20 equally sized slots of 0.5ms.

  A type 1 sub-frame consists of 2 consecutive slots, as shown in Figure 31.

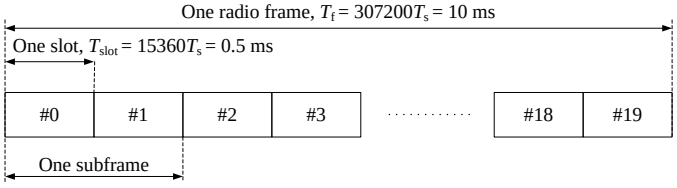

Figure 32 - LTE Frame Type 1

T_s, The basic time unit, corresponds to 30.72 MHz.

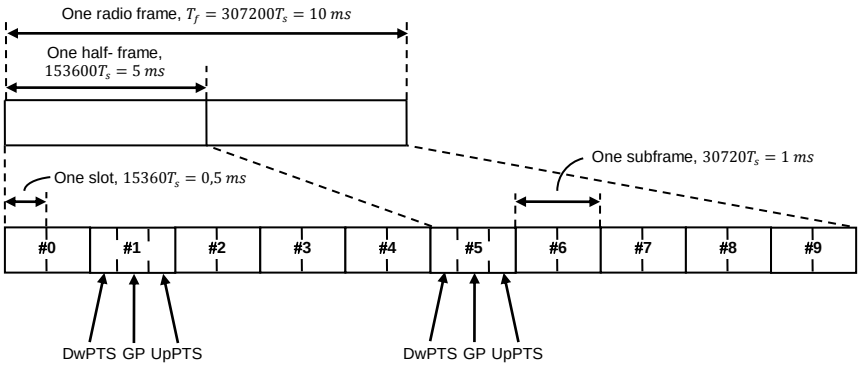

Figure 33 – LTE Frame Type 2

- ***Type 2.*** Type 2 frame structure is used for TDD. The 10 ms radio frame consists of 2 half-frames lasting 5 ms each.
  Each half-frame is further divided into five sub-frames lasting 1ms each (1 TTI), as shown in Figure 32.

All sub-frames, which are not special sub frames, are defined as composed by 2 slots of length 0.5 ms each.
The special sub-frames consist of the three fields:

- **DwPTS** (Downlink Pilot Timeslot),
- **GP** (Guard Period),
- **UpPTS** (Uplink Pilot Timeslot).

DwPTS, GP and UpPTS can have configurable individual lengths and a total length of 1ms.
The space delimited by 12 consecutive subcarriers in the frequency domain and 6 (or 7) OFDM consecutive symbols in the time domain is called Resource Block (RB).
The RB size is always the same for all bandwidths, and represents the minimum allocation unit for time-frequency resource allocation.
Table 2.2 reports the number of available RBs for different LTE bandwidths.

Channel Bandwidth (MHz)	1,4	3	5	10	20
Number of subcarrier	72	180	300	600	1200
Number of RBs	6	15	25	50	110

Figure 34 - RB configuration for different LTE channel bandwidths

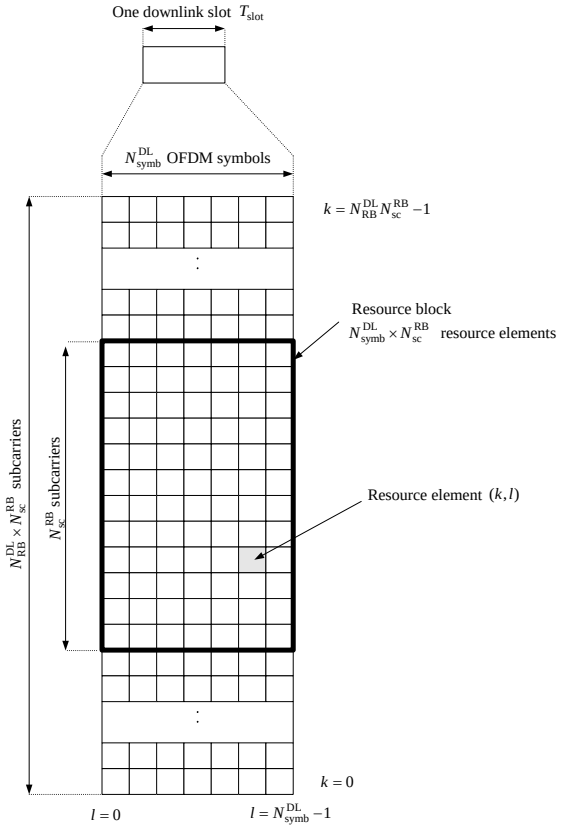

Figure 35 - Downlink resource grid

Figure 35 shows the downlink time-frequency resource grid composed by single resource elements, where:

- T_{slot} is the slot duration (e.g. 0.5 ms);
- N_{symb}^{DL} is the number of OFDM symbols in a downlink slot (6 or 7 symbols, depending on CPC length);
- N_{sc}^{RB} is the RB size in the frequency domain, expressed as a number of subcarriers (e.g. 12 subcarriers);

- N_{RB}^{DL} is the downlink bandwidth configuration, expressed as a number of RBs (e.g. 110 resource blocks for 20 MHz channel bandwidth).

The space delimited by one subcarrier in frequency domain and one OFDM symbol in the time domain is called Resource Element (RE).
In order to facilitate carrier offset estimation, channel estimation and time synchronization, specific REs are employed for transmission of reference signals. Reference signals are spread in time and frequency across the OFDMA space.
Channel response matrix on subcarriers bearing the reference symbols can be directly computed and, by interpolation, channel response matrix on remaining subcarriers can be inferred as well.

6.1.6. LTE uplink multiplexing scheme

The uplink frame structure resembles the downlink one for both FDD and TDD cases. The multiplexing scheme is also the same, as shown in Figure 36.

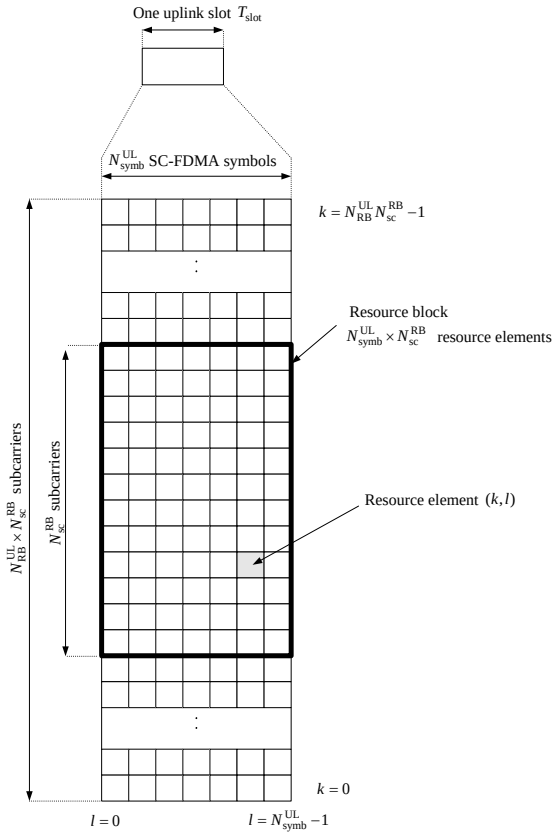

Figure 36 - Uplink resource grid

Where:

- N_{symb}^{UL} is the number of OFDM symbols in a uplink slot (e.g. 6 or 7 symbols);
- N_{RB}^{UL} is the uplink bandwidth configuration, expressed as a number of RBs (e.g. 50 subcarriers for 20 MHz channel bandwidth).

6.1.7. MIMO

Multiple Inputs, Multiple Outputs (MIMO) comprises different antenna configuration modes meeting the demands for higher data rate and better cell coverage without requiring an increase of either transmit power or bandwidth.

MIMO techniques can be applied both in downlink and uplink and specifically refers to the use of multiple antennas both at transmitter and receiver sides.

UE design constraints may limit use of multiple antennas for uplink direction (a mobile phone could be too small to fit multiple antennas and also its battery power could limit the total amount of antennas).

Each MIMO receiving antenna receives signals transmitted by all of the transmitting antennas (Figure 37).

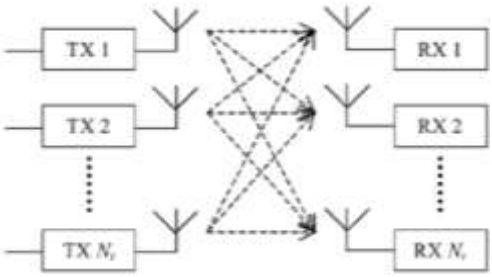

Figure 37 - Nt x Nr MIMO system representation

Assuming a time-independent narrowband channel, the (gain) element $h_{i,j}$ refers to a signal going from antenna i to antenna j.

Given a system formed by N_t TX antennas and N_r RX antennas the "channel matrix" H can be defined as:

$$H = \begin{bmatrix} h_{1,1} & \cdots & h_{1,N_t} \\ \vdots & \ddots & \vdots \\ h_{N_r,1} & \cdots & h_{N_r,N_t} \end{bmatrix}$$

Given $x = [x_1 \quad \cdots \quad x_{N_t}]^T$, where x_i is the signal transmitted on the i-th TX antenna, $y = [y_1 \quad \cdots \quad y_{N_r}]$, where y_j is the signal received on the j-th RX antenna and n a noise component, then the received signal can be defined as:

$$y = Hx + n$$

Variations of MIMO are:

- **Single Input Single Output (SISO)**, where $N_t = N_r = 1$;
- **Single Input Multiple Output (SIMO)**, where $N_t = 1$ and $N_r > 1$;
- **Multiple Input Single Output (MISO)**, where $N_t > 1$ and $N_r = 1$;

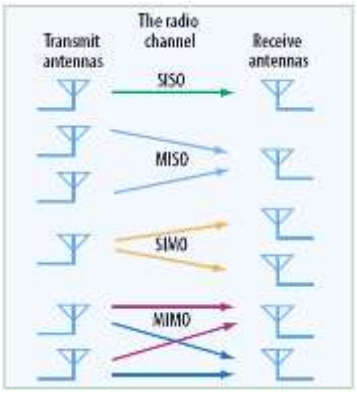

Figure 38 - MIMO configurations

6.1.8. MIMO configurations

MIMO technologies employed in commercial systems are a mix of three complementary aspects: diversity, spatial multiplexing and beam forming.

While it is not possible to achieve maximum diversity, spatial multiplexing and beam forming gains, each MIMO system is usually a trade-off between these three aspects:

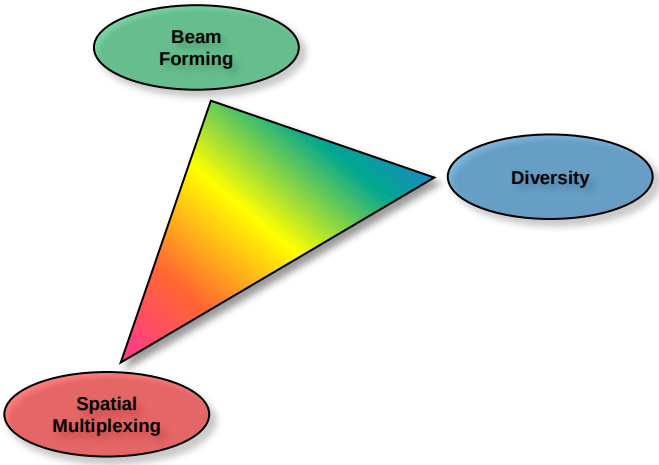

Figure 39 - Trade-off for MIMO systems

6.1.9. Diversity

Diversity improves the reliability of transmissions by using two or more communication channels exploiting their different characteristics. There can be several kinds of diversity:

- **Time diversity.** Time diversity is achieved by transmitting multiple versions of the same signal at different times.

  Alternatively, a redundant FEC code can be added to the signal thus resulting in the message being spread in time (using bit-interleaving) before being transmitted.

  Time diversity avoids error burst, leading to efficient error correction.
- **Frequency diversity.** The signal is transferred using different frequencies (e.g. OFDM) or spread over a wide spectrum affected by frequency-selective fading (e.g. frequency hopping or direct sequence spread spectrum).
- **Space diversity.** Spatial diversity requires a signal being transmitted over different propagation paths. In case of wired transmission, this can be achieved by transmitting via multiple wires.

  Regarding wireless transmissions, space diversity exploits the use multiple antennas, and can be classified according to the antenna configuration employed:
  - ***Transmit diversity***, if multiple transmission antennas are used (e.g. MISO systems);
  - ***Receive diversity***, if multiple receiving antennas are used (e.g. SIMO systems).

  If antennas are sufficiently spaced, the signal transmitted/received by each antenna can be considered independent from the other ones.
- **Polarization diversity.** It consists of multiple versions of a signal transmitted and received using different polarizations for each transmitting antenna. A diversity-combining receiver is required for signal decoding.
- **Multiuser diversity.** Multiuser diversity exploits the presence of different users in a fading environment.

It increases the throughput of the multiuser system exploiting the independence of fading statistics for different users.

Opportunistic user scheduling algorithms employed either at the transmitter or the receiver achieves multiuser diversity.

Such opportunistic schedulers allocate resources to users only when their channel conditions are favourable. Opportunistic scheduler therefore needs to have knowledge of SNR experienced by each user. Such scheduling requires a centralized control unit and also introduces signalling overhead (i.e. feedback reports about channel quality).

The system capacity of system exploiting multiuser diversity increases with the number of users, up to the so-called "saturation point".

- **Cooperative diversity.** It is a form of antenna diversity gain achieved by exploiting cooperation of distributed antennas belonging to different nodes.

  Diversity gain is obtained by decoding combination of relayed and direct signals in wireless multi-hop networks. Signals can either be relayed by dedicated relay nodes or just by other user nodes.

Regardless of the diversity type, each replica (copy) of the transmitted signal experience different fading and interference levels, thus the probability of having all signals in deep fading holes simultaneously is always significantly reduced with respect to a non-MIMO system. This kind of protection is defined as "diversity gain", which does not increase directly the transmission data rate, but positively affects the reliability of transmissions or alternatively the required total transmission power.

The maximum diversity gain achievable is $N_t \times N_r$, where N_t and N_r are the number of TX and RX antennas respectively.

6.1.10. Spatial multiplexing

In a MIMO system, received signal on each RX antenna is a superposition of the signals transmitted by the TX antennas. If those signals show sufficiently different

spatial signatures, the receiver can benefit of multiple spatially separated channels. Such channels share the overall SINR and avoid the throughput saturation.
Independently encoding, modulating and transmitting different data blocks using different antennas achieve SINR sharing.
A $N_t \times N_r$ MIMO configuration supports transmissions of at most $M = \min(N_r, N_t)$ simultaneous streams, given sufficiently high SINR values. M Value results to be the maximum rank of the channel matrix H and also the maximum achievable "multiplexing gain".
If inter-stream interference is considered, the value $r = rank(H) \leq M$ - "channel rank"- specifies the number of parallel streams supported by the channel.
If we define the value m as "transmission rank", i.e. the number of actually transmitted parallel streams, m needs to be limited by the channel rank r in order to avoid destructive spatial interference among transmitted data streams.
While beam forming, diversity and single-antenna transmissions are all single-rank schemes ($m = 1$), spatial multiplexing provides the possibility to use transmission ranks higher than one ($1 \leq m \leq M$) thus achieving high multiplexing gain and increasing the system peak rate if sufficiently high SINR are available.

6.1.11. Beam forming

Beam forming is a signal processing technique enabling to control the angle of signal transmission or reception using multiple antennas.
In transmission, transmit beam forming allows to focus transmit energy over a precise angular direction (e.g. the direction where most of UEs are located).
In reception, receive beam forming allows to calibrate a group of antennas in order to predominantly receive from a chosen angular direction (e.g. the direction where the eNodeB is located).
In order to implement a beam forming system, arrays of small non-directional antennas are used.

Figure 40 - Beam Forming

A beam-forming pattern can be fixed or adaptive and specifies the radiation pattern of the antenna array. The beam pattern specifies the angular direction (phase) and the transmit power (amplitude) of the beams in order to maximize the total power over the RX antenna.

Beam pattern types are:

- **Switched Beam Forming**. The beam pattern is chosen between a finite numbers of fixed predefined patterns. The angular direction of the beams is electrically calculated and a suitable fixed beam is selected.
- **Adaptive Beam Forming.** The beam pattern is adjusted in real-time reacting to UE movements. The complexity and the cost of such a system is obviously higher than switched beam forming one.

7. LTE ADVANCED

The LTE-Advanced technology will support the IMT Advanced requirements for 4G cellular systems. It will constitute an improvement of the existing rel. 8 specifications, with backward compatibility support for LTE. The targets of LTE and LTE-Advanced are compared in [Table 2]

LTE is a great step forward in the evolution of wireless networks.

The possible improvements of this technology, on the way to LTE-Advanced, can be carried out on two different directions: provide higher resource availability or use the existing resources in a more efficient way. However, the spectral efficiency reached with LTE, is already close to the theoretical limits, so the improvements introduced with LTE-Advanced are mainly due to the extension of the existing concepts than to the introduction of new access technologies[10].

Thus LTE-Advanced is a natural evolution of LTE.

	LTE	LTE ADVANCED
data rate	300 Mbps	300 Mbps
UL peak data rate	75 Mbps	500 Mbps
Max. available bandwidth	20 MHz	100 MHz
DL MIMO support	Up to 4 layers	Up to 8 layers
UL MIMO support	Single layer	Up to 4 layers
DL Peak Spectrum Efficiency (Bps/Hz)	15	30
UL Peak Spectrum Efficiency (Bps/Hz)	3,75	15

Table 2 - LTE vs LTE-A requirements

In order to reach the targets shown in [Table 2] and meet the IMT Advanced requirements, the following concepts are introduced with LTE-Advanced:

- Carrier aggregation of non-contiguous spectrum (up to 100 MHz)

[10] *An higher spectral efficiency is however possible, with the availability of more spatial layers for MIMO operations*

- Extension of cell coverage and capacity through relay nodes
- Increased MIMO support with more spatial layers
- Distributed antennas operating at the physical layer

Of course, besides the opportunities that these aspects bring, the challenges must also be taken in into account.

The aggregation of non-contiguous spectrum requires the support of an efficient control information signalling and a suitable resource-scheduling algorithm.

The presence of relay nodes requires an efficient allocation of access and backhaul resources, in order to obtain high performance and mitigate the interference.

The increasing number of spatial layers requires am efficient transmission mode selection algorithm.

These elements, and many others, need to be considered in the design process of the 4G LTE-Advanced technologies.

7.1.1. Relay support for LTE-Advanced

Ubiquitous high-data-rate coverage is a key element for next generation wireless networks. All the users, including those in the most unfavourable channel conditions (i.e., cell-edge users), expect a high-data-rate service. An efficient distribution of capacity all over the cell is then required.

With conventional architectures this could be accomplished by the replication of eNodeB. However the dense deployment of Base Stations would result in a costly and complex solution, because the eNodeB require fibre access to the backbone infrastructure.

With LTE-Advanced, wireless multi-hopping through relay nodes is introduced as a solution to provide a more efficient use and distribution of OFDM radio resources. Relay nodes have less cost and functionality than eNodeB.

Their full wireless connectivity, both towards the backhaul (here represented by the DeNB, which is the eNodeB whose service is enhanced by the deployment of a relay node) and towards the access network (the User Terminals), makes them a cost-effective solution to extend cell coverage and capacity.
Through relay deployment, new multi-hop networking architectures are introduced in the capabilities of LTE-Advanced. In this context, the radio resource management plays an important role to efficiently exploit the possibilities of relay-enhanced cells.
Currently, protocols and specifications to regulate relay nodes operations are still being investigated.

7.1.2. Relay Nodes

For multi-hopping operations, a relay node is seen as a UE from the Base Station perspective and it is seen as a Base Station from the UE perspective [4]. A UE should associate with a relay node, if this choice is profitable in order to increase the cell capacity: the amount of resources that a two-hop transmission requires, must be smaller than what required by a single hop transmission, where the eNodeB is directly addressed. This is a condition which is verified whereas the stronger signal that the user receives from the relay node, as opposite from the signal received from the eNodeB, allows a (cell-edge) user to use higher modulations or complex MIMO schemes.
The advantage is clear if we consider that the employment of the highest modulation available in LTE (64QAM) allows the same transmission to occupy about 1/10 of resources than with the lowest modulation (QPSK) [5].
However some critical condition could happen: a UE could receive a better signal from the relay node, but the relay-to-DeNB signal could be worst than the UE-to-eNodeB signal (the UE is sited between the DeNB and the relay node). A proper channel assignment algorithm must be able to spot and avoid a similar situation.
Relay nodes operate on a decode-and-forward scheme [4]. In contrast with repeaters, which simply amplify and forward the received waveform (interference and noise included), a relay node decodes the received signal and later transmits the relevant data to the next hop. Moreover, repeaters need

to operate within the Cyclic Prefix length (practically without delay), thus a portion of the output waveform causes a positive feedback in the repeater receiver interface. Relay nodes are not affected by this problem, since they are not subject to the delay constraint.

The signal forwarding may be explicit or transparent. With the transparent scheme a relay node decodes the transmission and send it to the receiver, which will reply with an HARQ ACK/NACK message. The relay node intercepts the HARQ feedback and, if necessary manages the retransmission. With the explicit scheme, first the relay node acknowledge the transmission to the sender, through HARQ ACK/NACK feedback, than, once the data has been correctly decoded, it sends it to the destination. By doing so the relay node acts like a checkpoint for the transmissions and decouples the sender and the receiver, making it possible to apply the paradigm by which it is considered as a BS from the UE and as a UE from the DeNB. The advantage of the transparent scheme is that it would allow the receiver to combine the waveforms transmitted both by the sender and the relay node. However this scheme is highly infeasible (for instance the relay node would be required to intercept control information messages in order to know the modulation and MIMO transmission mode of the subsequent transmissions), and the explicit one is adopted.

With the architecture of the relay-enhanced cell, the interference issue must be taken into account: if the same frequencies are used on the backhaul link (DeNB-to-Relay link) and on the access links (Relay-to-UEs links), simultaneous transmission and reception at the relay node would cause a strong interference.

For the above reason, radio resources need to be assigned to relay nodes, so that they do not transmit and receive simultaneously on overlapping frequency resources. This is referred as the relay-duplexing problem and it can be addressed by properly regulating the link scheduling activity through an efficient link-scheduling algorithm.

7.1.3. Distributed antenna system

Distributed Antenna Systems (DAS) are employed in LTE-A to increase coverage and/or transmission rate. In a DAS deployment, the eNodeB is connected via a fibre interface to a set of M spatially distributed antennas or *Remote Units* (RU), whose coverage may overlap. Some of these antennas may be co-located with the eNodeB itself. Transmissions are arranged in time slots called Transmission Time Intervals, (TTIs), whose duration is 1ms. Each RU transmits a dedicated *frame* on each TTI, and a centralized eNodeB scheduler via the fibre interface communicates the frame allocation to it. At the *logical* (MAC) level, the eNodeB scheduler allocates *frames,* i.e. vectors of (*Virtual*) *Resource Blocks* (RBs) to its associated UEs on each TTI. Each RB carries a fixed number of symbols, which translate to different amounts of bits depending on the *modulation and coding scheme* (MCS) used by the RU on that RB. In general, more information-dense modulations (e.g., 64QAM, yielding 6 bits per symbol) are favoured when a better channel to the UE is perceived. The quality of the wireless channel varies over time and is generally different from one RU to the other. For this reason, UEs report their perceived downlink channel states to the eNodeB scheduler as a Channel Quality Indicator (CQI). The CQI is an index in a standard table, computed by the UE according to the measured Signal to Noise Ratio (SNR), and determines the MCS that the eNodeB should use and the number of bits per RB, called Transport Block Size, TBS. Accordingly, we will sometimes use the word CQI to refer either of the latter two, trading a little description accuracy for conciseness and ease of reading. A single UE may receive transmissions from different RUs on the same time-frequency resource. Those transmissions are in fact *spatially separated*, and can therefore be reconstructed at the UE. A UE has a maximum number of *spatial layers* (or just *layers* for short) that it can decode simultaneously, which may be smaller or larger than the number of RUs that can target it at a given TTI. The number of layers changes over time, and is also reported by the UE to the eNodeB scheduler.

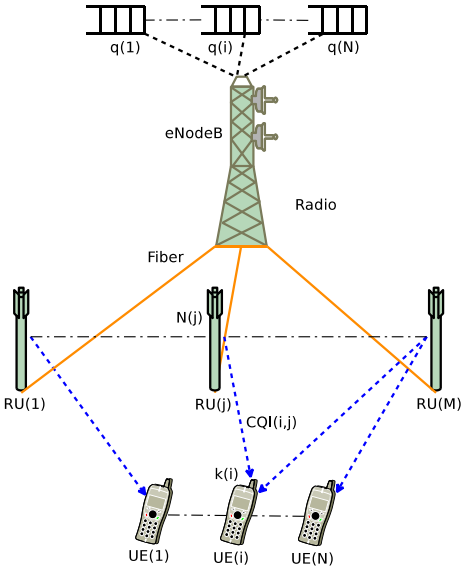

Figure 41 - a DAS deployment

UE traffic is physically buffered at the eNodeB. In order to build a schedule in a TTI, the eNodeB scheduler selects *which UEs* are going to be targeted, by which RU(s), using *which MCS* to guarantee reliable transmission, and employing how many RBs. These decisions are made, either sequentially or jointly, based on the reported CQIs, the backlog and type of traffic of each UE, the QoS requirements, etc. The goal of such decisions is, in general, to maximize the cell throughput.

8. LTE AND LTE ADVANCED ALGORITHMS FOR RESOURCE SCHEDULING

The first of scheduling algorithms analysed in this work is about exploiting absolute time for scheduling downlink *voice* flows in LTE networks.

Voice has been selected as the first target for our LTE network optimization studies mostly because of the LTE shifting towards a full packet-switched network, which should support high quality voice services at least resembling the ones provided by circuit, based networks. Moreover, well-established procedures exist to estimate the (subjective) quality perceived by the user from (objective) end-to-end network-level measurements such as loss and delay, i.e. the E-Model or Mean Opinion Score (MOS) [17]-[18]. Another reason is that behaviour of a voice application at the *receiver* side, and, more specifically, its play-out (PO) buffering mechanism is somewhat predictable, so that it is possible to infer the (non negligible) contribution of the voice receiver to loss and delay.

We devise a scheme, called OptiMOS, supposed to work at the MAC layer downlink scheduler, which assigns to each packet a deadline equal to the *presumed* playback instant at the receiver. Thus, late packets (those which the receiver would drop), can be dropped directly at the scheduler, and do not waste radio resources. Likewise, early packets can be delayed until their actual playback point without affecting user perception, thus increasing the chance that packets with tighter deadlines make it to the receiver on time. The playback point at the receiver is estimated by emulating, for each flow, a simplified *adaptive* receiver buffer, whose playback point is varied dynamically from one talkspurt to another. At the end of a talkspurt, simple computations allow OptiMOS to infer the *optimal-a-posteriori* playback instant, i.e. the one that would have guaranteed the highest MOS for that talkspurt ([21],[22]). The history of the optimal playback instants is then used to infer the *new* presumed playback instant for the subsequent talkspurt. The optimal-a-posteriori playback instant is in turn computed using E-Model formulas. These require mouth-to-ear delay, which can only be computed if the packet generation and (presumed) playback times are known, which in fact requires clock synchronization between the sender and scheduling point.

The first desirable attribute of OptiMOS is that it is inexpensive. The whole implementation framework requires less than 100 byte per flow, and the time overhead is also negligible, since its overhead is *constant* with respect to the number of flows. Second, it is *robust* to synchronization errors: clock errors in the order of ±10ms are tolerated without any appreciable quality degradation. Third, it can work in conjunction with *any* deadline-based scheduler, e.g. the well-known Earliest Due Date (EDD), or – for wireless networks having user- and time-dependent channel conditions – with schedulers that also take into account the channel state (e.g., [12]). Fourth, it is not tailored to a specific technology. The only architectural requirements are: i) that the *packet source* and the *scheduling* point are synchronized; ii) that OptiMOS is able to read the packet timestamps on arrival and the wall clock on their departure; iii) that *each packet* contains RTP timestamps (i.e., that packets are not fragmented *before* they get to the scheduler, although they can indeed be fragmented *by* the latter) and, iv) that the downstream segment of the path only includes a PO buffer (plus, possibly, delays that can be either measured online (e.g., those due to ARQ retransmissions) or estimated (e.g., propagation/processing delays). OptiMOS works with any access network with a coordinated point-to-multipoint scheduling under the above hypotheses, e.g. WiMAX, LTE or HSDPA, or wireless LANs such as 802.11e HCCA. For the sake of concreteness, here we evaluate it in the framework of an LTE cell.

Few works so far have advocated using the MOS in network QoS solutions. Some propose solutions at the *application* level, and rely on the receiver conveying feedback to the sender, which in turn reacts by adapting its sending pattern (e.g., [0[20]]). Such techniques can only react on a round-trip time timescale, which is an order of magnitude larger than the packet transmission timescale. Other works ([23], [25]) advocate using MOS for allocating resources inside the network. Some (e.g., [23][24]) solve a cross-layer optimization problem, whose objective is to optimize the sum of the MOS for a set of users. The MOS is assumed to be a non-linear function of the *data rate* of continuous, adaptive-rate applications. Such a definition is non-standard and leaves out VoIP, which is intermittent and non adaptive. In [25] a formula that links a MOS to the data rate of web browsing is inferred, and a scheduling algorithm for LTE is proposed that maximizes the utility of web downloads.

OptiMOS is implemented on downlink side a LTE enhanced NodeB (*DN*). Packets arriving at the DN are encapsulated into MAC Protocol Data Units (PDUs), possibly after concatenation and/or fragmentation, and are queued into per-flow MAC queues. PDUs are associated with *internal deadlines*. The latter can be read by a deadline-aware *scheduler* (which uses them to sort PDUs) and by a *buffer management scheme*, which drops PDUs whose deadline has expired. MAC queues can either be FIFO or sorted by PDU deadline. Once a PDU is scheduled for transmission, it is handled to the physical layer which takes care of delivering it to the receiver, possibly using retransmissions, timeouts, ACK/NACK, etc. We assume that the DN knows if and when a PDU has been successfully delivered to the receiver. We assume that the DN can read a wall clock, which has a limited skew with that of the senders, in the order of some milliseconds. As far as VoIP applications are concerned, we assume that speech frames are encapsulated into RTP messages, and that RTP timestamps are taken by the sender's wall clock as well. Each receiver may employ whichever *adaptive* PO buffering scheme. We assume that the DN knows the codecs used by the VoIP applications and their settings, e.g. their period. This information can be acquired by the DN at flow setup (e.g., as part of an admission control procedure), or – in some cases – could also be inferred by the latter by looking at the RTP header and payload. This is achieved by means outside of the scope of this paper. Furthermore, we assume that this information is stable during the flow lifetime, or at least that the DN can be made aware of any modification to it. Finally, although OptiMOS only takes into account VoIP flows, we remark that the simultaneous presence of other traffics (e.g., video or TCP) at the DN does not affect its performance.

OptiMOS

The OptiMOS framework is shown in Figure 49. For ease of exposition, we first describe OptiMOS making some simplifying assumptions: specifically we will initially assume that *one* speech frame is included in each RTP packet, and that packets are not fragmented at the DN, so that one MAC PDU carries exactly one speech frame. The extensions to the general case of *n* speech frames per RTP packet and/or packet fragmentation into *m* PDUs will be discussed later.

RTP packets are stamped with the speech frame generation time on creation. Call $g_{i,k}$ the generation time of the k^{th} frame in the i^{th} talkspurt, $i,k \geq 0$, and $a_{i,k}$ the arrival time at the DN of the IP packet carrying it. The network delays packets, and may also drop them and/or deliver them out of sequence. A packet arriving at the DN at time $a_{i,k}$ is encapsulated in a PDU, which is assigned a deadline $d_{i,k}$ computed as:

$$d_{i,k} = po_i - \left(a_{i,k} - g_{i,k}\right) - dl$$

Where po_i is the estimated *PO delay* at the receiver for the i^{th} talkspurt. The latter is defined as the (constant) interval between the generation and PO instant of any successfully played speech frame in that talkspurt. po_i is inferred by simulating an *optimal PO buffer (OPB)*, described in the next subsection. dl is an *estimate* of the downlink delay due to the radio interface and processing at the receiver, i.e., of all the future delay of the PDU *except* the part accountable to the receiver PO buffer. For instance, in a network employing ARQ cycles of fixed length, dl could be set as follows:

$$dl = T_{proc} + T_{radio} + \left(m - 1\right) \times T_{RTX}$$

where T_{proc} represents an estimate of the processing delay at the physical and MAC layers at the receiver, T_{radio} is the time to send a PDU through the downlink segment (i.e., the time it takes for the physical layer to deliver the PDU to the receiver, not including ACK/NACK generation and reporting), T_{RTX} is the time for a retransmission cycle and m is the number of transmissions that we want to be able to perform before dropping a PDU.

The DN buffer manager drops PDUs if their sojourn time in the MAC queue exceeds their deadline. As we will see later on, this is because the speech frames they carry are likely to be discarded anyway at the receiver PO buffer, hence transmitting them would just waste radio resources. A PDU scheduled at time $t_{i,k}$ is transmitted by the physical layer, and, after a (variable) downlink delay $dl_{i,k}$, it makes it to the PO buffer at time $q_{i,k} = t_{i,k} + dl_{i,k}$. Note that $dl_{i,k}$ can be measured at

the DN. Other than $dl_{i,k}$, the *only* additional delay that a speech frame undergoes before being played out is the one of the PO buffer at the receiver. We define the *network delay* of speech frame (*i,k*) as:

$$d_{i,k} = q_{i,k} - g_{i,k}$$

The sequence of $d_{i,k}$ is fed as an input to the OPB simulator. The latter simulates an optimal, *non-causal* adaptive PO buffer, and its purpose is to identify – *a posteriori*, i.e. once talkspurt i is over – what the optimal PO instant po_i should have been, i.e. the one that would have warranted the highest MOS for that talkspurt. The history of past optimal PO delays po_j, $0 \le j \le i$, is then used to infer the new PO delay po_{i+1} for the incoming talkspurt.

8.1.1. Optimal pay-out buffer simulator

An optimal PO buffer ([21],[22]) is simulated at the DN. The purpose of including such a component is that we want OptiMOS to predict what a *clever* adaptive PO buffering algorithm would do. This allows OptiMOS to delay early packets more (by assigning them a later deadline), and to have the buffer manager drop those with a low chance of being played out. Hereafter, we describe it in more detail.

Consider the example illustrated in [Figure 43], which shows the transmission, buffering and PO phases of a talkspurt with five VoIP frames. Let $p_{i,k}$ be the time when the *k*-th frame of talkspurt i is (virtually) passed to the decoder, i.e. played out, respectively. As $p_{i,k} - p_{i,j} = g_{i,k} - g_{i,j}$, for each *k*, *j* that are actually played, the *PO delay* po_i is constant for all the frames in a talkspurt which are actually played. Without loss of generality, we assume that $q_{i,k} = \infty$ for lost frames. The PO buffer discards frames received too late, i.e. such that $d_{i,k} > po_i$, which implies that frames with $d_{i,k} = q_{i,k} = \infty$ are also discarded. Frames discarded by the PO buffer contribute to the loss rate L_i, defined as the ratio between the number of discarded frames and the number of frames in the talkspurt. In the example of [Figure 43], $L_i = 1/5$

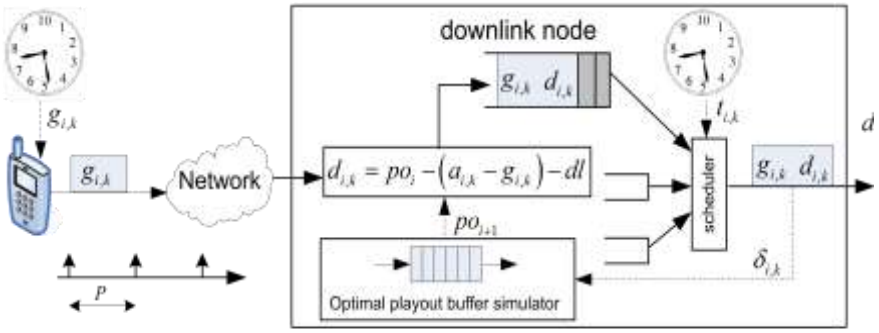

Figure 42 - The OptiMOS framework

because all frames but the first one are received "on time". Thus, the only degree of freedom of the OPB is the PO delay po_i. We define the *optimal PO delay* $po_i^{opt} \in \{d_{i,k}\}$ as the value of the PO delay that maximizes the R factor. As already pointed out, the actual formula to compute the latter depends on the specific codec used. However, by definition, the MOS is always obtained via a *non-increasing* function of po_i and L_i. Furthermore, L_i is itself a non-increasing function of po_i. Therefore, there exists an optimal value po_i^{opt} that achieves the maximum R factor for the talkspurt. The latter can be computed by searching the discrete set of network delays in a talkspurt $\{d_{i,k}\}$. Such a PO buffer algorithm is *adaptive*, since it sets the playback point on a per-talkspurt basis. Furthermore, it is obviously *non-causal*, since it requires the delay of all speech frames to be known before selecting the playback point.

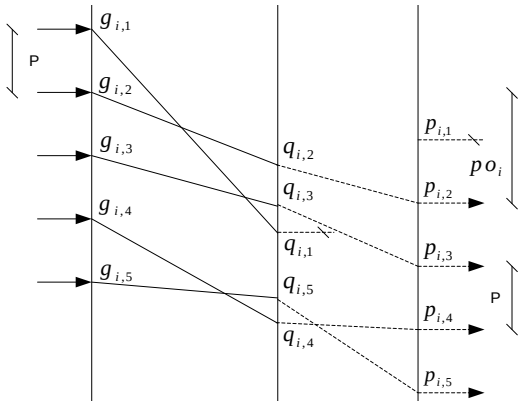

Figure 43 - VoIP frame re-ordering due to (non-causal) PO buffering. Solid lines represent network delays, dashed lines PO buffering delays

Within OptiMOS, the OPB is used as follows: after talkspurt i, the optimal PO delay po_i^{opt} is computed. Then, an exponential average is used to infer a likely PO delay for the *next* talkspurt, i.e., the $(i+1)^{th}$:

$$po_{i+1} = a \times po_i + (1- a) \times po_i^{opt}$$

with $0 < a < 1$. po_{i+1} is then used to set the deadlines for the packets in the next talkspurt. Thus, PO instants are selected based on past history. For the *first* talkspurt, when no estimate is available yet, a default value is used at the DN. The latter can be based on information regarding the call endpoints, where available. The a parameter should instead reflect how dynamic the upstream network conditions are.

8.1.2. Details and discussion

Having explained the basics of OptiMOS, we discuss some details, which allows us to remove some of the hypotheses and to make it both more general and more efficient. First of all, an RTP packet may include n speech frames, $n \geq 1$, obviously all belonging to the same talkspurt. In this case, the RTP timestamp refers to the last one, and the generation time of each one can be easily inferred from the period P, which is assumed o be known. In this case, we differentiate between the packet *deadline* and *discard time* (which had been assumed to be equal thus far). The deadline is computed based on the generation instant of the *first* speech frame; however, as far as *dropping* is concerned, we set the discard time of the packet to the presumed PO instant of the *last* speech frame. This means that we allow the PDU to be transmitted as long as it may still carry some useful content. Finally, when inserting speech frames in the OPB simulator, each one is associated to the correct generation time. Furthermore, depending on how large RTP packets are and on how the maximum PDU length is set in the downlink, one RTP packet may end up being fragmented into several PDUs at the DN (although we do not expect this to be common). In this case, all the PDUs should be associated to the same deadline and discard time, and the network delay of the packet should be set based on the transmission time of the *last* PDUs of a packet. As the network may drop packets and cannot be assumed to preserve their sequence, it might be possible that the *first* packet of a talkspurt i to arrive at the DN is in fact the k_0^{th}, $k_0 \geq 0$, at time a_{i,k_0}. Note that, since RTP packets carry sequence numbers: i) we know that k_0 other packets are missing; ii) we can compute their generation time $g_{i,j} = g_{i,k_0} - P \times (k_0 - j)$, $0 \leq j < k_0$, and iii) we can compute a *lower bound* on their network delay as $a_{i,k_0} - g_{i,j}$, directly upon arrival of packet k_0. On their arrival, these packets will thus be assigned an earlier deadline than those of packets already queued in the system. If MAC queues are FIFO,

PDU deadlines will not be monotonically increasing for a flow, which however should not be a problem for a scheduler.

OptiMOS requires the DN to distinguish talkspurts. For most codecs, the packet generation period is constant, which makes such a task easy. In fact, RTP packets carry both timestamps and sequence numbers. Thus, for those codecs which do not generate information during silence periods, one can easily observe that two consecutive packets should belong to different talkspurts if their sequence numbers are k_1, k_2, and their timestamps are such that $ts(k_2) > ts(k_1) + P \times (k_2 - k_1)$. In practical cases, a safety margin (e.g., in the order than one period) should be added to the right-hand side of the above inequality to account for jittery sources. For those codecs, instead, which send *reduced* information during silence periods, the onset of silences can be detected by examining the RTP payload size.

As far as scalability is concerned, we underline that OptiMOS does not rely on comparisons among flows, so it requires a constant overhead per PDU transmission. Its most complex operation is computing the optimal PO delay at the end of a talkspurt. This requires the R factor to be computed for each possible network delay, i.e. a $O(M)$ time overhead, assuming M speech frames in a talkspurt. Although M is not expected to be a very large number in practical cases (the *average* talkspurt length being in the order of 1s, i.e. 30-100 frames per second, depending on the codec), we can limit the time overhead without relying on this assumption. We do so by *quantizing* the network delays at a rather coarse resolution, e.g., 10ms or more, which, as we will show later on by simulation, comes with no appreciable quality degradation. Instead of storing all the *network delays*, we provision a fixed number of $B+1$ integer *delay counters*. Counter l, $0 \le l \le B-1$, is related to a network delay range equal to $[i \times Q + C, (i+1) \times Q + C[$, Q being the quantization interval and C being a delay offset which can be set to the fixed delay component along the path if the latter is known (otherwise $C = 0$ is a safe estimate). The *last* counter, i.e. the B^{th}, is related to delay range $[B \times Q + C, \infty[$.

When a speech frame arrives at the OPB simulator, the related counter is increased, which requires constant time. When computing the optimal PO delay,

the cost is $O(B)$. Assuming 10ms as a quantization interval, and 500ms as a maximum delay, we obtain $B = 50$, which makes this scheme affordable. Finally, delay ranges need not be of the same size, nor constant over time. In fact, progressive delay ranges may be a viable solution, especially for long-range calls. Furthermore, optimal values of Q and C for a flow can be dynamically estimated by examining the delay distribution as the connection progresses. We do not pursue this issue further in this paper, leaving it for future work.

8.1.3. Performance Evaluation

In this section we evaluate the performance of OptiMOS in an LTE cell. We show that OptiMOS actually boosts the MOS of VoIP flows. Second, delay quantization in the OPB simulator and errors in the clock synchronization (both as large as ±20ms) do not significantly affect the voice quality, which makes this scheme computationally affordable and robust. Third, OptiMOS yields qualitatively similar results with different PO buffering schemes.

All the scenarios have been simulated for 200 seconds. A single LTE cell is simulated, with an eNodeB at its centre and a variable number of receivers experiencing varying channel conditions. VoIP sources use the GSM AMR Narrow Band codec, which generates a 32-byte speech frame every 20ms, and are located upstream a core network, which introduces a variable delay. VoIP sources employ VAD. Talkspurts and silences are distributed according to Weibull functions with shape and scale (1.423, 0.824) and (0.899, 1.089) respectively [21] Header compression is also employed. EDD is used as a scheduler, with either a fixed or a dynamic, OptiMOS-computed deadline. The RLC layer at the eNodeB has been configured with the *Unacknowledged Mode*, with a fixed PDU size of 40 bytes (which implies no IP packet fragmentation). The physical layer has been configured in order to reduce the whole system bandwidth to 72 subcarriers (about 1MHz) in order to approach cell saturation within a manageable number of users. The physical layer is a two-state Markov chain, with a 0.5 state transition probability. In one state, the channel quality stays constant, in the other a new Channel Quality Indicator is extracted from a uniform distribution. As far as OptiMOS is concerned, the initial PO delay and a value hardly affect the steady-state performance in the considered scenarios. A reliable study on how to set both parameters should take into account network dynamics in real-life conditions, which is outside the scope of the present work. In the simulation, we set these values to $po_0 = 200ms$ and $a = 0.75$.

As a first experiment, we show that time synchronization errors – within up to ±20 ms, hardly affect the performance of OptiMOS. A set of 100 users is split into two groups: a first half experiencing a 50 ms average core network delay, and a second half experiencing 100 ms average core network delay. Figure 44 shows the

performance of OptiMOS with an increasing time synchronization error between the clocks of the sources and the DN. Specifically, adding them a value taken from a uniform interval of increasing width corrupts the timestamps of the packets. The receivers use the OPB algorithm. The figure clearly shows that the performance is hardly affected by such errors, at least within widths of up to 20ms. On the other hand, the varying core network delay clearly influences the MOS. All the following simulations are carried out assuming perfect synchronization for simplicity, and we stick to a single core network delay distribution for all flows, with values uniformly distributed in [50;80] ms.

Figure 45 shows the average MOS that OptiMOS achieves with 1 and 2 speech frames per packet. The scenario is with 120 VoIP users. As the figure shows, aggregating frames is detrimental to the performance, as it imposes a larger constant delay at the source, which decreases the MOS.

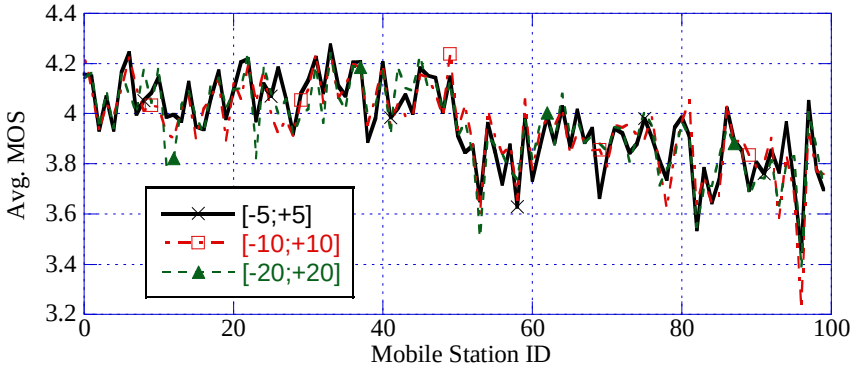

Figure 44 - Average MOS against synchronization error

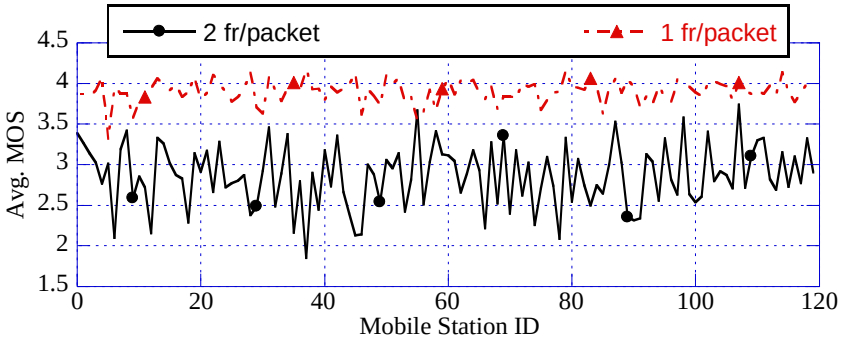

Figure 45 - Average MOS for two values of speech frame aggregation

We now compare OptiMOS to a fixed deadline assignment, with different PO buffers. The first comparison scenario is with VoIP-only traffic. 120 VoIP sources transmit their traffic to mobile receivers in the LTE cell. Note that this is around the saturation point, i.e. adding more users leads to remarkable performance degradation. Preliminary simulation carried out with EDD show that the best performance with a static deadline is obtained at 100ms, which is the one we select for comparison. We simulate identical receivers with three different PO buffers: the OPB, the one in [22], which approximates the latter in a causal setting, and the adaptive PO buffer used in H323plus [26] employed by Ekiga. We included it as its playback adaptation algorithm is completely different from the other two. Figure 46 shows the average MOS for both OptiMOS and a static 100ms deadline for all the users, with the three PO buffers. The figure clearly shows that the PO buffer has a major impact on the performance, which is a known fact. However, the *worst* performance result achieved by OptiMOS (in conjunction with an H323plus buffer) is still appreciably better than the *best* result obtained with a static deadline (in conjunction with the OPB). Figure 47 shows that this is because OptiMOS tries to compensate the upstream delay so as to decrease the end-to-end delay.

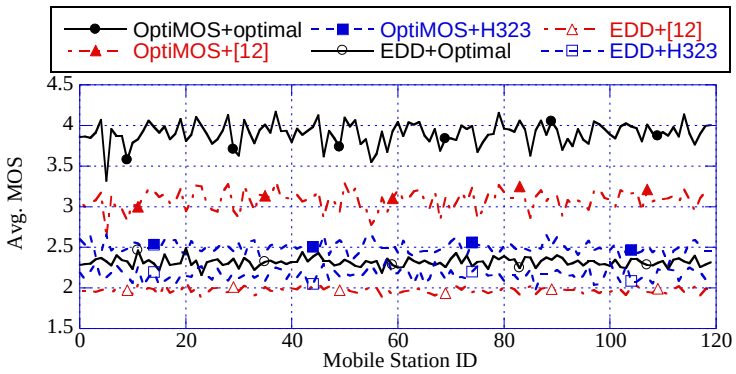

Figure 46 - Comparison between OptiMOS and a fixed 100ms deadline with several PO buffer - VoIP traffic only

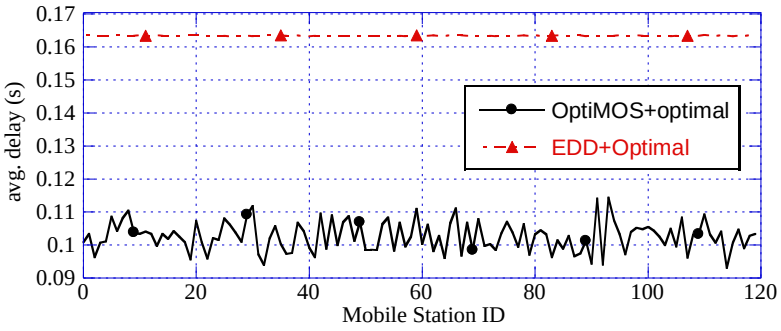

Figure 47 - Mouth to ear delay for the speech frames

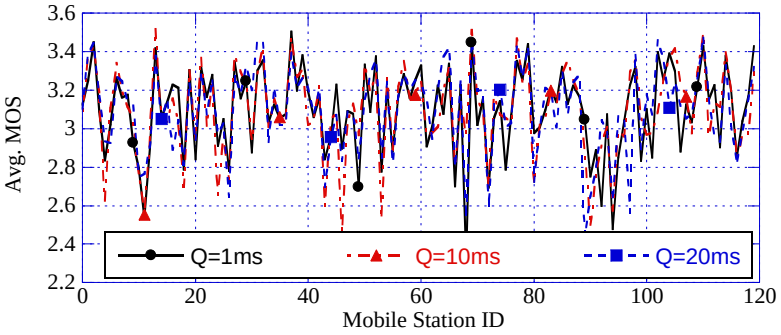

Figure 48 - Performance of OptiMOS with different delay quantization intervals

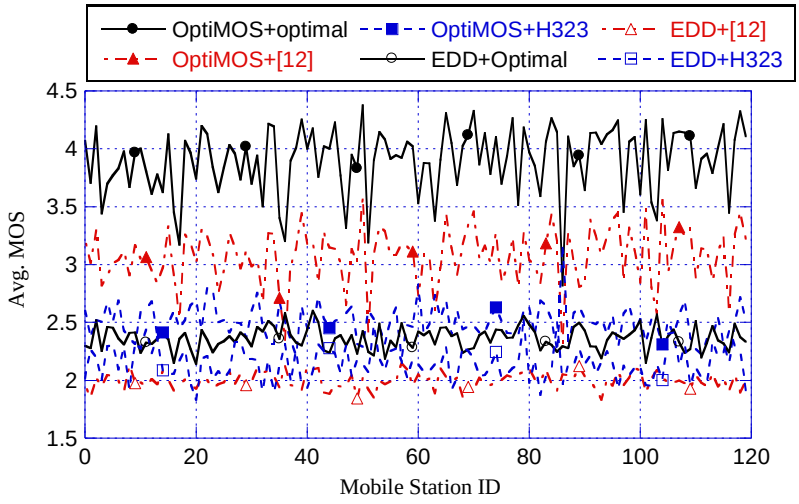

Figure 49 - Comparison between OptiMOS and a fixed 100ms deadline with several PO buffers - VoIP and FTP

The above simulations are carried out with a delay quantization interval $Q = 1ms$ for OptiMOS. Figure 48, which reports results with the PO buffer in [21], shows that increasing Q (up to 20ms) has a negligible impact on the performance of OptiMOS. The results for the other PO buffers are qualitatively similar and omitted for the sake of readability.

The situation does not change appreciably if other traffics are brought into the picture. In Figure 49 we report the same analysis, adding 80 FTP sources, which transmit infinite length files to the mobile users. FTP packets are given an “infinite” deadline, i.e. they are transmitted at a lower priority than VoIP traffic. As the figure shows, the only appreciable effect is a slight increase in the variability of the MOS among the users.

9. A LINK SCHEDULING ALGORITH FOR LTE-ADVANCED NETWORKS

In this chapter we describe a link-scheduling algorithm for a LTE-A multi-hop network infrastructure. The considered network architecture has a two-hop levels configuration and we assume users are only connected to relay nodes (no users are directly connected to the DeNB). We assume that users always have a more favourable condition in two-hop transmission through their anchor relay node, than in a single-hop transmission where they directly address the DeNB. A general representation of the network is shown in [Figure 50]

The number of relay nodes and the number of users is supposed to be configurable and the considered scenario is always on a single cell (single DeNB).

In this study we assume that the traffic direction is uplink, however it will be shown also that the algorithm we propose can be easily applied to the downlink case.

9.1.1. Relay Duplexing Problem

Relay nodes cannot receive and transmit at the same time using the same frequencies because this would result in strong interference. This physical limitation requires access and backhaul links being scheduled on separate resources. Two different approaches are possible:

- FDD relaying where access and backhaul are assigned different frequency resources but overlapping time resources
- TDD relaying where access and backhaul are assigned the same frequencies (i.e., full spectrum availability) and the separation is made in time domain

In the following we will assume to use TDD relaying as a duplexing mechanism.

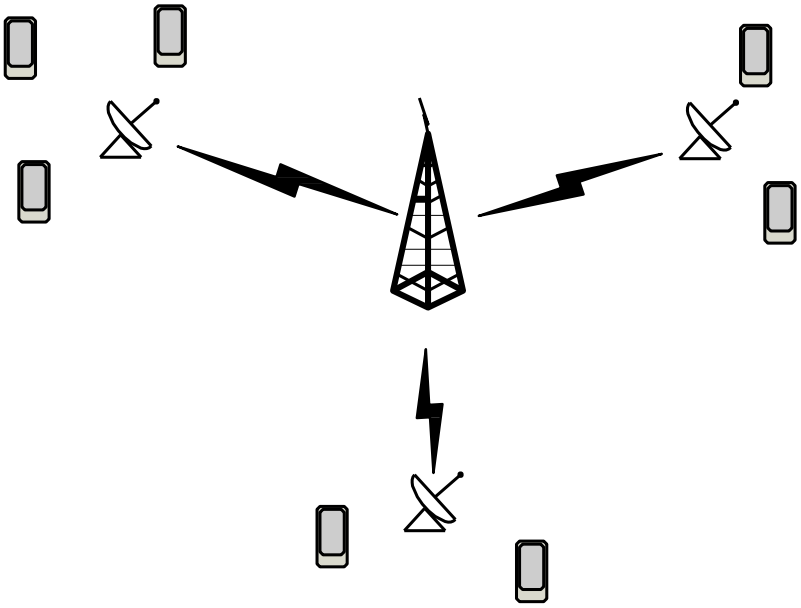

Figure 50 - Relay Scenario Setting

9.1.2. Algorithm objective

We assume to use LTE frame structure type 1 (LTE FDD mode), thus resources are assigned to access and backhaul links with LTE sub-frame granularity (the duration of a sub-frame is one TTI).

The link scheduling algorithm being studied must decide, for each relay, at each TTI, which link, between access or backhaul, is more profitable to have the sub-frame assigned, in order to maximize cell throughput, while meeting the interference constraint.

In LTE-A operations, the relay node is seen as a Base Station from the users' perspective and it is seen as a user terminal from the DeNB perspective. A relay node receives its resources from the DeNB and decides how to assign them to its users. In this context the link-scheduling algorithm must be centralized and executed by the DeNB.

9.1.3. Feedback assumptions

According to LTE uplink operations, user equipment and relay nodes must send feedback control information to their anchor relay node and DeNB respectively. For the execution of the algorithm Buffer Status Reports and uplink channel quality information must be available (the latter is estimated using reference signals). For convenience sake, in the following, the **U**plink **C**hannel **Q**uality will be referred with the term UCQ, and expressed as the number of bytes the channel allows to transmit into a single resource block (effectively the modulation and coding scheme that is suitable for the channel condition). This channel measurement is assumed to be wide-band.

Relay nodes have full users' feedback availability (channel quality and BSRs for each connected user). This knowledge is useful also at the DeNB to take the link scheduling decision, so we assume that aggregate control information summarizing users' status can be sent from the relay node to the DeNB. The use of this signalling is currently not mentioned in LTE rel. 10 standard documents.

9.1.4. Link-scheduling-problem formulation

The backhaul sub-frame is the system bottleneck as it is a single sub-frame that must withstand all of the transmissions that come from the access links, where a dedicated sub-frame is available for each relay node. Managing this resource is critical in order to obtain the highest cell throughput.

For this reasons, when the backhaul links are activated, it could be disadvantageous to let all of the relay nodes (or some of them) to transmit altogether. The amount of resources each relay node would dispose of (i.e., the available portion of the shared sub-frame), could be such that it would be more profitable for some of them to activate the access links, as this decision would generate more throughput.

This problem must be addressed by properly identifying those relay nodes to which backhaul resources allocation (or revocation) is beneficial from the throughput perspective.

The following unwanted scenarios underline the importance of scheduling access and backhaul activity properly:

• Backhaul resources are no longer available and a relay node must activate its access links where it has poor capacity

• Backhaul resources have been assigned to a relay node which had a considerable capacity in its access portion of the network as well, resulting in a low-gain decision (the previous situation may happen as a direct consequence of this one)

The proposed algorithm is able to avoid both situations.

The solution to the link-scheduling problem is at first given using an analytical formulation, which leads to optimal results. Then, to make the process feasible, an heuristic approximation of the optimal model is proposed.

9.1.5. Analytical formulation

As a first approach, the problem can be dealt with analytically and an optimal solution (in terms of optimal access/backhaul interleaving) can be found, given the input parameters at a certain TTI.

The integer linear programming problem that leads to the optimal interleaving pattern for a certain TTI takes the following parameters as input:

K	Number of relay nodes in the cell
bsr_relay_i	BSR of each relay in bytes (i over the set K)
ucq_relay_i	UCQ of each relay (i over the set K)
Q_i	Queue size for each relay in bytes (i *over* the set K)

num_upr_i Number of users per relay (i over the set K)

$bsr\_ue_{i,j}$ BSR for each user of each relay in bytes (i over the set K, j over the set num_upr_i)

$ucq\_ue_{i,j}$ UCQ for each user of each relay (i over the set K, j over the set num_upr_i)

B The number of resource blocks in a frame

The variables are:

tx_relay_i Number of bytes relay i will transmit to its DeNB this TTI (integer)

$tx\_ue_{i,j}$ Number of bytes user j will transmit to its relay i (integer)

b_i Behavior of each relay: 1 if relay i will transmit, 0 if relay i will receive (binary)

The objective function to be maximized, is the cell throughput:

$$\max \sum_{i=1}^{k} (tx\_relay_i - \sum_{j=1}^{upr_i} tx\_ue_{i,j})$$

The backhaul frame space utilization is subject to the following constraint, which expresses that the sum of blocks needed for relay nodes to transmit cannot exceed the total number of available blocks:

$$\sum_{i=0}^{k} \left\lceil \frac{tx\_relay_i}{cq\_relay_i} \right\rceil \leq B$$

A similar condition applies to access frame space: all of the users of a certain relay share the number of available blocks for their transmission:

$$\forall i, \sum_{j=1}^{upr_i} \left\lceil \frac{tx\_ue_{i,j}}{cq\_ue_{i,j}} \right\rceil \leq B$$

The total amount of bytes a relay can transmit cannot exceed its queued data:

$$\forall i, tx\_relay_i \leq \ bsr\_relay_i \cdot b_i$$

The corresponding constraint applied to users' queued data is:

$$\forall i, \forall j \in upr_i, tx\_ue_{i,j} - tx\_relay_i + bsr\_relay_i \leq Q_i$$

A constraint to prevent users to send bytes to the relay while its queue is full is also required:

$$\forall i, \sum_{j=1}^{upr_i} tx_{ue_{i,j}} - tx_{relay_i} + \ bsr_{relay_i} \leq Q_i$$

9.1.6. Idealizations

This analytical model produces the optimal link scheduling solution at each TTI.
It is supposed though, that the DeNB, which is the node where the solving process must take place, is aware of the status of the whole cell and particularly of channel and buffer status of every user terminals. Gathering the required data to get this cell-extended knowledge on the DeNB, is an unfeasible process, because a considerable amount of resources would be required for its distribution.
Moreover, solving the binary optimization problem within a single TTI would be a time-challenging task.
For the above reasons the analytical approach is unsuitable for a real environment application, but through the examination of this model, some key elements to produce a feasible heuristic algorithm can be extracted.

The analytical model is also useful to evaluate the heuristic algorithm performances:

it gives the optimal interleaving of access and backhaul links activation that leads to the maximum throughput, so heuristic algorithm results will be compared against the optimal ones in order to understand if the approximations introduced to make the link scheduling decision process feasible are acceptable.

9.1.7. Proposed Algorithm

We designed and evaluated a distributed algorithm, running on both eNodeBs and Relays, derived from the analytical model shown in section B, whose aim is to maximize the system throughput and therefore to exploit its full potential. At each transmission time interval (TTI, 1ms in LTE-A) and for each relay node, the decision whether to activate access or backhaul links is based on the throughput they could generate, given a certain backhaul resources availability.

The algorithm executes in three main steps:

• *Access step*, performed by each relay node, during which users feedbacks about their experienced channel conditions and data backlogs are gathered. From those data the access capacity is computed and signalled to the eNodeB

• *Backhaul step*, performed by the eNodeB, during which backhaul capacities are evaluated taking into account all Relay nodes current channel qualities.

• *Link Scheduling step*, performed by the eNodeB, during which the link scheduling decision is made according to the two previously obtained capacities by comparing them.

9.1.8. Access Step

Relay nodes have a complete knowledge of their connected-users status (buffer and channel status). This knowledge comes from feedback information that each relay node receives according to LTE procedures. The DeNB also needs to be acquainted on users status to take a proper link scheduling decision.
In the access step of the algorithm then, each relay node receives feedback information from its users and aggregates it into a summarized control information (it will be referred with the term access capacity in the following) to be signalled to the DeNB. The task of the access step then, is to compute and provide the access capacity feedback on each relay node.

In the proposed algorithm each relay node, at each TTI, aggregates users feedbacks by the execution of a maxSINR resource scheduler, which outputs the total capacity (expressed in bytes) that the access links of this relay are capable of. In a nonspecific manner access capacities can embody different meanings depending on the objective that must be reached with their employment (e.g., throughput maximization as in this excerpt). So any algorithm can be applied to produce them.

9.1.9. Access Capacity Computation Algorithm

The access step routine, which is executed on each relay node, considers each physical resource block of the sub-frame and, for each user of this relay, computes its capacity on this single block usage. The amount of bytes the considered user can allocate in this single resource block is obtained as the minimum between its UCQ (bytes/block) and its BSR (bytes). Then, the user with the highest capacity is selected and the obtained value is ceiled with the available relay queue space (bytes). This PRB is assigned to the user and the ceiled block capacity is added to the access capacity for this relay. At last, variables must be updated for next loop iteration (e.g., buffer-status-variable of the user must be decremented with the assigned capacity, relay-queue-space variable must be decremented as well). The algorithm is quite simple because, if we consider any user terminal, we can assume that there is no reason why it won't transmit data unless its buffers are empty.

9.1.10. Backhaul Step

In this step the DeNB receives standard feedback, plus the extra signalling about access capacities, from each relay node. This latter information about access links status is not supported in current LTE rel. 10 standard definitions, but it is a key element for the DeNB to be consistent on evaluating and producing a link scheduling decision.

The DeNB must assign backhaul resources to relay nodes while meeting the following conditions:

• a relay node receives resources only if, with that assignment, it will have a backhaul capacity higher than its access capacity (we refrain from wasting backhaul resources, if it can be known in advance that the backhaul link won't be activated for this relay node)

- The backhaul sub-frame is used efficiently
- Backhaul resources are assigned to a particular relay node while taking care that the choice is beneficial for cell throughput increment (the throughput gain that the assignment implies must be evaluated and taken into account)

After backhaul capacities have been estimated for each relay node, the link-scheduling decision is taken by simply comparing, for each one of them, its access and backhaul capacity and activating links accordingly (link with higher capacity wins).

9.1.11. Backhaul capacities computation algorithm

Because of the constraints listed above, it is necessary to define an ordering policy so that the highest priority is given to those relay nodes towards which resources assignment produces the highest gain in terms of cell throughput. By sorting relay nodes accordingly we can ensure that relay nodes, which are the most valuable for transmission, have a larger amount of resources at their disposal.

The inputs are relay nodes BSRs, relay nodes UCQs and relay nodes access capacities.

As a first step relay nodes which are incapable of getting a higher backhaul capacity than their access one, even if the whole remaining sub-frame is allocated to them, are removed from the list of relay nodes to be served (in the following this list will be referred as "relay list").

Then relay nodes in relay list are ordered on a gain basis and resources are assigned to the relay node with the highest score, which will be subsequently removed from relay list.

In order to reduce PRB fragmentation, the last assigned resource block is deallocated to a relay node if, even without it, the condition of having the backhaul capacity higher than the access capacity is still verified. At the end of the routine then, if some resource blocks are still available, they are assigned to those relay nodes who had been deprived before.
A list containing the residual amount of bytes that were not transmitted due to last block deallocation, is kept to perform the (possible) final reallocation.
This process is repeated until the relay list is not empty and some backhaul sub-frame resources (PRBs) are still available.
The output is a list of backhaul capacities for each relay node, which is evaluated to take the link scheduling decision.
According to what is shown in the pseudo code, three parameters are calculated to establish the relay nodes ordering (and priorities): capacity, gain and score.
The capacity of the relay node i is obtained as:

$$capacity_i = \min(BSR_i, CQI_i * rPRB)$$

from capacity$_i$, relay node gain is computed as:

$$gain_i = capacity_i - acces\_capacity_i$$

9.1.12. Performance Evaluation

For performance evaluation simulations, the following algorithms have been compared:

1. **GainOverBlocks** This algorithm is based on the analytical formulation described in [9.1.5] and embodies the algorithm described in [9.1.7].
2. **Static link-scheduling (*s-alt*)** this algorithm activates access and backhaul links alternatingly at each TTI, for all the relay nodes (e.g., at odd TTIs all the relay nodes use their access links, and at even TTIs all the relay nodes use their backhaul links, as described in [4]). It is defined static because the link scheduling strategy is always the same. When a link is selected for transmission, resources are assigned to nodes on a maxSINR criterion (in order to get the highest throughput).
3. **Static MBSFN patterns** (ID: sMBSFN) the usage of MBSFN patterns to configure sub-frames for DeNB-to-RN downlink transmissions (and, as a consequence, for the corresponding uplink transmissions), has been introduced with the 3GPP TS 36.216 document [6] since version 1.0.0, dated September 2010.

pattern/subframe	0	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0	0	0
3	0	1	1	1	0	0	0	0	0	0
4	0	1	1	1	0	0	1	0	0	0
5	0	1	1	1	0	0	1	1	0	0
6	0	1	1	1	0	0	1	1	1	0

Table 3 - MBSFN Patterns

In these patterns, the sub-frames marked with 1 are backhaul sub-frames, while the sub-frames marked with 0 are access sub-frames. Note that the ID of each pattern is equal to the number of backhaul sub-frames it contains.

In the graphs of the following sections, the number after the algorithm ID is the pattern ID, e.g., sMBSFN6 is used when simulating with MBSFN pattern with id 6.

MBSFN evaluations have been performed under the following hypothesis:

- All relay nodes use the same pattern

- A relay node maintains its pattern for the whole simulation

9.1.13. Traffic types

Performance evaluations have been carried out using the following traffic classes (for each scenario, all users belong to the same traffic class):

Full Buffer (ID: *fbuf*) Full buffer traffic for the user terminals. In the graphs, the number following the ID is the dimension of each user's buffer.

Constant Bit-Rate (CBR) (ID: *cbr*) Constant bit rate traffic. The number following the ID expresses the traffic rate, which, with this traffic class, is measured in *B/TTI.*

CBR on/off (ID: *bursty*) Constant Bit-Rate traffic with on and off periods. The number following the ID is the *rate/TTI* during the on periods (in the following the bursty rate will be expressed in *B/TTI on*).

VoIP (ID: *voip*) VoIP traffic. The trace is built using the Weibull distribution [[21].

9.1.14. Relay Simulations settings

All relay simulations have been made with the configuration described in this section. However, some particular scenarios, which are interesting to emphasize distinctive aspects of certain algorithms, constitute an exception and the configuration used for their modelling is described in the appropriate sections.
The table below contains the values of the main parameters for the base configuration.

TTI	10000
Relay Nodes	5
Number of users	[20, 20, 20, 20, 20]
Sub-frame PRBs	50
Relay queues size	4000B
Relay channel status	LOS

Table 4 - Relay simulation settings

The LOS condition for relay-nodes channels means that only the three highest (out of fifteen) UCQ values (corresponding to the MCSs which can carry the highest amount of data) are considered when the relay channel status is updated.
As for the users connected to each relay node, in the base configuration they are grouped into four categories depending on their channel status

1. **Good (channel) users**: users with high channel quality, where high means that they are more likely to use higher UCQs values than standard users. Terminals roaming at a short distance from the relay node compose this group of users. In the base configuration this group has three users for each relay (with ID 1, 2, and 3)
2. **Bad (channel) users**: users with low channel quality, where low means that they are more likely to use lower UCQs values than standard users.

These are intended to be cell-edge users. In the base configuration this group has two users for each relay (with ID 4 and 5)

3. **Quick users**: users with high mobility whose channel quality is more change- able and ranges the whole set of possible UCQ values. In the base configuration this group has two users for each relay (with ID 6 and 7)
4. **Standard users** (in the base configuration this group has thirteen terminals for each relay)

In a scenario configuration, the relay queue size is a very important parameter, which prevents users to deliberately transmit data to their relay node if its buffers are full. Its role is particularly emphasized in the analytical formulation (9.1.5) and in the access step routine (9.1.8). The value of 4000B is suitable (according to the number of PRBs and the UCQ values used in the simulator) to provide enough data for backhaul transmissions and to avoid excessive buffering at the relay nodes.

The usage of 50 PRB for each sub-frame corresponds to a spectrum allocation bandwidth of 10Mhz.

The configuration used for a certain scenario is summarized in the scenario ID with the following syntax: trafficClass-TTIs, e.g., cbr44-10000 stands for 10000 TTIs were simulated with CBR 44Bps traffic.

9.1.15. Simulation Results

In this section, the performance of algorithms described in 9.1.12 is evaluated
The throughput graphs for all the algorithms in the different scenarios are shown in Figure 51. Even if the sAlt algorithm has a good performance it can't take decisions adaptively to the users channel conditions. About MBSFN patterns, we can notice that in all the simulated scenarios the traffic saturates the cell, but the sMBSFN6 pattern always has a better performance than the others. This confirms the important role of the backhaul sub-frame as a system bottleneck: it must carry the whole amount of traffic generated on the access links, where an access sub-frame is available for each relay node. Then, if the users traffic is high, it is reasonable to expect that a more frequent scheduling of the backhaul link lead to a higher throughput. As a confirmation, we can observe from the graphs that the higher is the number of backhaul sub-frames in the MBSFN pattern, the higher is the throughput. As a conclusion, the dynamic link scheduling solution, is a much more profitable choice than the static ones.

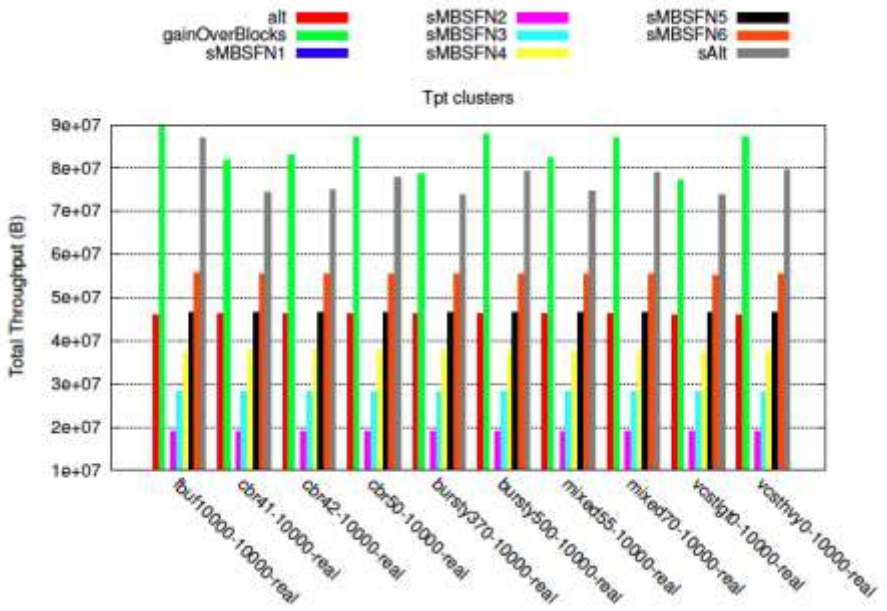

Figure 51 - Throughput performance

10. CONCLUSIONS

In this work we carried out a simulation study of the LTE and LTE Advanced technologies. We evaluated the systems performances under different traffic scenarios and by varying the values of a set of relevant parameters. Our analysis showed that by designing and implementing mechanisms at the MAC layer of the LTE and LTE systems, they can efficiently support the QoS required by real times traffics such as VoIP or increase the overall system performances.

Specifically:

We presented OptiMOS, a framework for dynamically assigning deadlines to voice flows so as to maximize the user-perceived quality.
OptiMOS simulates what an optimal adaptive PO buffer would do, and assigns deadlines to voice packets so that the latter are transmitted just in time for playout, or dropped if they have already passed their expected playout point. Simulation results show that OptiMOS increases the quality of experience of the users (assessed in terms of average MOS) remarkably, regardless of which adaptive PO buffering scheme they use, and is robust to synchronization errors and quantization of the time resolution in the computations

The link-scheduling problem for relay-enhanced networks in LTE Advanced has been addressed. It has been explained that this entails to regulate a relay node transmission activity (access and backhaul transmissions), in order to avoid overlapping time-frequency resource usage, which would lead to strong interference.
We have shown that a dynamic link scheduling approach, which takes into account the traffic load over the access and backhaul links, is definitely a more effective solution than a static one, which determines the relay nodes transmission activity in disregard of the network condition. Moreover, we have shown that, with the introduction of a little overhead (a control information represented by a number, to be signalled from each relay to the DeNB), it is possible to define a dynamic link-scheduling algorithm whose throughput performance is close to the theoretical optimum.

11. ACKNOWLEDGEMENTS

We would like to thank Vincenzo Pii for his related master thesis, and Prof. Luciano Lenzini and Ing. Giovanni Stea for their valuable support as my Ph.D. advisors. The subject matter of this paper includes description of results of a research project carried out by the Dipartimento di Ingegneria dell'Informazione of the University of Pisa on behalf of, and funded by, Telecom Italia S.p.A., who reserve all proprietary rights in any process, procedure, algorithm, article of manufacture, or other result of said project herein described.

REFERENCES

[1] 3GPP, “Overview of 3gpp release 8 v0.2.2,” January 2011.

[2] C. Gessner, “Umts long term evolution (lte) technology introduction,” Sept. 2008.

[3] 3GPP, “The 3gpp home page.” http://www.3gpp.org/ .

[4] R. Schoenen, R. Halfmann, and B. H. Walke, “An FDD Multihop Cellular Network for 3GPP-LTE,” VTC Spring 2008 - IEEE Vehicular Technology Conference , pp. 1990–1994, May 2008.

[5] R. Schoenen, W. Zirwas, and B. Walke, “Capacity and coverage analysis of a 3GPP-LTE multihop deployment scenario,” in Communications Workshops, 2008. ICC Workshops 08. IEEE International Conference on , pp. 31–36, Ieee, May 2008.

[6] 3GPP, “3gpp ts 36.216 physical layer for relaying operation,” December 2010.

[7] Qualcomm Europe, “Specifying blank subframes for efficient support of relays.” 3GPP document R1-083817, October 2008. TSG-RAN WG1 #54bis, Prague.

[8] Ericsson, “Efficient support of relays through mbsfn subframes.” 3GPP document R1-084357, November 2008. TSG-RAN WG1 #55, Prague.

[9] I. WINNER, “IST-4-027756 Final assessment of relaying concepts for all CGs scenarios under consideration of related WINNER L1 and L2 protocol functions,” 2007.

[10] M. Kaneko and P. Popovski, “Radio resource allocation algorithm for relay aided cellular OFDMA system,” in ICC’07. IEEE International Conference on Communications, 2007 , pp. 4831–4836, IEEE, 2007.

[11] W. Nam, W. Chang, S. Chung, and Y. Lee, “Transmit optimization for relay-based cellular OFDMA systems,” in ICC’07. IEEE International Conference on Communications, 2007 , pp. 5714–5719, IEEE, 2007.

[12] M. Andreozzi, G. Stea, A. Bacioccola, R. Rossi, “Flexible Scheduling for Real-Time Services in High-Speed Packet Access Cellular Networks”, in Proc. European Wireless 2009, Aalborg, DK, May 17-20 2009

[13] D. Veitch, J. Ridoux, S. Babu Korada, “Robust Synchronization of Absolute and Difference Clocks over Networks”, IEEE/ACM Trans. on Networking, 17(2): pp. 417-430, Apr. 2009.

[14] The Network Time Protocol website, http://www.ntp.org/

[15] IEEE 1588 Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, http://ieee1588.nist.gov

[16] G. M. Garner et al. ”IEEE 802.1 AVB and Its Applicatin in Carried-Grade Ethernet”, IEEE Comm. Magazine, Dec. 2007, pp.126-134

[17] ITU-T Recommendation G.107. The Emodel, a computational model for use in transmission planning. Dec. 1998.

[18] ITU-T Recommendation G.108. Application of the Emodel: A planning guide. Sep. 1998.

[19] ITU-T Recommendation G.113. Transmission impairments due to speech processing. Feb. 2001.

[20] Z. Qiao, L. Sun, N. Heilemann, E. Ifeachor: “A new method for VoIP quality of service control use combined adaptive sender rate and priority marking”, in Proc. IEEE ICC 2004, Paris, France, June 20–24, pp. 1473–1477.

[21] A. Bacioccola, C. Cicconetti, G. Stea, “User level performance evaluation of VoIP using ns-2", in Proc. NSTOOLS'07, Nantes (FR), Oct. 22, 2007.

[22] L. Atzori, M.L. Lobina, M. Corona: “Playout buffering of speech packets based on a quality maximization approach”. IEEE Trans. on Multimedia 8(2): 420-426 (2006)

[23] S. Khan, S. Duhovnikov, E. Steinbach, M. Sgroi, W. Kellerer, “Application-driven Cross-layer Optimization for Mobile Multimedia Communications using a Common Application Layer Quality Metric”, in Proc. IWCMC’06, Vancouver, CA, July 3-6, 2006

[24] A. Saul, G. Auer, “Multiuser Resource Allocation Maximizing the Perceived Quality”, EURASIP Journal on Wireless Communications and Networking, Vol. 2009, Ariticle ID 341689

[25] P. Ameigeiras, *et al.*, “QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems”, Computer Communications, 33(5): 571-582, March 2010

[26] http://www.h323plus.org/, accessed Dec. 2009