Survey in two chapters for Hermes

WP2 and WP4 members

22nd January 2008

Contents

The wireless metropolitan network standard for WiMAX, IEEE 802.16, defines various high speed mechanisms that provide wireless last mile broadband access in Metropolitan Area Networks (MANs) at a cost much lower than traditional cable, DSL or T1 technologies. A typical scenario for the use of WiMAX is for it to provide broadband Internet access to various users in one or more buildings via rooftop antennae. This emerging technology provides a very attractive alternative to the 3G technology which is based on cellular networks. The low cost of WiFi deployment is obtained at the cost of much smaller coverage. The WiMAX is part of a global standardization effort of the IEEE that involves not only the local WiFi networks (IEEE 802.11) but also regional networks (IEEE 802.22).

WiMAX, supports several advanced techniques, such as Adaptive Antenna System (AAS), Adaptive Modulation and Coding (AMC) and Convolutional Turbo Code (CTC). IEEE 802.16e defines five QoS classes: i. Unsolicited Grant Service (UGS) for constant-bit-rate, delay-and-jitter-sensitive applications such as Voice over IP, ii. real-time Packet Service (rtPS), also specified for streaming applications but with higher priority on all other classes, iii. Extended rtPS (ErtPS) adds a bound on the jitter, iv. non-rtPS (nrtPS) for elastic applications and v. best-effort (BE).

The implementation of these QoS classes takes place at the MAC layer via a classifier and a scheduler. It operates at the flow level, defined by a service flow ID and a connection ID pair, uplink or downlink direction and a set of QoS metrics.

This general survey is composed of two parts which correspond to workpackages WP2 and WP4 of the WINEM project.

In chapter 1 we survey issues related to information-theoretic formulation of capacity in IEEE802.16 WiMAX, with a focus on the PHY layer. Many optimization problems can be stated due to the numerous possible choices in the optimization criterion (minimum sum power, maximimal sum rate, fairness and all its variants, ...), which come with related algorithms. In chapter 2, we present WiMAX capacity under the networking point of view, and related radio resource management issues. The bulk of scheduling and CAC solutions proposed by researches for IEEE 802.16 is being presented. We first provide an overview of the main features proposed by the standard to support QoS and then outline the challenges that should be addressed when designing a new scheduling or CAC solution. Along with the description of each proposal, a comparison outlining the advantages and limits of each solution is being presented. Then a focus is made on the MAC layer and an Erlangian approach of system capacity optimization is presented.

For an easier reading a list of all acronyms used in this report is provided on the following page.

Acronyms

AAS : Advanced Array Systems
AMC : Adaptive Modulation and Coding
ARQ : Automatic Repeat-reQuest
BE : Best Effort
BLER : BLock Error Rate
BRGM : Bandwidth Request-Grant Mechanisms
BS : Base Station
CAC : Call Admission Control
CBR : Constant Bit Rate
CCI : CoChannel Interference
CDMA : Code Division Multiple Access
CID : Connection ID
CoS : Class of Service
CP : Common Part
CS : Convergence-Specific
CSI : Channel State Information
CTMC : Continuous Time Markov Chain
DCCP : Datagram Congestion Control Protocol
DL : DownLink
DRR : Deficit Round Robin
DSL : Digital Subscriber Line
EDF : Earliest Deadline First
ertPS : Extended Real-Time Polling Service
FDD : Frequency Division Duplexing
FEC : Forward Error Correction
FET : First Exit Time
FFT : Fast Fourier Transform
FQ : Fair Queuing
FUSC : Fully Used SubChannelization
GPRS : General Packet Radio Service
HARQ : Hybrid Automatic Repeat-reQuest
LOS : Line-Of-Sight
LR : Latency-Rate
MAC : Media Access Control
MANs : Metropolitan Area Networks
MCS : Modulation and Coding Scheme
MDT : Mean Delay Time
MIMO : Multiple Input, Multiple Output
MTU : Maximum Transmission Unit
NLOS : Non-Line Of Sight
nrtPS : Non-Real-Time Polling Service

OFDM : Orthogonal Frequency Division Multiplexing
OFDMA : Orthogonal Frequency-Division Multiple Access
PDUs : Protocol Data Units
PHY : Physical Layer
PMP : Point-to-MultiPoint
PS : Processor Sharing
PUSC : Partially Used SubChannelization
QoS : Quality of Service
RR : Round-Robin
RTO : Retransmission TimeOut
rtPS : Real-Time Polling Service
RTT : Round-Trip Time
SC : Single Carrier
SCTP : Stream Control Transmission Protocol
SDUs : Source Data Units
SF : Service Flow
SFID : Service Flow Identifier
SINR : Signal-to-Interference-and-Noise-Ratio
SNR : Signal-to-Noise-Ratio
SSs : Subscriber Stations
STBCs : Space Time Block Codes
TCP : Transport Control Protocol
TDD : Time Division Duplexing
TDMA : Time Division Multiple Access
TFRC : TCP-Friendly Rate Control
UGS : Unsolicited Grant Services
UL : UpLink
WiFi : Wireless Fidelity (IEEE 802.11b)
WiMAX : Worldwide Interoperability for Microwave Access
WRR : Weighted Round-Robin

Chapter 1

Information theoretic capacity of WiMAX

Tijani Chahed (GET/INT), Laura Cottatellucci (Eurecom), Rachid Elazouzi (LIA), Sophie Gault (Motorola), Alberto Suarez Real (Eurecom)

1.1. System description

The first version of WiMAX, for fixed broadband access in the 10-66 Ghz range, was started in 1998 and was completed in October 2001. It was amended in version 802.16a to behind 2-11 GHz in January 2001. Version 802.16d, completed in January 2004, brings some enhancements in the uplink. Version 802.16e is mainly about mobility and asymmetric links.

The WiMAX Forum [WiMb] was created to promote inter-operability between proposals and products which produce many options in PHY and MAC layers, both in licensed ranges (2.5-2.69 and 3.4-3.6 Ghz) and unlicensed ones (5.725-5.85 Ghz) [GHO 05]. And hence a confusion on performance, in terms of capacity/rates and coverage.

The design of 2-11 GHZ PHY layer is driven by Non-Line Of Sight (NLOS) communications. Standards a and d define three possible access mechanisms : SC (Single Carrier), OFDM (256 carriers, with multiple access based on TDMA) and OFDMA (2048 carriers, multiuser OFDM by giving a subset of carriers to individual users).

A variable channel bandwidth, an integer multiple of 1.25, 1.5 and 1.75 MHz with a maximum of 20 MHz, has been adopted for global implementation. But this choice is being narrowed down by WiMAX Forum.

1.1.1. *Subchannelization*

This refers to the dedicated allocation of blocks of subcarriers to users and not single subcarriers, following either a distributed mode, which improves frequency diversity and robustness, or the adjacent mode, which increases multiuser diversity.

In the distributed subcarriers allocation, full channel diversity is obtained by distributing the allocated subcarriers to subchannels using a permutation mechanism. This mechanism is designed to introduce frequency diversity, thus minimizing the performance degradation due to fast fading which is characteristic of mobile environments. In addition to that, WiMAX standards [802 04, 802 05] specify two different distributed allocation modes: The FUSC (Fully Used Subchannelization) mode where all subcarriers are used to form subchannels in each cell, and the PUSC (Partially Used Subchannelization) mode where the frequency band is divided into three segments.

For illustration, with an FFT size of 1024 and after reserving the pilot and guard subcarriers, a FUSC allocation will correspond to 16 subchannels of 48 data subcarriers each, while a PUSC allocation will correspond to 30 subchannels, each containing 24 data subcarriers. Note that assigning subcarriers to subchannels in PUSC is a bit complicated, as it employs two permutations:

– An outer permutation divides the subcarriers into six major groups of clusters using a specific renumbering sequence.

– An inner permutation operates separately on each major group, distributing subcarriers to subchannels within the group and is based on the FUSC permutation with distinct parameters for the odd and even major groups.

This is illustrated in Figure 1.1, where two groups are assigned to one segment corresponding to a sector of the cell. Note that a segment can also be allocated to a cell in an omni-directional setting.

The adjacent allocation corresponds to the WiMAX AAS (Advanced Array Systems) mode, designed to support MIMO techniques and adaptive modulation. Note that, in order to achieve a frequency diversity, mobiles using adjacent allocation may hop rapidly between different subchannels during their communication times.

1.1.2. *Adaptive Modulation and Coding*

The idea behind Adaptive Modulation and Coding (AMC) is to transmit at the highest possible data rates when the channel is good, and low rates when the channel is bad, so as to avoid excessive dropped packets.

Versions a and d define 7 modulation and coding combinations, for robustness versus rate trade-offs, depending on channel and interference conditions. These are

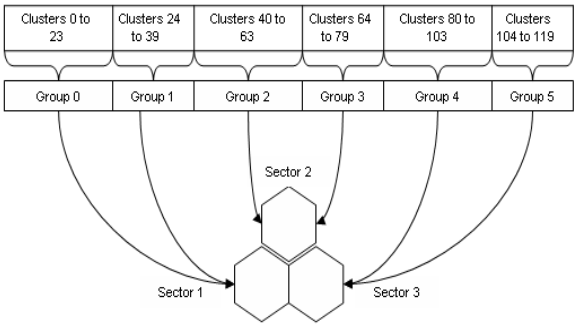

Figure 1.1. *Construction of groups and segments in the PUSC allocation mode.*

the same as in 802.11 a and g, the only difference being that WiMAX uses outer Reed-Solomon code concatenated in an inner convolutional code. Interleaving is used to reduce the effect of error bursts. Turbo coding is optional (increases capacity but also delay and complexity). The difference between the UpLink (UL) and DownLink (DL) is in the length of the preamble (to help the receiver with synchronization and channel estimation): it is short in the UL, long in the DL [GHO 05].

The AMC controller tunes transmit power, transmit rate (constellation) and coding rate, as a function of Signal-to-Interference-and-Noise-Ratio (SINR). The performance depends on many factors : BLock Error Rate (BLER), ARQ/HARQ and power control vs. waterfilling for instance (more power to stronger channels, not always true in practice where some savings are possible [AND 06]).

1.1.3. *Diversity*

Four types of diversity exist in OFDMA-based WiMAX : multiuser diversity (between users), spatial diversity, frequency diversity (between subcarriers) and time diversity (by allowing latency).

Note that although these diversities bring gains in capacity, they are not necessarily additive. For instance, multiuser diversity gain reduces in WiMAX because of spatial diversity and the need to assign users contiguous blocks of subcarriers.

1.1.4. *MAC functionalities*

The MAC layer is composed of Convergence-Specific (CS) and Common Part (CP) sublayers. CS maps transport layer specific traffic to flexible, any-traffic MAC layer units. CP is responsible for fragmentation of MAC SDUs into PDUs, QoS, scheduling and retrasmissions. Details of scheduling and reservation management are left undefined in the standard.

The MAC frame structure is as follows (see Reference [KIM 05b]). The MAC frames are composed of two main TDMA subframes, one for the downlink and an other one for the uplink. For the needs of the connectivity study, we focus only on the uplink subframe, itself partitioning in four TDMA subframes. The first three are reserved for the CAC : initial and maintenance connection. The last one carries the data transmission through numerous time slots. The whole capacity is greatly improved by using, for all these subframes, several OFDMA frequencies. Moreover, the ranging intervals can manage a large number of contending connection because each of the three time slots uses CDMA technique. This allows to share the channel resources through all contending nodes as well as minimize the collision probability. Figure 1.2 shows all these specificities. On this, only a single time slot and its OFDMA subchannels concerns the CAC process. the figure 1.2 sketches it as "Periodic Ranging and Bandwidth Request Ranging interval". All our study aims at characterizing the arrival, collision and queuing process for the requests which arrives in this interval.

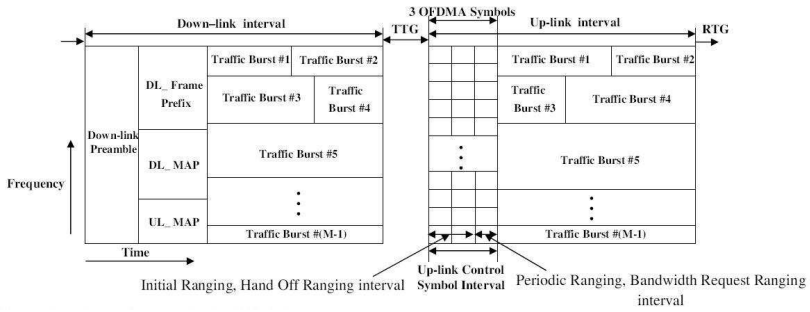

Figure 1.2. *IEEE 802.16e MAC frame format*

The bandwidth request principles are as follows. Once a node is roamed to its cell, it can engage a bandwidth request. The procedure depends on the node state : if the node is silent, it uses the contention time slot in the Bandwidth Request Ranging Interval. Else, when the node is still transmitting, the request is achieved by using

an aggregate or incremental bandwidth request in its data reserved time slots. The incremental one is required when a node needs more resources. The other one allows to reevaluate, often periodically, the node needs.

As soon as a node wants to send data, it chooses one of the N codes composing the dedicated bandwidth request code family, and proceeds to its demand by transmitting its coded request through the bandwidth request ranging interval. These requests follow a backoff process in case of collision in the selected code. A collision occurs if two or more nodes have chosen the same code in the same ranging interval.

Note that the communication way used in the IEEE802.16e standard is far more complex than any other wireless communication. Thus, the base station (BS) has to manage, CDMA coding and decoding, resource allocation, flow scheduling, etc, from one TDMA frame to the following one. So, the incoming connection requests wait some MAC frames before receiving any response. In fact, the mobile waits its bandwidth response until a timeout threshold. The IEEE802.16e standard version defines the timer T_3 as the maximum MAC frame number that a contending node can wait before considering that its request has been lost on the wireless channel, or in the BS request queue.

1.1.5. *Optional features*

Space Time Block Codes (STBCs) are optional and can be implemented in the DL to provide increased diversity. Its implementation is quite probable in WIMAX as the latter shall adopt two-antenna transmit diversity using Alamouti code. Receive diversity is also envisioned as it requires no extra transmission effort [good for cooperation].

Intelligent, ie, adaptive, antennas are optional (to improve spectral efficiency of the system). Point-to-multipoint (PMP) frames are defined both for the UL and DL.

1.2. Achievable rates and resource allocation in single cells: Problem formulation

This section is dedicated to define the concept of capacity or maximum achievable rate in single cell OFDMA networks and to illustrate the dual problem of minimum transmitted power under target rate constraints. The optimization problem of determining the maximum achievable rate is inherently related to the resource allocation problem. In single-cell OFDMA networks the resource allocation problem consists in assigning the subcarriers to the active users in the system and in determining the corresponding transmit powers.

The problem of maximizing the achievable rate under power constraint and its dual problem of minimizing the transmit power under target rate constraints correspond to optimize the system from two different perspectives: a users' perspective and a network perspective. From the users' perspective, the aim is to achieve a certain QoS characterized by simply a rate (elastic traffic) or a rate and additional delay constraints (rigid traffic) keeping the transmit powers as low as possible. The network perspective aims at maximizing the capacity of the system under maximum power constraints.

The problems consist in allocating both power and bandwidth (subcarriers) so that the constraints on powers or rates are satisfied and at the same time the system is optimized with respect to the network perspective (i.e. the maximum capacity is achieved) or to the users perspective (i.e. the desired rate are achieved with minimum cost in terms of energy). We assume that the problem admits a solution. The existence of a solution should be guarantee by the call admission control performed at higher layers. Thus, in the most general framework the problem consists in jointly allocating subcarriers and powers.

In determining the the fundamental limits of the system we do not take into account the fairness issue, extremely relevant in practical systems. Therefore, in Section 1.2.2 we reformulate the optimization problems enforcing fairness criteria.

1.2.1. *General formulation*

In this section we consider an OFDMA system with K users and N tones in the downlink channel. The base-station and each user are equipped with a single antenna. We assume that the OFDMA system is designed in such a way that each tone has flat frequency response. The channel gain for user k on tone n is denoted by h_{kn} and a system is impaired by additive white Gaussian noise with variance σ_{kn}^2. Let S_k be the set of tones allocated to user k. Each tone is allocated to at most one user, i.e. $S_j \cap S_k = \emptyset$ for $j \neq k$ and $\cup_{k=1}^{K} S_k \subseteq \{1, 2, \ldots, N\}$. Let p_{kn} be the power allocated to user k on tone n and $\gamma_{kn} = |h_{kn}|^2/\sigma_{kn}^2$. The SNR of user k on tone n is $p_{kn}\gamma_{kn}$.

In the rate maximization problem, the total transmitted power is constrained to be not greater than P_{tot} and the objective consists in maximizing the sum rate. The problem can be formulated as follows.

$$
\begin{aligned}
\text{maximize} \quad & \sum_{k=1}^{K} \sum_{n \in S_k} \log_2(1 + p_{kn}\gamma_{kn}) \\
\text{subject to} \quad & \sum_{k=1}^{K} \sum_{n \in S_k} p_{kn} \leq P_{tot}, \\
& S_j \cap S_k = \emptyset \quad \forall j \neq k \\
& \cup_{k=1}^{K} S_k \subseteq \{1, 2, \ldots, N\} \\
& p_{kn} \geq 0 \quad \forall k \text{ and } \forall n
\end{aligned} \tag{1.1}
$$

In the dual power minimization problem, each user requires a minimum transmitting rate R_k and the objective is to minimize the total used power. It can be mathematically formulated as:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{k=1}^{K} \sum_{n \in S_k} p_{kn} \\
\text{subject to} \quad & \sum_{n \in S_k} \log_2(1 + p_{kn}\gamma_{kn}) \geq R_k \forall k \\
& S_j \cap S_k = \emptyset \quad \forall j \neq k \\
& \cup_{k=1}^{K} S_k \subseteq \{1, 2, \ldots, N\} \\
& p_{kn} \geq 0 \quad \forall k \text{ and } \forall n
\end{aligned} \tag{1.2}
$$

The first problem is more appropriate for bursty applications, as data traffic, whilst the second would be more suitable for fixed-rate applications, such as voice traffic.

For the uplink channel, the rate maximization problem can be formulated in a similar way. The unique global power constraint in (1.1) is substituted by a set of individual power constraints, one constraint for each of users. More specifically, we require $\left\{ \sum_{n=1}^{N} p_{kn} = P_k, ; k = 1, \ldots, K \right\}$.

By making use of the duality of the Gaussian multiple-access and broadcast channels [JIN 04] it can be shown that, given a set of minimum required rates, the total energy required is the same for uplink and downlink.

In general, the optimization problems described above are not convex. It is necessary to find the optimal subset of subcarriers for each of the users, and the problem turns into a combinatorial problem with exponential complexity in N.

In order to simplify the problem of resource allocation, two approaches are possible: either to solve the joint optimization problem with a suboptimum approach or to split it into two sub-problems, frequency allocation and power allocation. In the uplink, both power allocation and subcarriers assignment can be done in a centralized or in a distributed way.

1.2.2. *Fairness*

It may be advisable to consider fairness criteria to perform the resource allocation. In fact, the sum rate maximization techniques assign subcarriers to the users with the best channel gain, and when path loss gaps among users are large (likely scenario in a wireless environment), most of the resources are assigned to a small subset of users, and the ones that experience low channel gains may receive no data. Several different optimization criteria can be adopted to enforce a fairer behaviour of the system. The most relevant fairness criteria are illustrated in this section.

– Max-min problem. The objective is to maximize the worst user capacity. The problem is formalized as follows

$$\begin{aligned} &\text{maximize} && \min_k \sum_{n \in S_k} \log_2(1 + p_{kn}\gamma_{kn}) \\ &\text{subject to} && : \text{same constraints as in (1.1).} \end{aligned}$$

This formulation provides maximum fairness between users, but it is not well suited to scenarios with users requiring different rates corresponding to different service levels.

– Proportional fairness. The objective is still the maximize sum capacity, but a set of constraints is imposed to guarantee that proportional rates among the different users are maintained for each channel realization. The problem is formalized as follows

$$\begin{aligned} &\text{maximize}_{p_{k,n}} && \sum_{k=1}^{K} \sum_{n \in S_k} \frac{1}{N} \log_2(1 + p_{kn}\gamma_{kn}) \qquad (1.3) \\ &\text{subject to:} && \text{same constraints as in (1.1),} \\ & && \text{and the additional constraints} \\ & && R_1 : R_2 : \ldots : R_K = \delta_1 : \delta_2 : \ldots : \delta_K \\ & && \text{where } \{\delta_i\}_{i=1}^{K} \text{ is a set of fixed values} \\ & && \text{ensuring proportional fairness among users.} \end{aligned}$$

– Hard Fairness. Within this strategy each user transmits at his own desired rate, independently of the actual channel realization. Indeed, this formulation corresponds to the minimum sum power problem.

1.2.3. *Unified approach*

Let us introduce the utility function $U_k(r)$ of user k. The utility function is required to be a nondecreasing function of the data rate. A unified framework for the above mentioned problems in the downlink can be formulated as follows

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{K} U_k(R_k) \\ \text{subject to} \quad & R_k = \sum_{S_k} \log_2[1 + p_{kn}\gamma_{kn}] \\ & \sum p_{kn} \leq P_{tot}, \quad p_{kn} \geq 0 \end{aligned} \tag{1.4}$$

The general problem statement in (1.5) boils down to the classical sum rate maximization problem when $U(r) = r$. When fairness is introduced the slope of the utility curve should decrease as the data rate increases. This property prevents from assigning the most of resources to a small subset of users with high channel gains. If $U(r) = \log(r)$ the unified framework (1.5) reduces to a proportional fairness model similar to the one considered in (1.3). Utility functions of the form $U(r) = -\frac{r^{-\alpha}}{\alpha}, \quad \alpha > 0$ can also be considered. The parameter α determines the degree of fairness. Stricter fairness requirements are enforced as α increases. The max-min fairness is obtained by letting $\alpha \to \infty$.

In the following, we assume a continuous spectrum of subcarriers. We denote by D_k the frequency band assigned to the user k and we assume that bands assigned to different users are not overlapping. Then, the unified resource allocation framework is formulated as follows

$$\begin{aligned} \text{maximize} \quad & \sum_{k=1}^{K} U_k(R_k) \\ \text{subject to} \quad & R_k = \int_{D_k} \log_2[1 + \beta p(f)\gamma_k(f)]\mathrm{d}\, f \\ & \int_0^B p(f)df = P_{tot}, \quad p(f) \geq 0 \end{aligned} \tag{1.5}$$

where $\gamma_k(\cdot)$ is the ratio between $|h_k(f)|^2$, the power spectrum of the channel of user k, and the power spectrum of the noise $N_k(f)$, i.e. $\gamma_k(f) = \frac{|h_k(f)|^2}{N_k(f)}$ and $\beta = \frac{1.5}{-\ln(-5\mathrm{BER})}$.

This approach enables to obtain only upper bounds on the performance of practical OFDMA systems since the minimum granularity of subcarriers is finite in actual systems.

From this unified framework several approaches steam:

– Subcarrier assignment. It is obtained by assuming a uniform power allocation over the entire available frequency band, i.e. $p(f) = 1$

$$\text{maximize} \qquad \sum_{k=1}^{K} U_k(\int_{D_k} \log_2[1 + \beta\gamma_k(f)]\mathrm{d}\, f)$$

– Power allocation. For a given subcarrier assignment $\{S_k\}$ the power spectrum $p(f)$ is optimized in order to

$$\text{maximize} \qquad \sum_{k=1}^{K} U_k(\int_{D_k} \log_2[1 + \beta p(f)\rho_k(f)]df)$$

$$\text{subject to} \qquad \int_0^B p(f)df = P_{tot}, \quad p(f) \geq 0$$

– Joint subcarrier assignment and power allocation. The simultaneous optimization of subcarrier and power allocation is performed for the problem defined by objective function and constraints in (1.5).

1.3. Fundamental algorithms for maximizing the achievable rates in a multiuser OFDM cell.

The algorithms presented in this section solve the problem of power allocation, considering two different quantities to maximize. Multiuser extension are designed for multiple access channels (MAC), i.e., the uplink transmission.

In OFDM systems, waterfilling is the only algorithm that solves the general problem of maximizing (Shannon) capacity subject to finite power constraint.

1.3.1. *Waterfilling for capacity-achieving Gaussian inputs.*

The optimization problem which is considered in this part is to find the optimal power allocation which maximizes the sum capacity in the general context of parallel

Gaussian channels. The solution is given by the well-known waterfilling algorithm. Waterfilling principle takes advantage of the problem structure by decomposing the channel into orthogonal modes, which greatly reduces the optimization complexity. This idea can also be extended to the multiuser case under the "aggregate sum capacity"[1] objective.

1.3.1.1. *Model*

In a generic K-user Gaussian vector multiple access channel (MAC), the output signal **y** received at the base station (BS) can be expressed as follows

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{H}_i \mathbf{x}_i + \mathbf{z}$$

where

– $\mathbf{H}_i$ is the time-invariant channel matrix (uplink channel between user i and BS),

– $\mathbf{x}_i$ is the input signal transmitted by user i,

– **z** is the additive Gaussian noise vector with a covariance matrix denoted as $\mathbf{S}_z$.

Channels are assumed to be known to both the transmitters and the receiver. Furthermore, there is no cooperation between the transmitters. Transmitted signals $\{\mathbf{x}_i\}$ are assumed to be independent and satisfy the power constraints

$$\text{tr}\left(\mathbf{E}[\mathbf{x}_i \mathbf{x}_i^T]\right) \leq P_i \ .$$

Let $\mathbf{S}_i = \mathbf{E}[\mathbf{x}_i \mathbf{x}_i^T]$, the power constraint becomes $\text{tr}\left(\mathbf{S}_i\right) \leq P_i$.

1.3.1.2. *Single-user waterfilling*

For a single-user Gaussian vector channel, the signal expression is reduced to

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \ .$$

Therefore the sum capacity maximization problem is

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}\log|\mathbf{H}\mathbf{S}\mathbf{H}^T + \mathbf{S}_z| - \frac{1}{2}\log|\mathbf{S}_z| \\ \text{subject to} & \text{tr}(\mathbf{S}) \leq P \\ & \mathbf{S} \geq 0 \ . \end{array}$$

1. The aggregate sum capacity is defined as the sum of all capacities over parallel channels and over users.

The problem resolution is as follows.

1) The first step is to take the eigenvalue decomposition of the noise covariance matrix $\mathbf{S}_z$, which is symmetric positive definite

$$\mathbf{S}_z = Q\Delta Q^T$$

where Q is an orthogonal matrix and Δ is a diagonal matrix containing the eigenvalues.

The problem can then be rewritten

$$\text{maximize } \frac{1}{2}\log|\hat{\mathbf{H}}\mathbf{S}\hat{\mathbf{H}}^T + \mathbf{I}|$$

where $\hat{\mathbf{H}}$ is the normalized channel matrix: $\hat{\mathbf{H}} = \Delta^{-1/2}Q^T\mathbf{H}$.

2) The second step is to take the singular value decomposition of $\hat{\mathbf{H}}$

$$\hat{\mathbf{H}} = F\Sigma M^T$$

where F and M are orthogonal matrices and Σ is a diagonal matrix containing singular values $\{h_1, \ldots, h_r\}$, r being the rank of $\hat{\mathbf{H}}$.

Let us define $\hat{\mathbf{S}} = M^T\mathbf{S}M$, we have $tr(\mathbf{S}) = tr(\hat{\mathbf{S}})$. The problem is then equivalent to

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\log|\Sigma\hat{\mathbf{S}}\Sigma^T + \mathbf{I}| \\ \text{subject to} \quad & \text{tr}(\hat{\mathbf{S}}) \leq P \\ & \hat{\mathbf{S}} \geq 0\ . \end{aligned}$$

or in scalar form, with $\hat{\mathbf{S}} = diag\{p_1, \ldots, p_r\}$

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\log\left[\prod_{i=1}^{r}(h_i^2 p_i + 1)\right] \\ \text{subject to} \quad & \sum_{i=1}^{r} p_i \leq P \\ & p_i \geq 0\ . \end{aligned}$$

After introducing Lagrange multiplier λ and solving more general Karush-Kuhn-Tucker (KKT) conditions, we get the expressions of the optimal power to transmit on subcarrier i ($i = 1 \ldots N$)

$$p_i = \max(\lambda - \frac{1}{h_i^2}, 0) \tag{1.6}$$

where λ is the waterfilling level settled so as to satisfy the power constraint.

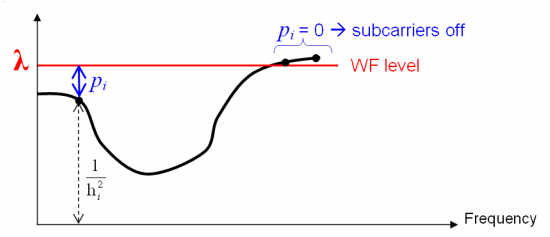

Figure 1.3. *Waterfilling principle: allocated power corresponds to the height of water that has been poured (between black curve and red line, but only when red line is above the black curve).*

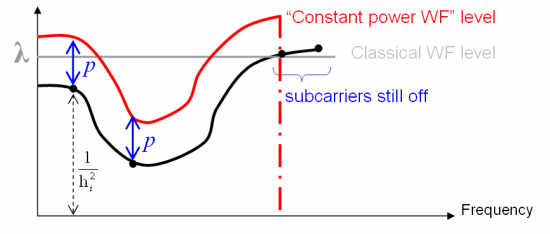

Figure 1.4. *Constant Power Waterfilling principle: switched-off subcarriers are the same as in classical waterfilling scheme, but power is uniformly allocated on the remaining subcarriers.*

1.3.1.3. *Waterfilling variants*

Several waterfilling variants exists.

In the context of fading channels ([GOL 97]), if subchannels statistics are known, Shannon capacity is shown to be achieved by waterfilling over time, which is also known as *statistical waterfilling*.

Constant power waterfilling ([YU 06]) was thought in order to simplify transmitter design. If we dwell upon Shannon capacity formula which is roughly $\log(1 + \text{SNR})$, we can observe the capacity is more sensitive to SNR when SNR is low. Power has thus to be particularly well allocated to low SNR subchannels. Constant power waterfilling simply exploits this observation and consists in allocating zero power to subchannels that would receiver zero power in exact waterfilling and constant power in subchannels that would receive positive power in exact waterfilling.

Another practical aspect is that waterfilling should be considered jointly with bitloading. Once power allocation is performed, bitloading traditionally follows and consists in selecting a modulation and coding scheme adapted to the resulting SNR so

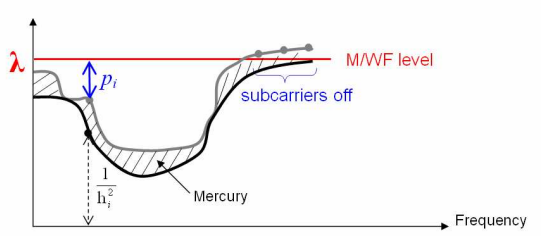

Figure 1.5. *Mercury/waterfilling principle: an intermediate step where mercury is poured comes before waterfilling. The allocated power still corresponds to the height of water.*

that system constraints (e.g. target BER) are satisfied. An SNR margin, corresponding to the gap of any practical systems w.r.t. Shannon theoretical capacity, is introduced. This gap generally includes an error margin, which is a safety factor included to protect the modem's performance in case of unanticipated channel degradation. Usual gap values can be found for the design of DSL modems. For instance, an uncoded modem requires an SNR gap of 9.8 dB to operate at a symbol error probability of 10^{-7}, gap that can be reduced by the addition of a coding gain. A remark can be made is that since different constellations can be loaded, different gap values should be used over subcarriers. In spite of this, a constant gap[2] is considered.

Mercury/waterfilling ([LOZ 06]) presented in detail in subsection 1.3.2 partially solves this paradox. This variant's naming results from an analogy with waterfilling: a layer of mercury[3] is first poured (i.e. it lays under water) and the mercury height on each subchannel is actually fitted to the loaded constellation.

1.3.1.4. *Multi-user waterfilling*

The idea of waterfilling can be generalized to multiple access channels. In such channels, the sum capacity is maximized when using successive interference cancellation (SIC) ([CHE 93]). Let us explain in a few words the principle of SIC in a basic two-user case. We assume user 1 has a higher priority than user 2, therefore the receiver decodes the signal sent by user 2 first, considering the signal transmitted by user 1 as noise. Then the receiver regenerates the signal from user 2, substracts it from the received signal, and finally decodes the signal sent by user 1.

---

2. The gap computation is based on an estimate of the symbol error probability when using QAM on Gaussian channels.

3. Due to higher density w.r.t. water, poured mercury level is not systematically horizontal.

If we come back to the general model of a K-user multiple access channel (see §1.3.1.1), the sum capacity under SIC assumption has the following expression

$$\begin{aligned} & \frac{1}{2}\log\frac{|\sum_{i=1}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|}{|\sum_{i=2}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|}+\frac{1}{2}\log\frac{|\sum_{i=2}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|}{|\sum_{i=3}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|}+\ldots \\ & \qquad\qquad +\frac{1}{2}\log\frac{|\mathbf{H}_K\mathbf{S}_K\mathbf{H}_K^T+\mathbf{S}_z|}{|\mathbf{S}_z|} \\ = \quad & \frac{1}{2}\log\frac{|\sum_{i=1}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|}{|\mathbf{S}_z|} \\ = \quad & \frac{1}{2}\log|\sum_{i=1}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|-\frac{1}{2}\log|\mathbf{S}_z| \end{aligned}$$

Then the sum capacity maximization problem becomes

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\log|\sum_{i=1}^{K}\mathbf{H}_i\mathbf{S}_i\mathbf{H}_i^T+\mathbf{S}_z|-\frac{1}{2}\log|\mathbf{S}_z| \\ \text{subject to} \quad & \text{tr}(\mathbf{S}_i)\leq P_i \\ & \mathbf{S}_i\geq 0 \qquad \forall i=1,\ldots,K. \end{aligned}$$

The problem solution is as follows.

$\{\mathbf{S}_i\}$ is an optimal solution to the rate-sum maximization problem if and only if $\mathbf{S}_i$ is the single-user waterfilling covariance matrix of the channel $\mathbf{H}_i$ with $\mathbf{S}_z+\sum_{j=1,j\neq i}^{K}\mathbf{H}_j\mathbf{S}_j\mathbf{H}_j^T$ as noise, for all $i=1,2,...,K$ ([YU 04]).

The iterative waterfilling algorithm is as follows.

The idea of iterative waterfilling is still presented by Yu in [YU 04], as an efficient numerical algorithm to compute the optimal input distribution that maximizes sum capacity on a Gaussian multiple access channel with vector inputs and a vector output. The numerical algorithm can be implemented in an iterative way, to compute the set of rate-sum optimal input covariance matrices.

Algorithm: Iterative waterfilling

repeat
  for $i = 1\ to\ K$
    $\mathbf{S}'_z = \sum_{j=1, j\neq i}^{K} \mathbf{H}_j \mathbf{S}_j \mathbf{H}_j^T + \mathbf{S}_z$;
    $\mathbf{S}_i = arg\max_S \frac{1}{2} \log |\mathbf{H}_i \mathbf{S} \mathbf{H}_i^T + \mathbf{S}'_z|$;
  end
until sum rate convergence.

The two main results, proved in [YU 04], are the following:

– Using iterative waterfilling algorithm, the sum rate converges to the sum capacity, and $(\mathbf{S}_1, ..., \mathbf{S}_K)$ converges to an optimal set of input covariance matrices for the Gaussian vector multiple access channel.

– After one iteration of iterative waterfilling algorithm, the $\{\mathbf{S}_i\}_{i=1,...,K}$ achieve a total data rate $\sum_{i=1}^{K} r_i$ that is at most $(K-1)\, m/2$ nats away from the sum capacity, where m is the number of output dimensions.

The algorithm can be used to find the set of optimal covariance matrices that achieve the sum capacity of a Gaussian vector multiple access channel. This set of K covariance matrices gives a set of $K!$ corner points of a capacity pentagon, each corresponding to a different decoding order. Upper and lower bounds on the entire capacity region can be derived from these corner points.

1.3.2. *Mercury/waterfilling for maximizing achievable rates with arbitrary input constellations.*

1.3.2.1. *Why Mercury/waterfilling rather than Waterfilling?*

The well-known waterfilling algorithm solves the problem of maximizing capacity, which is defined as the maximal mutual information assuming all possible input distributions. In case of Gaussian channels, Shannon capacity is reached when the input signal has a Gaussian distribution; nevertheless, inputs are usually drawn from discrete constellations and thus Gaussian inputs cannot be realized in practice. For this reason, a way to carry out waterfilling is once the power allocation decided, to compensate on each subcarrier all aspects due to practical implementation (among which the use of discrete constellations) by using an SNR gap. Then the number of bits transmitted by user k on subcarrier n is given by the generic formula

$$R_{k,n} = \log_2 \left(1 + \frac{\mathrm{SNR}_{k,n}}{\Gamma} \right) \tag{1.7}$$

where $\mathrm{SNR}_{k,n}$ is the signal-to-noise ratio of user k on subcarrier n received at the BS side and Γ is the so-called SNR gap. Finally the modulation and coding scheme

Constellation	Coding rate	Spectral efficiency (bits/symbol)
QPSK	$1/2$	1
QPSK	$3/4$	1.5
16QAM	$1/2$	2
16QAM	$3/4$	3
64QAM	$1/2$	3
64QAM	$3/4$	4.5

Table 1.1. *Typical MCS table used in WiMAX systems*

(MCS) is selected among a given list of associated constellation and coding rates as provided in table 1.1 for instance.

A clear paradox of waterfilling is that Gaussian inputs are assumed since the optimized quantity is capacity, whereas Gaussian inputs are not used in practice. The problem should therefore be formulated differently, under the assumption of finite constellations.

1.3.2.2. *Single user Mercury/waterfilling*

Basic Mercury/waterfilling is a power allocation algorithm derived in [LOZ 06] for the single user context [4]. It aims at maximizing the sum mutual information over parallel Gaussian channels under power constraint assuming precisely arbitrary input distributions. These inputs can indeed be drawn from discrete constellations such as QPSK or 16QAM. We assume unit-variance Gaussian noise in the sequel.

Let γ_i be a measure of the channel strength on subcarrier i defined as $\gamma_i = |h_i|^2$, where h_i the normalized channel gain, as defined in 1.3.1. Let $I_i(\rho_i)$ be the input-output mutual information on the i^{th} channel, where the SNR denoted ρ_i is equal to $p_i\gamma_i$. The power allocation problem can be expressed as following

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n} I_i(p_i\gamma_i) \\ \text{subject to} \quad & \sum_{i=1}^{N} p_i \leq P \\ & p_i \geq 0 \ . \end{aligned}$$

This optimization problem has been solved thanks to a recent result of information theory ([GUO 05]) on the expression of the mutual information derivative, linked to

4. Please note that the multiuser case has not been solved

Constellations	MMSE function
BPSK	$\text{MMSE}^{\text{BPSK}}(\rho) = 1 - \int_{-\infty}^{\infty} \tanh(2\sqrt{(}\rho\xi) \frac{e^{-(\xi-\sqrt{(}\rho))^2}}{\sqrt{(}\pi)} d\xi$
QPSK	$\text{MMSE}^{\text{QPSK}}(\rho) = \text{MMSE}^{\text{BPSK}}(\frac{\rho}{2})$
4PAM	$\text{MMSE}^{\text{4PAM}}(\rho) = 1 - \int_{-\infty}^{\infty} \frac{\left(e^{-8\rho/5}\sinh(6\sqrt{(}\rho/5)\xi) + \sinh(2\sqrt{(}\rho/5)\xi)\right)^2}{e^{-8\rho/5}\cosh(6\sqrt{(}\rho/5)\xi) + \cosh(2\sqrt{(}\rho/5)\xi)^2} \frac{e^{-\xi^2-\rho/5}}{10\sqrt{(}\pi)} d\xi$
16QAM	$\text{MMSE}^{\text{16QAM}}(\rho) = \text{MMSE}^{\text{4PAM}}(\frac{\rho}{2})$

Table 1.2. *MMSE expressions for different constellations*

the minimum mean square error (MMSE) expression:

$$\frac{d}{d\rho} I(\rho) = \text{MMSE}(\rho)$$

The problem solution is as follows.

Basically Mercury/waterfilling algorithm consists in allocating power p_i^* to subcarrier i. The set of powers $\{p_i^*\}_{i=1,\ldots,N}$ is given by

$$p_i^* = 0 \quad \text{if } \gamma_i \leq \eta \tag{1.8}$$

$$\gamma_i \text{MMSE}_i(p_i^* \gamma_i) = 0 \quad \text{if } \gamma_i > \eta \tag{1.9}$$

where the threshold η has to be set so as to satisfy the power constraint

$$\sum_{i=1,\gamma_i>\eta}^{N} \frac{1}{\gamma_i} \text{MMSE}_i^{-1}\left(\frac{\eta}{\gamma_i}\right) = P \, .$$

The function MMSE_i is the MMSE function corresponding to the input constellation loaded on the i^{th} subcarrier.

A barrier to Mercury/waterfilling's practical implementation is its great computational load due to the non-linear nature of MMSE function, as shown on table 1.2. These values should be tabulated for various constellations, to be used in the algorithm implementation.

A graphical interpretation is the following. In order to visualize the two successive steps of mercury and water pouring, we define the function $G_i(.)$ as following

$$\begin{aligned} G_i(\zeta) = \quad & 1/\zeta - \text{MMSE}_i^{-1}(\zeta) \quad \text{if } 0 \leq \zeta \leq 1 \\ & 0 \qquad\qquad\qquad\quad \text{if } \zeta > 1 \, . \end{aligned}$$

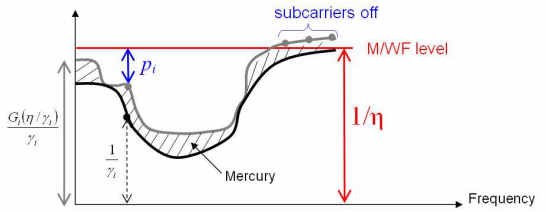

Figure 1.6. *Differents steps in Mercury/waterfilling algorithm.*

For Gaussian inputs, $G_i(\zeta) = 1 \; \forall \zeta$.

The algorithm can be decomposed into the following steps:

1) Plot $1/\gamma_i$ over the subcarriers.

2) Fix a value for η.

a) Pour mercury until its height reaches $G_i(\eta/\gamma_i)/\gamma_i$ on each subcarrier.

b) Waterfill until the water level reaches $1/\eta$. The water height over the mercury gives p_i^*.

3) Check if the power constraint is satisfied by summing all $\{p_i^*\}$. If not, tune η and go back to step 2.

1.4. Resource allocation algorithms in a single-cell OFDMA network.

Algorithms that optimize resource allocation among users in Wimax are not specified in the standard.

At the PHY layer, the typical procedures for allocating resources consist of the following steps. Let us focus on the downlink channel. Users first estimate and feedback the channel state information (CSI) to the base station (BS). Then, the BS performs the allocation algorithm based on the CSI knowledge and assign subcarriers to each user and the corresponding powers. Finally, it starts transmitting according to the defined resource allocation. In the uplink, the procedure is similar. The BS can directly estimates the channel of each user. Thus, no feedback from the user terminal is required. By using the CSI the BS performs resource allocation and informs each user about its subcarriers allocation and the corresponding transmitting power.

Examples of possible allocation methods are : Minimum sum power, Maximum sum rate, Fair allocation, Proportional fair, Maximum fairness (max-min).

1.4.1. *Minimum Sum Power*

Power Allocation for a Single User

Let us consider first the power minimization problem defined in (1.2) with a single active user in the system. In fact, this simpler problem gives a better understanding and provides a bit allocation technique which can be used for the multiuser case. Let us denote by $f_k(r)$ the received power required to reliably transmit at rate r. The required received power is determined by taking into account the actual coding and modulation schemes and the bit error rate constraints. Due to practical constraints, the number of bits per channel use in each subcarriers must be an integer. Then, the objective is to minimize $\sum_{n=1}^{N} \frac{1}{|h_n|^2} f(r_n)$ under the constraint $R = \sum_{n=1}^{N} r_k$. In this case an optimal approach is based on a greedy algorithm which assigns one bit at a time by choosing the tone requiring the minimum energy. Several algorithms have been proposed for solving this problem with a common structure.

Initialization $\quad \forall n, r_n = 0, \Delta P_n = \dfrac{f(1) - f(0)}{|h_n|^2}$

Iterations (R times)

$\hat{n} = \arg\min_n \Delta P_n$

$r_{\hat{n}} = r_{\hat{n}} + 1$

$\Delta P_{\hat{n}} = \dfrac{f(r_{\hat{n}} + 1) - f(r_{\hat{n}})}{|h_n|^2}$

Table 1.3. *Power allocation algorithm for a single user.*

Extension to the Multiuser Case

In the multiuser case, users are not allowed to share a subcarrier. This creates a dependency among users and rends the greedy algorithm described in the previous item suboptimal.

Lagrangian Relaxation Algorithm

In the multiuser case, it is required an optimization over discrete variables which implies an exhaustive search. Historically, this problem has been tackled by relaxation methods: in order to simplify the optimization some of the constraints are relaxed. We propose here the relaxation method in [WON 99] where the requirement of integer bit loads is relaxed and $\rho_{k,n}$, a sharing factor for the subcarriers is introduced. The problem can be formalized as follows

$$\min_{c_{k,n}\in[0,M],\rho_{k,n}\in[0,1]} \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{\rho_{k,n}}{|h_{k,n}|^2} f_k(r_{k,n})$$

$$\text{subject to:} \quad \sum_{n=1}^{N} \rho_{k,n} c_{k,n} = R_k \qquad \forall k$$

$$\sum_{k=1}^{K} \rho_{k,n} = 1 \qquad \forall n \tag{1.10}$$

where c_{nk} is the number of bits of the k-th user that are assigned to the n-th subcarrier. Note that a feasible point satisfying the constraints of the original problem with integer bit load and disjoint allocation of subcarriers is also a feasible point satisfying the constraints in (1.10). Since in the formulation of the problem (1.10) with relaxation the optimization is done over a wider set of feasible points, the solution to the minimization problem with relaxation is only a bound for the solution to the original problem. By writing $r_{k,n} = c_{k,n}\rho_{k,n}$, the problem is transformed into a convex optimization problem. The details of this solution can be found in [WON 99].

In general, the obtained solutions $r^*_{k,n}$ and $\rho^*_{k,n}$ could be directly suitable for the original problem without relaxation. In fact, $\{r^*_{k,n},\, n = 1, \ldots N,\, k = 1, \ldots K\}$ could be not integers and $\{\rho^*_{k,n},\, n = 1, \ldots N,\, k = 1, \ldots K\}$ could indicate a time sharing solution. Additionally, a quantization of these values could not satisfy the individual rate constraints any longer. These problems are typically circumvented by some heuristic approach. In practice, the resource allocation problem based on Lagrangian relaxation is performed in several steps. As first step the subcarrier allocation problem is solved by applying the multiuser Lagrangian relaxation algorithm. As second step, the subcarriers that should be optimally shared by several user, i.e $\rho \in (0, 1)$ for some n, are assigned to the users with the biggest ρ_{kn}. Finally, bit and power allocation is performed by applying the algorithm for the minimum sum power allocation in Table 1.3 to each single user k and considering only the subcarriers assigned to user k.

The proposed scheme allows a reduction of the total transmitted power of 5-10dB compared to OFDM without adaptive modulation, and 3-5dB with respect to OFDM with adaptive modulation and bit allocation, but no adaptive subcarrier allocation.

Algorithms based on Lagrange Dual Decomposition

The relaxation of the constraints adopted in [WON 99] modifies the original OFDMA system, so it may introduce significant loss in optimality. Moreover, the algorithm

is computationally intensive because of its slow convergence rate and it is difficult to be implemented. In [KIV 03] a efficient algorithm was proposed by approximating the channels with flat fading over the whole available bandwidth and assuming $\widetilde{\gamma}_{k,n} = \frac{1}{N}\sum_n \gamma_{k,n}$ instead of $\gamma_{k,n}$. The minimization of the power consumption with constraints on the bit rate and transmission rate is solved in two steps. In the first step the *number of subcarriers* to be allocated to each user is determined using the signal-to-noise ratio γ_k. In the second step the best assignment of subcarriers to users is performed. The performance of this approach depends on the labelling of subcarriers. Then, the resulting resource allocation strategy is neither unique nor stable. Furthermore, the algorithm is not extended to frequency selective fading channels.

In [SEO 06] an approach based on Lagrange dual decomposition is proposed for the more general case of weighted sum power minimization. This approach is extremely efficient and results in the advanced solution to the sum power minimization problem at the time this book was written. The sum power minimization problem is there generalized as follows. A weight λ_k is assigned to each user and the objective function to be minimized is

$$\sum_{k=1}^{K} \lambda_k \sum_{n \in S_k} p_{kn} \tag{1.11}$$

under the same constraints of problem (1.2).

The problem can be solved by considering the Lagrangian

$$\mathcal{L}(\{p_{k,n}\},\{R_{k,n}\},\boldsymbol{\mu}) = \sum_{k=1}^{K} \lambda_k \sum_{n=1}^{N} p_{k,n} - \sum_{k=1}^{K} \mu_k \left(\sum_{n=1}^{N} R_{k,n} - R_k \right)$$

where $\boldsymbol{\mu} = (\mu_1, \ldots \mu_K)$ are the Lagrangian multipliers and $R_{k,n} = log_2(1+p_{k,n}\gamma_{k,n})$. Then, the Lagrangian dual function is given by

$$\begin{aligned} g(\boldsymbol{\mu}) &= \min_{\{p_{k,n}\},\{R_{k,n}\}} \mathcal{L}(\{p_{k,n}\},\{R_{k,n}\},\boldsymbol{\mu}) \\ &= \sum_{n=1}^{N} g_n^{'}(\boldsymbol{\mu}) + \sum_{k=1}^{K} \mu_k R_k \end{aligned}$$

where

$$g_n^{'}(\boldsymbol{\mu}) = \min_{\{p_{k,n}\}} \left(\sum_{k=1}^{K} \lambda_k p_{k,n} - \sum_{k=1}^{K} \mu_k R_{k,n}\right) \qquad \forall n = 1, \ldots N. \tag{1.12}$$

Note that the minimization of $g(\boldsymbol{\mu})$ reduces to the N disjoint optimizations in (1.12). Additionally, the objects of the minimizations in (1.12) are convex functions of $p_{k,n}$.

The dual Lagrange optimization problem

$$\begin{array}{ll} \text{maximize} & g(\boldsymbol{\mu}) \\ \text{subject to} & \mu_k \geq 0 \qquad \forall k = 1, \ldots K \end{array} \tag{1.13}$$

can be solved by convex optimization approaches since, as well-known, $g(\boldsymbol{\mu})$ is concave. In general, the optimum solution to the dual Lagrangian problem does not provide the optimal solution to the original minimization problem in (1.11). Adopting the solution of problem (1.13) as solution of the problem defined in (1.11) determined the so called duality gap. However, in [SEO 06] is observed that for the specific problem under consideration the duality gap becomes smaller and smaller as the number of subcarriers increases. Then, for practical systems the number of subcarriers is typically large enough that we can consider the duality gap negligible.

Thanks to the convexity of the argument of the $\min$ operator in 1.12, for a fixed value of $\boldsymbol{\mu}$, $g'_n(\boldsymbol{\mu})$ is given by

$$g'_n(\boldsymbol{\mu}) = \min_k \left\{ \lambda_k \left(M_k - \frac{1}{\gamma_{k,n}} \right)^+ - \frac{\mu_k}{2} \log_2 \left(1 + \left(M_k - \frac{1}{\gamma_{k,n}} \right)^+ \gamma_{k,n} \right) \right\} \tag{1.14}$$

where $M_k = \frac{\mu_k}{2 \log 2 \lambda_k}$ and $(x)^+ = \max(0, x)$. After solving (1.14) for all n we can determine $g(\boldsymbol{\mu})$ for the fixed value of $\boldsymbol{\mu}$. The optimal $\boldsymbol{\mu}^*$ maximizing $g(\boldsymbol{\mu})$ can be efficiently obtained by using the ellipsoid method, until every user's rate converges. A subgradient value that can be used in the application of the ellipsoid method is

$$d_k = R_k - \sum_{n=1}^{N} R^*_{k,n} \qquad k = 1, \ldots K$$

where $\{R^*_{k,n}\}$ are the solutions of the system 1.12. Table 1.4 summarizes this algorithm.

The overall optimization requires $\mathcal{O}(K^2)$ runs of the optimization problem with complexity $\mathcal{O}(NK)$. Hence, the total complexity is $\mathcal{O}(NK^3)$, instead of $\mathcal{O}(NK^N)$ in case of exhaustive search.

1.4.2. *Sum rate maximization*

The problem of sum rate maximization in flat fading downlink channels has been investigated by Li and Goldsmith [LI 01] in an information theoretic setting and extended to parallel flat fading downlink channels by Tse [TSE 00].

Initialization

Choice of $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and a $K \times K$ matrix $\mathbf{P} = \mathbf{P}_0$ such that

$\mathcal{E}(\boldsymbol{\mu}_0, \mathbf{P}_0) = \{\mathbf{z} | (\mathbf{z} - \boldsymbol{\mu}_0)^T \mathbf{P}_0^{-1} (\mathbf{z} - \boldsymbol{\mu}_0) \leq 1\}$ contains the optimum $\boldsymbol{\mu}$

Fix $\varepsilon > 0$ (accuracy) and $m = 1$

repeat

$m = m + 1$

for $n = 1, ...N$

for $k = 1, ...K$

$$aux_k = \lambda_k \left(M_k - \frac{1}{\gamma_{k,n}} \right)^+ - \frac{\mu_k}{2} \log_2 \left(1 + \left(M_k - \frac{1}{\gamma_{k,n}} \right)^+ \gamma_{k,n} \right)$$

endfor

select $k^* = \operatorname{argmin}_k (aux_k)$

$$p_{k^*,m} := \left(M_{k^*} - \frac{1}{\gamma_{k*,n}} \right)^+ \text{ and } p_{k,n} := 0 \; \forall k \neq k^*$$

$$R_{k^*,m} := \frac{1}{2} \log_2 \left(1 + \frac{p_{k^*,m}}{\gamma_{k*,n}} \right) \text{ and } R_{k,n} := 0 \; \forall k \neq k^*$$

endfor

evaluate subgradient $\mathbf{d} = (d_1, \ldots d_k),$ where $d_k := R_k - \sum_n R_{k,n}$

if $\sqrt{\mathbf{d}^T \mathbf{P} \mathbf{d}} < \varepsilon$

then return $\{p_{k,n}\}$

else update ellipsoid

$$\widetilde{\mathbf{d}} := \frac{\mathbf{d}}{\sqrt{\mathbf{d}^T \mathbf{P} \mathbf{d}}}; \quad \boldsymbol{\mu} := \boldsymbol{\mu} - \frac{1}{m+1} \mathbf{P} \widetilde{\mathbf{d}}; \quad \mathbf{P} =: \frac{m^2}{m^2 - 1} \left(\mathbf{P} - \frac{2}{m+1} \mathbf{P} \widetilde{\mathbf{d}} \widetilde{\mathbf{d}}^T \mathbf{P} \right)$$

until $\sqrt{\mathbf{d}^T \mathbf{P} \mathbf{d}} < \varepsilon$

Table 1.4. *Sum power minimization in downlink by Lagrangian duality.*

Several approaches have been proposed that decouple the sum rate maximization problem into two disjoint problems: subcarrier allocation and power allocation for

a single user. As an example we propose the work in [JAN 03]. Each subcarrier is assigned only to the user with the best channel gain. Once the subcarrier allocation has been performed, the power is allocated to each of the subcarriers. This last problem reduces to power allocation for a single user and it is solvable by using standard Lagrange multiplier techniques. The solution is given by the classical waterfilling method, i.e.

$$\begin{cases} p_{k_n^*} & = \sigma^2 \left[\frac{1}{\lambda_0} - \frac{1}{|h_{k_n^*}|^2}\right]^+ \quad n = 1, \ldots, N \\ p_{k,n} & = 0 \qquad k \neq k_n^* \end{cases}$$

being λ_0 a threshold to be determined from the total power constraint. There is no explicit method to calculate λ_0, the waterlevel and a numerical search method may result computationally too intensive. In such a case an equal power allocation in all the subcarriers may be used. This approximation is based on the observation that waterfilling and equal power may yield marginal performance differences.

The most efficient sum rate maximization algorithm which performs subcarrier assignment and power allocation jointly in a optimum way is in [SEO 06] and it is based on the Lagrange duality. The rationale behind this algorithm is the same as for the algorithm on the sum power minimization presented in Table 1.4. For a detailed description and discussion of the algorithm the interested reader is referred to [SEO 06]. We summarize it in Table 1.5.

1.4.3. *Fair allocation*

In this section we consider the utility model (1.4). For k user with achievable transmission rate R_k, the corresponding utility is given by $U_k(R_k)$, where $U_k(\cdot)$ is a nondecreasing and typically concave function. Extensions of the fair allocation to nonconcave utility functions are proposed in [SON 05a].

Subcarrier allocation

By assuming a fixed power allocation $\{\mathbf{p}[1], \mathbf{p}[2], \ldots, \mathbf{p}[K]\}$ the joint problem of resource allocation reduces to subcarrier allocation. The latter can be expressed by a nonlinear integer programming problem. The disjoint subsets S_k^* of tones assigned to users are a solution to the problem (1.4) if they satisfy the following constraints:

$$\begin{aligned} \left.\frac{\mathrm{d}U_k(r)}{\mathrm{d}r}\right|_{r=R_k^*} c_k^{\boldsymbol{p}}[n] \quad &\geq \quad \left.\frac{\mathrm{d}U_\ell(r)}{\mathrm{d}r}\right|_{r=R_\ell^*} c_\ell^{\boldsymbol{p}}[n], \qquad \forall n \in S_k^* \qquad \forall k, \ell \\ R_k^* \quad &= \quad \sum_{j \in S_k^*} c_j^{\boldsymbol{p}}[n] \Delta f \end{aligned} \tag{1.15}$$

Initialization

Fix $\mu = \mu_0$ and $a \in \mathcal{R}^+$ such that

$[\mu - a, \mu + a]$ contains the optimumμ

Fix $\varepsilon > 0$ (accuracy)

repeat

for $n = 1, ...N$

for $k = 1, ...K$

$$aux_k = \frac{1}{2}\log_2\left(1 + \left(\frac{1}{2\log 2\mu} - \frac{1}{\gamma_{k,n}}\right)^+ \gamma_{k,n}\right)$$

endfor

select $k^* = \mathrm{argmin}_k(aux_k)$

$$p_{k^*,m} := \left(\frac{1}{2\log 2\mu} - \frac{1}{\gamma_{k*,n}}\right) \text{ and } p_{k,n} := 0 \; \forall k \neq k^*$$

$$R_{k^*,m} := \log_2\left(1 + \frac{p_{k^*,m}}{\gamma_{k*,n}}\right) \text{ and } R_{k,n} := 0 \; \forall k \neq k^*$$

endfor

evaluate $d = P_{tot} - \sum_n p_{k,n}$

if $|d| < \varepsilon$

then return $\{p_{k,n}\}$

else$\mu := \mu - \frac{1}{2}a \cdot \mathrm{sign}(d); \quad a := \frac{a}{2}$

until $|d| < \varepsilon$

Table 1.5. *Sum rate maximization in downlink by Lagrangian duality.*

where Δf is the subcarrier bandwidth and $c_j^{\boldsymbol{p}}[n]$ is the achievable transmission efficiency (data rate per Hertz) of user j on subcarrier n corresponding to the given power allocation, i.e. $c_j^{\boldsymbol{p}}[n] = \log_2(1+\beta p_{jn}\gamma_{jn})$. These optimality conditions are sufficient, but non necessary because the discrete optimization problem is not convex.

In particular, for continuous rate adaptation, and taking into account that $Pr\{c_k^{\boldsymbol{p}}[n] = c_\ell^{\boldsymbol{p}}[m]\} = 0$ for $(k,n) \neq (\ell,m)$, subcarrier n should be assigned to user $k^* = k^*(n)$ such that

$$k^*(n) = \underset{i\in\{1,2,\ldots K\}}{\operatorname{argmax}} \left\{ \left.\frac{\mathrm{d}U_i(r)}{\mathrm{d}r}\right|_{r=R_i^*} c_i^{\boldsymbol{p}}[n] \right\}$$

In an analogous way, if a linear utility function is considered with constant marginal utility $\frac{\mathrm{d}U_k(r)}{\mathrm{d}r}$, subcarrier n should be assigned to user k^* such that

$$k^* = \underset{i\in\{1,2,\ldots K\}}{\operatorname{argmax}} \{c_i^{\boldsymbol{p}}[n]\}$$

The utility based subcarrier assignment is a nonlinear combinatorial optimization problem. For this class of problems there exists no general approach to achieve optimality. In [SON 05b] a sorting-search algorithm for an OFDMA system with two active users is proposed (see Table 1.6). Then, the sorting-search algorithm is generalized to the case of K users by updating the subcarrier assignment of each pair of users iteratively by means of the subcarrier assignment algorithm for the two-user case. The computational complexity is nearly $(K-1)^2(N+1)\log_2 N$ which is still a low complexity if compared to the complexity of the exhaustive search N^K.

Power allocation

Under the assumption that the sets S_k of the subcarrier assignment are given the optimal continuous rate adaptation is presented in [SON 05a]. The optimal solution is a utility based waterfilling approach:

$$p_{kn}^* = \left[\frac{1}{\lambda}\frac{\mathrm{d}U_k(r)}{\mathrm{d}r} - \frac{1}{\beta\gamma_{kn}}\right]^+$$

with waterfilling level λ calculated in order to satisfy the power constraint.

Joint subcarrier assignment and power allocation

A joint optimum resource allocation is presented in [SON 05a]. The optimal resource allocation must satisfy the optimality conditions for both the subcarrier assignment only and power allocation only problems. The algorithm is proposed for continuous rate adaptation and it is shown in Table 1.7. The algorithm iteratively performs subcarrier assignment, power allocation, and update of the marginal utility. If the utility function is concave and the parameter μ in the update step is properly selected

sort $((c_2^{\boldsymbol{p}}[n]/c_1^{\boldsymbol{p}}[n]))$ in increasing order

get thresholds: $T[n], n \in \{1 \dots N+1\}$ in increasing order

$low = 1, high = N+1$

while $high - low > 0$

  $center \leftarrow \lfloor (low + high)/2 \rfloor$

  $T \leftarrow center$

  if $T - ((\dfrac{\mathrm{d}U_1(\mathrm{d}r_1)}{\mathrm{d}r} / \dfrac{\mathrm{d}U_2(r_2)}{\mathrm{d}r})) > 0$

    $high \leftarrow center$

  else

    $low \leftarrow center$

  end if

end while

choose the best T between low and $high$

(1.16)

Table 1.6. *Sorting search algorithm for an OFDMA system with two users.*

this approach attains the optimum solution. Similarly, an algorithm for joint resource allocation in case of discrete rate adaptation can be obtained by iterating between the sorting search subcarrier assignment and the greedy power allocation algorithm.

1.4.4. *Proportional fairness*

The aim of the proportional fairness optimization scheme is to find a trade-off between capacity and fairness. In contrast to approaches based on utility functions, the objective function is still the sum capacity but the proportional fairness is imposed through nonlinear constraints on the rates. In [SHE 05] a suboptimal algorithm for proportional fairness is proposed. Subcarrier allocation and power allocation are performed disjointly. First, the algorithm performs subchannel allocation assuming equal power distribution among subcarriers. Powers are allocated in two steps once the subchannels are assigned. The first step consists in applying the waterfilling algorithm to each user. This step provides the total power $P_{k,tot}$ to be allocated to user k as a linear function in p_{nk}, the power to be allocated to a predefined tone. The second step enables to determine the set of $P_{k,tot},\ k = 1, \dots, K$ which maximizes the total

Iterate until $\sum_{i\in[1,\dots K]} \frac{\mathrm{d}U_i\left(R_i^{(\ell)}\right)}{\mathrm{d}r}\left(R_i^{(\ell+1)}-R_i^{(\ell)}\right)\leq\epsilon$

1) Get new subcarrier assignment according to condition

$$\hat{m}(n)\leftarrow \arg\max_{i\in[1,\dots,K]}\left\{\phi_i^{(\ell)}c_i^{\boldsymbol{P}}[n]\right\}$$

2) Get new power allocation

$$p_{\hat{m}[n],n}\leftarrow\left[\frac{\phi_{\hat{m}(n)}^{(\ell)}}{\lambda}-\frac{1}{\beta\gamma_{\hat{m}(n),n}}\right]^+$$

$$R_i^{(\ell+1)}\leftarrow\sum_{k\in S_i}\log_2(1+\beta p_{in}\gamma_{in})$$

3) Update $\phi_i^{(\ell)}$ with positive step size $\mu\in(0,1)$

$$\phi_i^{(\ell+1)}\leftarrow(1-\mu)\phi_i^{(\ell)}+\mu\frac{\mathrm{d}U_i\left(R_i^{(\ell+1)}\right)}{\mathrm{d}r}$$

Table 1.7. *Joint resource allocation for continuous rate adaptation.*

Initialization: $R_k=0,\Omega_k=\varnothing,k=1,\dots,K,\ A=\{1,\dots,N\}$

For $k=1$ to K

find n satisfying $|\gamma_{k,n}|\geq|\gamma_{k,j}|,\ \forall j\in A$

$\Omega_k=\Omega_k\cup\{n\},\ A=A-\{n\}$ update R_k

While $A\neq\varnothing$

find k satisfying $R_k/\delta_k\leq R_i/\delta_i,\ \forall i\ 1\leq i\leq K$

for the found k, find n satisfying $|\gamma_{k,n}|\geq|\gamma_{k,j}|,\ \forall j\in A$

for the found n and k, let $\Omega_k=\Omega_k\cup\{n\},\ A=A-\{n\}$ update R_k

Table 1.8. *Proportional fairness in [SHE 05]: Subchannel allocation.*

rate under constraints on the total transmitted power and rate ratio. The expression of $P_{k,tot}$ and the system of equations for the constrained optimization of the total rate

$P_{k,tot}$ are in Table 1.9. Because their non linearity, the system of equations is solved by iterative approaches like the Newton-Raphson or quasi-Newton methods.

The subchannel assignment and the power allocation have a complexity $\mathcal{O}(KN)$ and $\mathcal{O}(K)$, respectively. Then, the complexity of the algorithm is linear in the number of users and tones in contrast to the exhaustive search which has a complexity in the order of K^N.

Let us express the total power constraint per user k by

$$P_{k,tot} = |S_k| p_{k,j} + \sum_{n \in S_k \diagup \{j\}} \frac{\gamma_{k,n} - \gamma_{k,j}}{\gamma_{k,n}\gamma_{k,j}} \qquad \text{for } j \in S_k$$

The total power and rate ratio constraints are given by

$$\frac{1}{\delta_1}\frac{|S_1|}{N}\left(\log_2\left(1+\gamma_{1,\ell}\frac{P_{1,tot}-V_1}{|S_1|}\right)+\log_2 W_1\right) = \frac{1}{\delta_k}\frac{|S_k|}{N}\left(\log_2\left(1+\gamma_{k,j}\frac{P_{k,tot}-V_k}{|S_k|}\right)+\log_2 W_k\right)$$

$$j \in S_k,\ \ell \in S_1$$

$$V_k = \sum_{n \in |S_k| \diagup \{j\}} \frac{\gamma_{k,n} - \gamma_{k,j}}{\gamma_{k,n}\gamma_{k,j}}$$

$$W_k = \left(\prod_{n \in S_k \diagup \{j\}} \frac{\gamma_{k,n}}{\gamma_{k,j}}\right)^{\frac{1}{|S_k|}} \qquad j \in S_k$$

With the power constraint

$$\sum_{k=1}^{K} P_{k,tot} = P_{tot}$$

Table 1.9. *Proportional fairness: total power constraints and capacity ratio constraints.*

1.4.5. *Max-min fairness*

The max-min fairness problem can be seen as a special case of proportional fairness. A specific algorithm has been proposed in [TOU 06] for the max-min fairness under the constraint of equal number of tones and users. The algorithm maximizes the minimum user rate over all possible allocations.

1.4.6. *Sum rate maximization in the uplink*

From a theoretic information perspective the problem of power allocation in a multiuser OFDM network has been investigated in [YU 04] and solved by an iterative waterfilling algorithm. The problem of maximum achievable rate in the uplink of an OFDMA single cell network can be modelled in a information theoretic setting as a FDMA Gaussian multiple access channel with intersymbol interference. It has been tackled in [YU 02].

The study of practical algorithms for resource allocation in uplink received much less attention than the resource allocation in the downlink OFDMA cell. The sum rate maximization problem with individual power constraints is investigated in [KIM 05a]. Subcarriers are allocated by applying a greedy algorithm and assuming a uniform power allocation or a waterfilling power allocation.

The greedy algorithm for joint subcarrier assignment and power allocation consists of the following steps:

Step 1 For each subcarrier n, which has not been allocated yet, and each user k calculate the assigned power p_{kn} that would be allocated to user k in subcarrier n assuming that are allocated to user k all the subcarriers already allocated to it in previous iterations and subcarrier n.

Step 2 Choose the pair (k^*, n^*) such that $(k^*, n^*) = \underset{(k^*,n^*)}{\operatorname{argmax}} p_{k,n}\gamma_{k,n}$.

Step 3 Repeat step 1 and step 2 until all subcarriers are allocated.

1.4.7. *Fair game-theoretic approach in the uplink*

In [HAN 05] an approach based on game theory is proposed with the aim of providing a minimum rate to each user while the overall system performance is optimized. The resource allocation is based on a cooperative game and make use of the Nash bargaining approach. The maximum sum rate problem is reformulated to include the constraints of the coalition problem. More specifically, a user k participates to a coalition only if a minimum rate $R_k^{(\min)}$ is guaranteed. The optimization problem is given by

$$\begin{aligned}
\text{maximize} \quad & \sum_{k=1}^{K} U(R_1, R_2, \ldots R_K) \\
\text{subject to} \quad & \sum_{k=1}^{K} \sum_{n \in S_k} p_{kn} \leq P_{tot}, \\
& S_j \cap S_k = \emptyset \quad \forall j \neq k \\
& \cup_{k=1}^{K} S_k \subseteq \{1, 2, \ldots, N\} \\
& p_{kn} \geq 0 \quad \forall k \text{ and } \forall n \qquad (1.17) \\
& \sum_{n=1}^{N} R_{k,n} \geq R_k^{(\min)} \quad \forall k. \qquad (1.18)
\end{aligned}$$

where $R_{k,n}\log_2(1 + p_{kn}\gamma_{kn})$, $R_k = \sum_n R_{k,n}$, and $U(R_1, R_2, \ldots R_K)$ is a utility function properly chosen according to the Nash bargaining approach. In [HAN 05] it is shown that a convenient definition for the Nash bargaining game is

$$U(R_1, R_2, \ldots R_K) = \prod_{k=1}^{K} (R_k - R_k^{(\min)}).$$

Assigned the set of minimum rate for all users $\{R_{min}, k = 1, \ldots K\}$ the Nash bargaining solution finds a rate allocation such that no other allocation leads to superior performance for some users without implying inferior performance for some other users. The problem reduces to the proportional fairness problem when $R_k^{(\min)} = 0$ In [HAN 05] an algorithm for two users is proposed. For the multiuser case the algorithm can be generalized in a non-centralized way, by a two-step iterative process. First, users are grouped in pairs (either randomly, or using the well known Hungarian method, which will optimize the grouping to reduce the convergence time), named coalitions, and the two-user algorithm is applied to each of the pairs. Then, players are regrouped and the process is iterated until convergence is achieved. The bargaining can be done at the base station without incurring in any signaling overhead between users and base station.

In [HAN 05] the Nash bargaining solution is compared with the maximal rate approach where the utility function is $U(R_1, R_2, \ldots R_K) = \sum_k R_k$ and the max-min fairness when the utility function is $U(R_1, R_2, \ldots R_K) = \min_k R_k$. In figures 1.7 presents the performance of the Nash bargaining approach as in the assessment in [HAN 05]. The performance for the two user case is presented and compared with

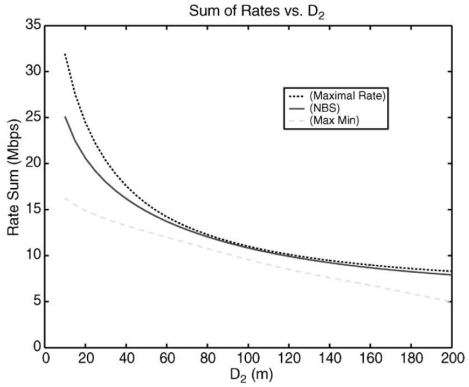

Figure 1.7. *Nash bargaining approach for fair resource allocation in uplink: sum rate versus BS-moving user distance.*

two other allocation strategies. We assume that a user is kept at a fixed distance from the base station while the other user changes position. The sum capacity and the individual rate are plotted as function of the distance d of the second user. The Nash bargaining approach show a large performance improvement compared to the the hard fairness approach (max-min). Although it still guarantees a minimum achievable rate to all the users, it provides a sum rate very close to the optimum achievable rate.

1.5. Enhancements in single-cell networks

1.5.1. *Multiple antenna arrays at the transmitters and the receivers*

Communication systems with multiple antenna arrays at the transmitters and the receivers are referred as multiuser MIMO (multiple input multiple output) systems and provide spatial multiplexing and diversity. They increase capacity thanks to the multiplexing gain obtained by the several inputs and outputs, and robustness via diversity, as several copies of each frame will be received at each receiver. A trade-off between capacity and robustness can also be envisaged.

The use of multiple antennas at both the transmitter and the receiver side can provide a huge increase in the throughput of wireless communication systems [TEL 95, GOL 99]. For example, in the case of MIMO Rayleigh channels with n_t transmitting antennas and n_r receiving antennas, perfect knowledge of the channel at the receiver and no channel knowledge at the transmitter, the ergodic capacity increase is known to be min(n_r, n_t) bits per second per hertz for every 3dB increase at high SNR while only one bit per second per hertz can be gained by increasing the SNR of 3 dB in

the additive white Gaussian channel with single transmitting and receiving antennas at high SNR [TEL 95].

Multiple antennas can be used to reduce the error probability at the receiver for a given data rate or to increase the data rate for a given error probability. The first effect is known as diversity gain, the latter is referred to as multiplexing gain or degree of freedom gain. More specifically, let us consider the capacity of an additive Gaussian noise at high SNR, given by $\log \mathrm{SNR}$, the multiplexing gain of a code with data rate R is defined by $r = \frac{R}{\log \mathrm{SNR}}$. In order to define the diversity gain let us consider the error probability P_e of a code with rate R at the output of a maximum likelihood detector: if P_e decays as SNR^{-d} then the code has diversity gain d. The fundamental tradeoff between multiplexing gain and diversity gain in a point-to-point system has been characterized in [ZHE 03]. For i.i.d. Rayleigh fading the best decay rate for a given multiplexing gain is given by

$$d^*_{n_t,n_r}(r) = (n_t - r)(n_r - r) \text{ for } r \text{ integer and } r \leq \min(n_t, n_r). \tag{1.19}$$

The entire curve $d^*_{n_t,n_r}(r)$ is piecewise linear joining the points in (1.19). The largest achievable multiplexing gain for a given diversity gain d is the inverse of $d^*_{n_t,n_r}(r)$ and it is denoted by $r^*(d)$. The maximal diversity gain is $n_t n_r$ achievable for $r \to 0$. The maximal multiplexing gain is $\min(n_t, n_r)$ attained for $d \to 0$.

The analysis is further complicated when we consider multiple access channels whose sources are equipped with n_t transmitting antennas and the destination is equipped with n_r receiving antennas. In this case the diversity gain and the multiplexing gain typical of MIMO systems are combined with the multiple-access gain of multiple access channels. The tradeoff among the three different kinds of gains is investigated in [TSE 04].

In the multiple access channel we consider K users, each of them equipped with n_t transmitting antennas. If the user k uses a code with rate R_k the multiplexing gain is

$$r_k = \log \frac{\mathrm{SNR}_k}{R_k}.$$

To analyze the diversity-multiplexing-multiple access tradeoff the minimal error probability for each user at the output of an individual maximum likelihood detector is required to decay at least as fast as SNR^{-d}. The t-uple $(r_1, r_2, \ldots, r_K)$ of the multiplexing gains is provided in [TSE 04]. In [TSE 04] the symmetric situation is also investigated, i.e. for a minimum multiplexing gain r common to all users, the t-uple of diversity gains is provided. In this case it is shown that the maximal multiplexing gain achievable by each user is $\min\left(n_t, \frac{n_r}{K}\right)$. Then, at least concerning the maximal multiplexing gain insightful results are available. Within the range of achievable multiplexing gains, the tradeoff on the performance can be divided into two regimes: (i) lightly loaded regime where the system behaves as if only one user is in the system,

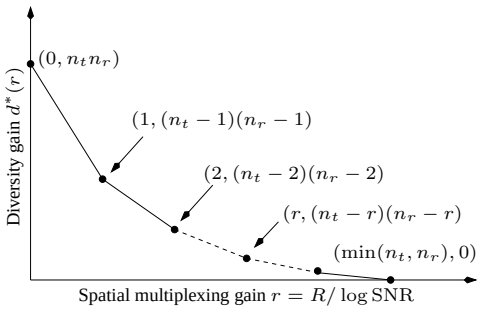

Figure 1.8. *Diversity multiplexing tradeoff for $n_t \leq \frac{n_r}{K+1}$ is the same as the single user curve*

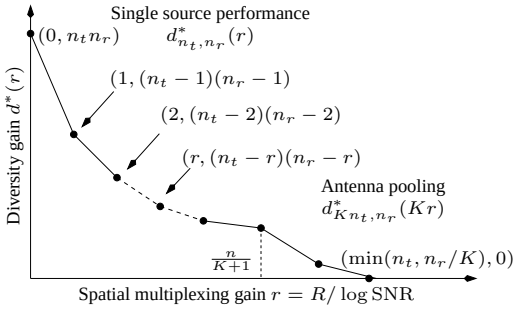

Figure 1.9. *Diversity multiplexing tradeoff for $n_t \geq \frac{n_r}{K+1}$. It is the same as the single user curve for $r \leq \frac{n}{K+1}$. For $r \geq \frac{n}{K+1}$ the curve corresponds to the antenna pooled curve.*

i.e. $d^*_{n_t,n_r}(r)$, and (ii) heavily loaded regime where the system behaves as if the K users pool up their transmit antennas together. The diversity multiplexing tradeoff is illustrated in Fig. 1.8 for $n_t < \frac{n_r}{K+1}$ and in Fig. 1.9 for $n_t > \frac{n_r}{K+1}$.

When we focus on OFDM/OFDMA systems, the analysis of the tradeoff diversity-multiplexing gains determined a floury of activities to determine codes achieving the above mentioned tradeoff.

Resource allocation algorithms have been developed for MIMO-OFDMA systems. In [PAN 04] the sum power is minimized subject to individual rate constraints for all users. A joint power allocation and subcarrier assignment is proposed based on dirty paper encoding. The same problem is investigated in [ZHA 05] but in this case the constraints are expressed not only in terms of data rate but also in terms of maximum bit error rate. In [LO 07] the dual problem of maximizing the the sum capacity under power constraint is investigated in a cross-layer optimization framework. Power

allocation and subcarrier assignment algorithms are discussed with and without fairness constraints. Resource allocation maximizing the sum capacity subject to total power and proportional rate constraints or weighted proportional rate constraints are provided in [XU 06] and [MOR 06], respectively.

1.5.2. *Bitloading*

Classical wireless multicarrier systems use the same and unique fixed input constellation across all subcarriers, thus the overall error probability is limited by the "poorest" subcarriers, i.e. subcarriers presenting the worst performance e.g. in terms of Signal-to-Noise Ratio (SNR). First devised for static channels in the context of xDSL transmissions, the principle of bitloading consists in loading *adapted*[5] constellation size on each subcarrier. The association of multicarrier modulation with bitloading is known under the acronym DMT for Discrete MultiTone. In wireline applications where bit loading is traditionally used, the channel can be assumed to be quasi-static and the bit and power allocations may not change for a long time. Therefore the algorithmic complexity may not be a problem. However, in wireless environments, the channel is time varying, and the loading algorithm must be computationally efficient so that the transmitter can update the bit and power distributions quickly enough to track the channel variations.

Many bitloading algorithms exist as different ways to implement the solution of different constrained optimization problems. These can be classified in many ways, for instance regarding the objective function that is optimized (power, channel capacity, bit error probability). This objective function is generally associated with a constraint. Common choices are the maximization of the "throughput" given a power constraint known as rate adaptive loading ([KAL 89]), and the minimization of the energy given a fixed throughput requirement, known as margin adaptive loading ([CHO 95]). In ([CAM 98]), optimality conditions are introduced. In rate adaptive loading, the power constraint can be either individual (uplink transmission) or global (downlink transmission). It can also deal with total transmitted power or maximum power spectral density, or even both ([BAC 02]). In both cases, an error rate constraint is obviously considered. This constraint can be on the mean error rate over all subcarriers or on each subcarrier. Regarded as a performance metric, different throughput definitions can be used: Shannon capacity, sometimes shifted by a gap to reach some error rate requirement, mutual information adapted to some input signal distribution, etc. From a practical point of view, these generally have to be rounded to map with integer bit allocation. If this notion is integrated from the problem formulation, this yields integer programming problems. "Greed" is another notion related to integer bit allocation; in this approach, closed form expressions of performance measures are

5. w.r.t. system parameters

not used and bits are basically allocated in a successive way (“bit-filling”) to subchannels on which the power increment required to transmit an additional bit is minimal, until the power constraint is reached. Based on this iterative method, several loading algorithms have been proposed in the literature ([HUG 87, FAS 03]). However, the algorithm complexity often makes it almost inapplicable for practical applications, especially when the system has large number of subcarriers. The complexity can be reduced by increasing the problem granularity, i.e. considering blocks of subcarriers instead of subcarriers taken individually.

1.6. Resource allocation in multicell OFDMA networks

As shown in the previous sections the investigation of single cell OFDMA networks has attracted many efforts and the understanding of a single cell system is thorough and has reached a mature stage. On the contrary, the research on multicell OFDMA networks is still in its infancy. This is mainly due to the very limited information theoretical knowledge about the interference channel, i.e. a channel where two users transmit independently to two different destinations, and each destination is interested in the information only of one user although it receives signals from both users.

In practice, the interference problem in multicell OFDMA networks is solved by frequency reuse. The full system band is divided in F disjoint bands and adjacent cells communicate on different bands. The number of different bands F is called reuse factor and a proper and careful deployment of the base stations minimizes the intercell interference. Under this conditions the intercell interference can be neglected or modelled as an additive white Gaussian noise and the resource allocation reduces to the resource allocation in a single cell problem. In this direction, the most of research efforts have been devoted to the optimization of the frequency reuse factor. It is well-known that the frequency reuse approach has major drawbacks as the huge spectral efficiency cost, the need of a costly planning for cell deployment, and a difficult re-planning when the introduction of additional cells in the network is required. In [wima], [JIA 07], [JOO 07], the frequency reuse method is refined in fractional frequency reuse techniques. In this latter approach, the full available band is assigned to users in the internal part of the cell while frequency reuse is applied only at the edge of the cells. This improves the spectral efficiency compared to the frequency reuse technique but still implies a considerable loss.

Recently, the concept of random frequency reuse has been introduced in [SAA]. This work focuses on the downlink and the resource allocation algorithm is developed for the downlink. However, the concept of random frequency reuse can readily be extended to the uplink channel. The resource allocation algorithm proposed in [SAA] enables a random factor reuse in a cell based on the actual channel conditions. A base station in a cell in a given tone is activated only if the global capacity of the

network increases by allocating a user in such a subcarrier. Under the assumptions of large, dense, interference limited (i.e. with negligible background noise compared to the interference) networks, and binary level of allocated powers the power allocation algorithm and the algorithm to determine the active cells for a given tone decouple. Utilizing the averaging properties of large dense networks on the interference level at each user, a knowledge of the global channel state information is not necessary for resource allocation. This makes feasible a distributed implementation of the algorithm based only on the knowledge of the channel gains between a base station and the end users. This dynamic spectral reuse allows a large improvement in performance with respect to fixed frequency reuse schemes.

Resource allocation in the uplink channel of a multicell OFDMA network is tackled in [MOR 07]. The resource allocation is based on relaxation methods. The multicell power allocation problem is reduced to a single cell resource allocation problem using as interference in a certain frame the estimate of the interference in the previous frame. However, no statistical model for the interference process is assumed and the instantaneous resource allocation in a given frame does not take into account the simultaneous allocations in other cells.

1.7. Achievable rates and resource allocation in OFDMA networks with relays

The relays capabilities of nodes in WiMAX enable the enhancement of the network capacity and coverage through cooperation among nodes and, eventually, cooperative diversity.

From an information theoretic point of view a basic relay channel consists of three nodes, a source, a relay, and a sink and was introduced first by van der Meulen in [MEU 71]. The source node transmits its data stream to the relay and the sink. The relay sends the information received by the source to the destination with a symbol interval. The fundamental limits of this basic model have been investigated first by Cover and El Gamal in [COV 79] and the potential of a relay channel to improve the overall performance have been shown.

In [COV 79] the analysis is limited to degraded relay channels and it is shown that the capacity is achieved by a block-Markov chain scheme with an infinite number of blocks and decode-and-forward (DF) strategies. In the DC strategy the relay decodes the source information, re-encodes it, and transmits it to the sink. Cover and El Gamal proposed also a quantized-and-forward strategy for general relay channels. In the quantized-and-forward strategy the relay quantizes its received signal and transmits a compressed version to the destination. The analog version of the quantized-and-forward approach is the well-know amplify-and-forward (AF) strategy where the relay simply retransmits its received signal. The most practical relay strategies are based on these schemes.

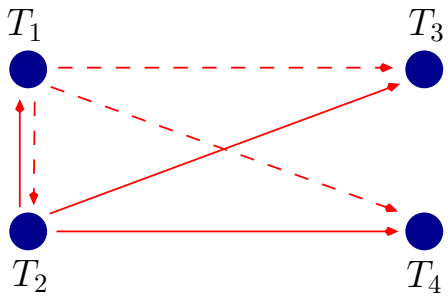

Figure 1.10. *Cooperative diversity*

In [REZ 04] capacity and power allocation of degraded Gaussian multirelay channels have been investigated assuming an infinite number of hops. Recently, the relay channel has been object of intensive studies. However, the exact capacity of Gaussian or general relay channels is still unknown. In [KOH 04] and [GAM 06] upper and lower bounds on the capacity of Gaussian relay channels are provided. Fading relay channels have been object of studies in [KRA 05, WAN 05a, HØS 05, YAO 05]. Relay channels with multiple antennas are investigated in [KRA 05, WAN 05a].

Beside the original scheme of the relay channel in [MEU 71], the relay channel with orthogonal components has received also attention (e.g. [LIA 05, HØS 05, GAM 05, KRA 04]). In this case the source transmits to the relay and the destination in channel 1 and the relay transmits to the destination in channel 2, with channel 1 and channel 2 being orthogonalized in the time-frequency plane. Relay channels with orthogonal components are the basic blocks for a special kind of spatial diversity referred to as cooperative diversity in literature ([SEN 03a, SEN 03b, LAN 00, LAN 01, LAN 03, LAN 04]).

In contrast to the classical systems with spatial diversity based on physical arrays, systems with cooperative diversity create and exploit space diversity using a collection of distributed antennas belonging to multiple terminals, each with its own information to transmit and relays capabilities. To illustrate the concept let us consider Fig. 1.10.

Let assume that T_1 and T_2 are the handsets and T_3 and T_4 are the base stations with eventually $T_3 = T_4$. T_1 and T_2 can listen to each other's transmissions to the base stations thanks to the broadcast nature of the channel and then jointly communicate their information. In this way T_1 and T_2 behave as a virtual array for the transmission of information both from T_1 and T_2.

In mobile multihop relay networks the resource allocation problem is further exacerbated by the possibility of cooperation among nodes. Compared to the resource

allocation problem for a single cell OFDMA network stated in Section 1.2, the resource allocation problem in relay networks requires the choice of convenient relay nodes, relay strategies (AF, DF, etc.), and power and subcarrier allocation. In case the users terminals may play the role of source/destinatination of the communication or the role of relay nodes the problem is formulated as follows.

Let $\mathcal{K} = \{1, 2, \dots K\}$ be the set of users nodes. Denote the base station as node $K+1$. Let $\mathcal{K}_+ = \{1, 2, \dots K+1\}$ be the extended set of nodes and let $\mathcal{N} = \{1, 2, \dots N\}$ be the set of tones. We assume that each of the K user nodes has both upstream and downstream communications with the base station. Let (s, d) be the source destination pair, or data stream. The set of data stream is $\mathcal{M} = \{(1, K+1), (2, K+1), \dots (K, K+1), (K+1, 1), (K+1, 2), \dots (K+1, K)\}$ with cardinality $|\mathcal{M}| = 2K$. Let us denote by $\boldsymbol{P}$ the $(K+1) \times N$ matrix with (k, n) element equal to p_{kn}, the power allocated to user k on tone n. This matrix can have at most two nonzero entries in each column, one for the source and one for the relay in the classic formulation of the relay channel. In case of relay channel with orthogonal component the matrix has at most a single component in each column. Similarly, let $\boldsymbol{R}$ be a $2K \times N$ matrix whose (m, n)-element is the data rate of stream m on tone n ($n \in \mathcal{N}$ and $m \in \mathcal{M}$). Since only one stream can be active in each tone, each column vector of $\boldsymbol{R}$ has at most one nonzero entry. If $\vec{1}$ denotes an N-dimensional column vector with all unit elements $(\boldsymbol{P}\mathbf{1})_i$, $i \in \mathcal{K}_+$, the i-th component of the vector $\boldsymbol{P}\mathbf{1}$ is the total power expended at node i-th over all tones. Similarly $(\boldsymbol{R}\mathbf{1})_m$, $m \in \mathcal{M}$, gives the total data rate of stream m, summing across all tones. Let $\vec{p}^{\,\mathrm{MAX}} = [p_1^{\mathrm{MAX}}, p_2^{\mathrm{MAX}}, \dots p_{K+1}^{\mathrm{MAX}}]^T$, where $\vec{p}^{\,\mathrm{MAX}}$ is the individual power constraint for node i. The joint optimization problem is

$$\begin{aligned} \max_{\boldsymbol{P}, \boldsymbol{R}} \quad & \sum_{m \in \mathcal{M}} U_m((\boldsymbol{R}\mathbf{1})_m) \\ s.t. \quad & \boldsymbol{P}\mathbf{1} \preccurlyeq \vec{p}^{\,\mathrm{MAX}}, \ \boldsymbol{R} \in \mathcal{C}(\boldsymbol{P}) \end{aligned}$$

where U_m is the utility function of data stream m, function of the achievable rate of stream m, $(\boldsymbol{R}\mathbf{1})_m$, and $\mathcal{C}(\boldsymbol{P})$ denotes the achievable rate region.

This general problem has been investigated in [NG 07] under the constraints that a stream $m \in \mathcal{M}$ is transmitted by the source and the relay on the same tone and the channel is slow fading, flat on each tone. It is shown that the global optimization problem can be decomposed into two suboptimum problems, namely a utility maximization problem, corresponding to a rate adaptation problem at the application layer

$$g_{appl}(\lambda) = \max_{\vec{t}} \sum_{m \in \mathcal{M}} (U_m(t_m) - \lambda_m t_m)$$

and a joint relay strategy (decode and forward, amplify and forward) and relay node selection and power and bandwidth allocation at the physical layer

$$g_{phy}(\lambda) \begin{cases} \max_{\boldsymbol{P},\boldsymbol{R}} \lambda_m \sum_{n \in \mathcal{N}} R(m,n) \\ \text{s.t.} \boldsymbol{P}\mathbf{1} \preccurlyeq \vec{p}^{\text{MAX}}, \boldsymbol{R} \in \mathcal{C}(\boldsymbol{P}). \end{cases}$$

Their proposal is a centralized utility-maximization framework, at the physical layer, in relationship with user traffic (cross-layer design). They make use of the pricing variables λ_m as weighing factors. The result is optimal bandwidth and power allocation bandwidth for each user as well as selection of best relay node and best relay strategy for each source-destination pair.

Chapter 2

WiMAX network capacity and radio resource management

Tijani Chahed (GET/INT), Ikbal Chammakhi Msadaa (Eurecom), Rachid Elazouzi (LIA), Fethi Filali (Eurecom), Salah-Eddine Elayoubi (France Telecom R&D), Benoit Fourestié (France Telecom R&D), Thierry Peyre (LIA), Chadi Tarhini (GET/INT)

2.1. Survey on RRM proposals

IEEE 802.16 BWA technology is emerging as a promising solution that provides QoS guarantees for heterogeneous classes of traffic with different QoS requirements. It offers the possibility of adapting the modulation and coding schemes based on the channel conditions and proposes a set of mechanisms such as packing and fragmentation to allow efficient use of the available bandwidth. The standard however leaves open the resource management and scheduling issues.
In this section, the majority of scheduling and CAC solutions proposed by researches for IEEE 802.16 systems during the last years is being presented. We first provide an overview of the main features proposed by the standard to support QoS and then outline the challenges that should be addressed when designing a new scheduling or CAC solution. Along with the description of each proposal, a comparison outlining the advantages and limits of each solution is being presented.

2.1.1. *IEEE 802.16 QoS support*

The IEEE 802.16 Standard [802 04] specifies the air interface for fixed BWA systems in the frequency ranges 10-66 GHz and sub 11 GHz. The standard covers both

the Media Access Control (MAC) and the physical (PHY) layers. The 802.16 MAC layer was designed to accommodate different PHYs and services, which address the needs of different environments. In this survey, systems of interest are those operating at frequencies below 11 GHz, where LOS is not required.

The basic topology of an IEEE 802.16-based network consists of one Base Station (BS) and one or more Subscriber Stations (SSs). In PMP, which is the only mode for sharing media considered in this survey, the SSs within a given antenna sector receive the same transmission broadcast by the BS—corresponding in general to the Internet Service Provider (ISP)—on the downlink channel (DL). Each SS is required to capture and process only the traffic addressed to itself (or to a broadcast or multicast group it is a member of). On the uplink channel (UL) however, the Time Division Multiple Access (TDMA) scheme is applied. Downlink and uplink channels are duplexed using one of the two following techniques: Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD). The main difference between the two duplex modes is that in FDD, the DL and UL use different frequencies, while in TDD both channels use the same frequency in different time intervals.

The standard defines a connection-oriented MAC protocol where all the transmissions occur within a context of a unidirectional connection. Each connection, identified by a unique Connection ID (CID), is associated to an admitted or active service flow (SF) whose characteristics provide the QoS requirements to apply for the PDUs exchanged on that connection. There are three types of service flows: (a) provisioned service flows for which the QoS parameters are provisioned for example by the network management system, (b) admitted service flows for which resources, mainly bandwidth, are reserved and (c) active service flows which are activated to carry traffic using resources actually provided by the BS. Each service flow is uniquely identified by a SFID; admitted and active service flows have also a CID. Service flows may be dynamically managed; they may be created, changed or deleted using DSA, DSC and DSD MAC management messages, respectively. The SF management procedure consists actually in exchanging DSx-REQ, DSx-RVD—sent by the BS when the transaction is SS-initiated—DSx-RSP and DSx-ACK messages, between the BS and the SS. Note that initiating the creation of a new service flow is a mandatory capability for a BS and an optional one for an SS. As mentioned above, a service flow defines the QoS that should be provided by the SS and BS to the packets traversing the MAC interface and which are associated to that SF. In order to facilitate the MAC SDUs delivery with the appropriate QoS constraints, the IEEE 802.16 Standard defines a classification process by which a MAC SDU is mapped to the associated connection and so to the SF related to that connection. The classification procedure is performed by classifiers consisting of a set of protocol-specific matching criteria.

Depending on the service to be tailored to each user application, the connection is associated with one of the following scheduling services supported by the 802.16 MAC protocol: Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS),

Extended Real-time Polling Service (ertPS)—introduced by the IEEE 802.16e-2005 standard [802 05], Non-real-time Polling Service (nrtPS), and Best Effort (BE). Each scheduling service is designed to meet the QoS requirements of a specific category of applications.

– UGS is designed to support real-time applications that generate fixed-size data packets at periodic intervals, such as T1/E1 and VoIP without silence suppression. The mandatory service flow QoS parameters for UGS service are listed in Table 2.1; this table summarizes, according to the scheduling service type, the QoS parameters that must be specified when establishing a new service flow. UGS connections never request bandwidth; the amount of bandwidth to allocate to such connections is computed by the BS based on the Minimum Reserved Traffic Rate defined in the service flow of that connection.

– rtPS is designed to support real-time applications that generate variable-size data packets at periodic intervals, such as moving pictures expert group (MPEG) video. Unlike UGS connections, rtPS connections must inform the BS of their bandwidth requirements. Therefore the BS must periodically allocate bandwidth for rtPS connections specifically for the purpose of requesting bandwidth. This corresponds to the polling bandwidth-request mechanism. This mechanism exists in three variants: unicast polling, multicast polling and broadcast polling. Only unicast polling can be used for rtPS connections.

– Extended rtPS is a new scheduling service introduced by IEEE 802.16e-2005 standard [802 05] to support real-time service flows that generate variable size data packets on a periodic basis, such as Voice over IP services with silence suppression. Like in UGS, the BS shall provide unicast grants in an unsolicited manner which saves the latency of a bandwidth request. However, unlike UGS allocations that are fixed in size, ertPS allocations are dynamic like in rtPS. By default, the size of allocations corresponds to current value of Maximum Sustained Traffic Rate at the connection. The SS however may request changing the size of the UL allocation.

– nrtPS is designed to support delay-tolerant applications such as FTP for which a minimum amount of bandwidth is required. The polling mechanism can be applied to nrtPS connections. However, unlike for rtPS, nrtPS connections are not necessarily polled individually—multicast and broadcast polling are possible—and the polling must be regular not necessarily periodic.

– BE is designed for applications that do not have any specific bandwidth or delay requirement, such as HTTP and SMTP. For BE connections, all forms of polling are allowed in order to request bandwidth.

The QoS parameters that must be specified when establishing a new service flow are listed in Table 2.1. The value of the Request/Transmission (Rx/Tx) Policy parameter—that should be specified in each service flow— offers the possibility to specify, for the corresponding service flow, options for PDU formation such as restriction on packing

and fragmentation capabilities as well as attributes affecting the bandwidth request types.

	UGS	rtPS	ertPS	nrtPS	BE
Maximum Sustained Traffic Rate	X	X	X	X	X
Minimum Reserved Traffic Rate	—	X	X	X	—
Maximum Latency	X	X	X	—	—
Tolerated Jitter	X	—	—	—	—
Traffic Priority	—	—	—	X	X
Rx/Tx Policy	X	X	X	X	X

Table 2.1. *Mandatory QoS parameters for each scheduling service*

Indeed to inform the BS of its uplink bandwidth requirement, the SS may send a stand-alone bandwidth request header or just piggyback the request on a PDU, which is an optional capability. Other mechanisms such as bandwidth stealing and the use of poll-me (PM) bit[1] are also specified by the IEEE 802.16 Standard. It is important to mention that, whatever be the bandwidth request mechanism in use, bandwidth is always requested by an SS on a per-connection basis, it is nevertheless granted by the BS to an SS as an aggregate of grants. Therefore, since the SS receives the allocated bandwidth as a whole in response to a per-connection requests, it cannot know which request is honored. The SS can then use the grant—specified in a Data Grant IE—either to send data or management messages or even to request bandwidth for any of its connections.

2.1.2. *Scheduling and connection admission control challenges*

When designing a new scheduler for 802.16 systems, it is important to understand the challenges faced not only in any wireless network but also in those that are specific to 802.16 technology. In this Section we focus mainly on the latter category of challenges. Our objective is to outline a set of features that are specified by the IEEE 802.16 standard and that should be supported by an 802.16 scheduler:

– The scheduler should satisfy the QoS requirements, illustrated in Table 2.1, of the different classes of service specified by the standard.

– The scheduler should fairly redistribute the available resources among the different service flows while taking into account their respective modulation and coding scheme.

1. A field of a specific subheader of a MAC PDU, used by the SS to request a bandwidth poll for a non-UGS connection.

– Since bandwidth allocation are made on a per-SS basis, a scheduler should be integrated in the MAC structure of an SS;

– When making a scheduling decision, the scheduler should take into account the resulting MAC and physical overhead; It should also take advantage of concatenation, packing, and fragmentation mechanisms, proposed by the standard, to make efficient use of the available resources.

– It should adopt a bandwidth polling (at the BS) and requesting (at the SS) policy;

– When considering the TDD mode, the amount of bandwidth allocated for uplink and downlink should be dynamically adapted to the traffic transmitted on each direction;

– The scheduler should take into account the dynamic aspect of a service flow. Indeed a SF may be added, updated, or deleted. The scheduler should also make the difference—in terms of resource allocation—between a provisioned SF, an admitted SF and an active SF;

All these challenges should be addressed when designing a new scheduling solution for IEEE 802.16 networks. The complexity of the proposed algorithm should nevertheless be implementation-friendly.

2.1.3. *Scheduling proposals*

As shown in Figure 2.1, the approaches adopted in literature when designing a scheduling solution can be divided into two main categories. The first one is a queuing-derived strategy where the authors focus on the queuing aspect of the scheduling problem and try to find the appropriate queuing discipline that meet the QoS requirements of the service classes supported by the IEEE 802.16 standard [802 04, 802 05]. In this first category, two kinds of structures are proposed: either simple structures consisting in general in one queuing discipline applied for all the scheduling services [CIC 07, CIC 06, SAY 06] or hierarchical structures consisting in two or multiple layers reflecting different levels of scheduling like in [CHA 06a, CHE 05, LIU 05, PER 06, SET 06, SUN 06, WON 03b, WON 03a]. In the second category, the scheduling problem is formulated as an optimization problem whose objective is to maximize the system performance subject to constraints reflecting in general the QoS requirements of different service classes [NAS 04, NIY 05a, NIY 06a, NIY 05b, NIY 06b, NIY 06d, NIY 06e, NIY 06c, SIN 06].

2.1.3.1. *Packet queuing-derived strategy*

Simple scheduling structures

Sayenko *et al* [SAY 06] believe that because there is no much time to do the scheduling decision, a simple one-level scheduling mechanism is much better than a

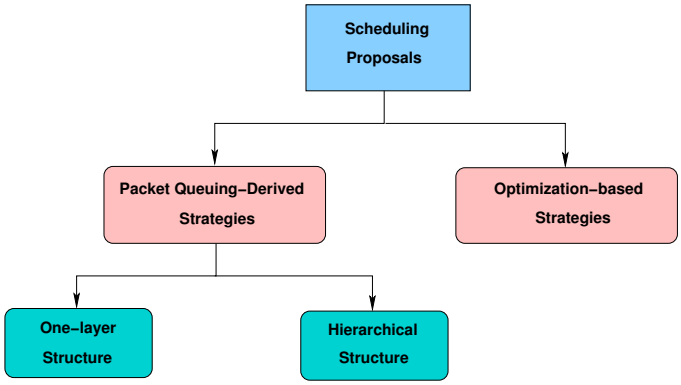

Figure 2.1. *Classification of the scheduling strategies*

hierarchical one. Therefore they proposed a scheduling solution based on the Round-Robin (RR) approach. They argued that there is no need to use disciplines like Fair Queuing (FQ) since the weights in such algorithms are floating numbers while the number of allocated slots, in 802.16 networks, should have an integer value. They also tried to outline the difference between the Weighted Round-Robin (WRR) discipline and the 802.16 environment. They insist on the fact that WRR may lead to a waste of resources because of its work-conserving behavior that does not fit the fixed-size frame of 802.16 that implies a non-work conserving behavior.
Based on the above considerations, the authors proposed in [SAY 06] a scheduling solution that consist in three main steps:

– Allocating for each connection the minimum number of slots that ensure the minimum reserved traffic rate with respect to the used modulation and coding scheme,

– Distributing the free slots between rtPS and nrtPS connections and then assigning the remaining to BE connections,

– Ordering the slots in such a manner the delay and jitter values are decreased.

– Estimating the overhead for UGS, ertPS, and in some cases nrtPS connections. It is not possible for rtPS and BE connections where it is more likely that the SDU size vary.

Note that [SAY 06] is one of the rare research works in which the overhead resulting from the scheduling decision, and packing or fragmentation capability is taken into account. However it is also worth mentioning that the authors consider a GPC mechanism and when ordering slots, they apply an interleaved scheme that is in contradiction with the frame structure specified by the standard.

In [CIC 06, CIC 07], Cicconetti *et al* conjecture that the class of latency-rate (LR) scheduling algorithms is particularly suited for implementing schedulers in 802.16

MAC since the basic QoS parameter required by a given connection is the minimum reserved traffic rate. Indeed the behavior of such algorithms is determined by two parameters which are the latency and the allocated rate [STI 98]. From this class, the authors have chosen the deficit round robin (DRR) algorithm. DRR is simple to implement ($O(1)$ complexity if specific allocation constraints are met) and provides, according to [CIC 06, CIC 07], fair queuing in presence of variable length packets[2]. It nevertheless requires a minimum rate to be reserved for each packet flow; so even BE connections should be guaranteed a minimum rate. Also since this algorithm assumes that the size of the head-of-line packet is known, it can not be applied by the BS to schedule uplink transmissions. For this reason the authors have made the choice of implementing it as SS scheduler and as a downlink scheduler at the BS, since both BS and SS know the head-of-line packets sizes of their respective queues. To schedule uplink transmissions—based on backlog estimation—they have selected the WRR algorithm which belongs, like DRR, to the class of LR algorithms.
The simulation study carried by Cicconetti *et al* [CIC 07] demonstrated that the performance of 802.16 systems, in terms of throughput and delay, depends on several metrics such as frame duration, the mechanisms used to request UL bandwidth, the offered load partitioning—how traffic is distributed among SSs, the connections within each SS, and the traffic sources within each connection.

Hierarchical scheduling structures
To the best of our knowledge, Wongthavarawat and Ganz [WON 03b, WON 03a] are the first authors who introduced a hierarchical structure of bandwidth allocation for 802.16 systems. This hierarchical scheduling structure, shown in Figure 2.2, combines strict priority policy, among the service classes, and an appropriate queuing management discipline for each class: EDF for rtPS, and WFQ for nrtPS. Fixed time duration is allocated to UGS connections and remaining bandwidth is equally shared among BE connections. In order to avoid starvation for lower priority connections, a policing module is included in each SS. It forces each connection to respect the traffic contract when demanding bandwidth. The proposed scheduling algorithm takes into account the queue size information and the service actually received by each connection. It also considers the arrival time and the deadline requirements of rtPS connections. However, the authors focused only on UL scheduling. They considered TDD mode and assumed that the durations of UL and DL subframes are dynamically determined by the BS but they did not specify how these proportions are fixed. The QoS architecture they proposed in [WON 03b] includes a token-bucket based admission control module that will be described in Section 2.1.4.

In [SUN 06], the authors proposed a two-layers scheduling structure composed of a BS scheduler and an SS scheduler. At BS scheduler, priority is given to schedule

2. This is in contradiction to what has been stated by Fattah and Leung in [FAT 02] where they qualify the fairness of DRR algorithm as “poor”.

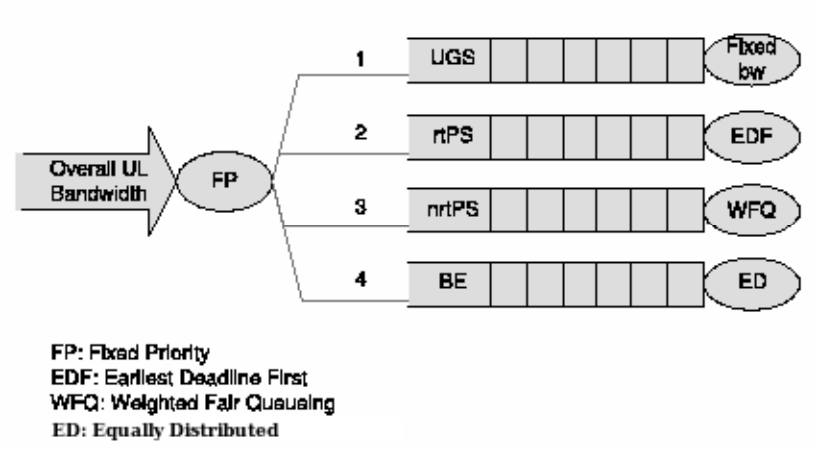

Figure 2.2. *Hierarchical structure for bandwidth allocation [WON 03b, WON 03a]*

data grants for UGS connections and bandwidth request opportunities for rtPS and nrtPS connections. The amount of bandwidth allocated in this phase is reserved during connections' setup. Data grants for rtPS, nrtPS are then scheduled taking into account the information contained into bandwidth request messages and their minimum requirements. Finally, the residual bandwidth, if any, is redistributed in proportion to pre-assigned connections' weights. The proposed SS scheduler considers a fixed priority scheme—1, 2, 3 and 4 for BE, nrtPS, rtPS and UGS scheduling service, respectively. Bandwidth is firstly guaranteed for UGS connections. rtPS packets are then scheduled based on their respective deadline stamps—corresponding to their $arrival\_time + tolerated\_delay$. Each nrtPS packet is associated with a virtual time calculated to guarantee the minimum reserved bandwidth and hence maintain an acceptable throughput. A simple FIFO mechanism is applied for BE queues.

Other scheduling structures focusing on delay requirements were proposed in literature. In [LIU 05] for instance, three schedulers were combined to meet the QoS requirements of different classes (Figure 2.3). Time sensitive traffic streams—consisting in UGS flows, rtPS flows and (n)rtPS polling flows—are served by Scheduler 1 that applies EDF algorithm. Minimum bandwidth reserving flows (nrtPS flows) are scheduled by Scheduler 2 using WFQ. The weights correspond to the proportion of requested bandwidth. WFQ algorithm is also applied by Scheduler 3 to serve BE traffics; weights nevertheless correspond in that case to traffic priorities specified by each BE connection. Other components of the proposed architecture (Figure 2.3) are then used to plan contention and reserved transmission opportunities according to the bandwidth availability and to the priorities assigned to each scheduler—the highest priority is assigned to Scheduler 1.

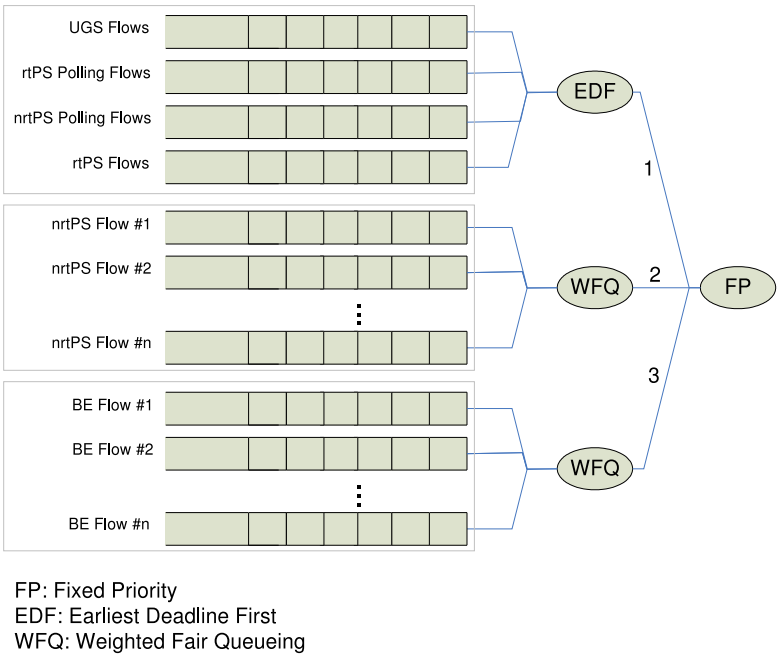

Figure 2.3. *3 schedulers [LIU 05]*

In [PER 06], a multimedia supported uplink scheduler is proposed. It includes a proportional fair (PF) BS scheduler and an earliest due date (EDD) SS scheduler. The BS scheduler (Figure 2.4) allocates resources first for the UGS service and then to poll SSs having at least one non-UGS connection: one slot is allocated in each frame for each SS having rtPS or nrtPS connections and one slot every three frames is allocated for SSs having only BE service connections. Finally, remaining OFDMA resources are proportionally allocated for SSs based on the received bandwidth requests. As can be seen from Figure 2.5, the EDD SS scheduler serves packets from the four traffic queues (UGS, rtPS, nrtPS and BE) in the order of the deadline assigned to each packet regardless of their scheduling service type.

Fragmentation, packing and PHS capabilities were considered in the packet-based scheduling strategy proposed in [SET 06]. As can be seen from Figure 2.6, the proposed scheduler combines a strict priority policy among the different service categories and a specific queuing management discipline for each class: fixed bandwidth, WRR and RR for UGS, (n)rtPS and BE, respectively. For WRR discipline, weights are determined according to the guaranteed bandwidth. Adaptive modulation and coding was also addressed in this work. However, a preliminary WRR/RR allocation was

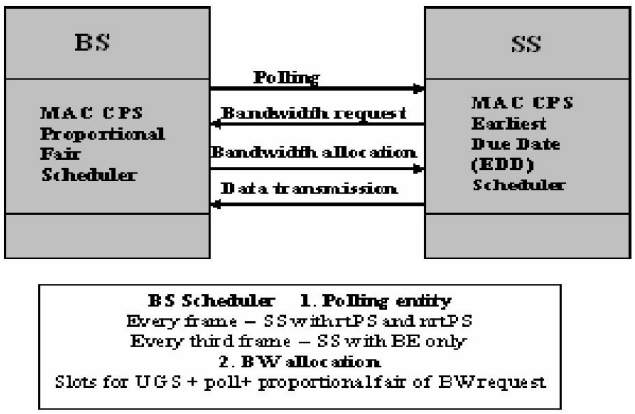

Figure 2.4. *Multimedia supported uplink scheduler[PER 06]: BS Scheduler*

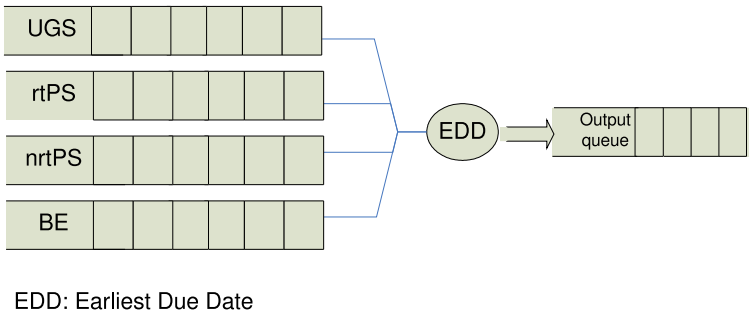

Figure 2.5. *Multimedia supported uplink scheduler[PER 06]: EDD SS Scheduler*

achieved assuming the use of the most robust burst profile while bandwidth was allocated taking into account the actual burst profile!! The admission control algorithm that manages the access of new connection—and based on which the minimum bandwidth requirements are guaranteed—was not described in this work.

To the best of our knowledge, [CHE 05] is the only research work that has proposed a scheduling algorithm considering simultaneously uplink and downlink bandwidth allocation in TDD mode. In first layer scheduling—of the two-layer hierarchical scheduling structure proposed in this work—Chen *et al* [CHE 05] have suggested the use of Deficit Fair Priority Queuing (DFPQ) algorithm instead of Strict Priority in order to avoid starvation for low priority classes. This first layer scheduling is based on two policies. The first one is a transmission direction-based priority where they chose to attribute to DL a higher priority than UL. The second policy is a service class-based

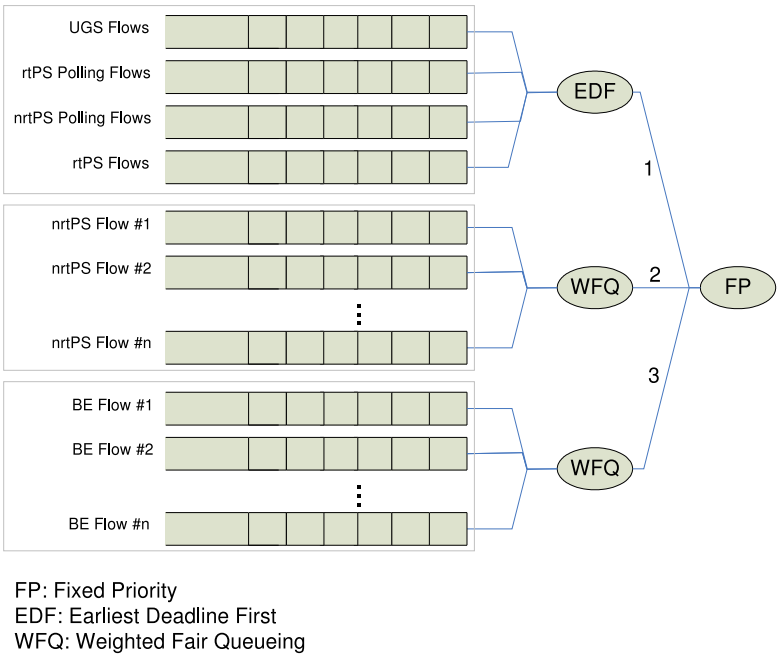

Figure 2.6. *Scheduler model [SET 06]*

priority applying the following scheme: rtPS>nrtPS>BE. As can be seen from Figure 2.7, the authors have combined these two policies using a strict priority scheme which assigns strict priority from highest to lowest to: DL_{rtPS}, UL_{rtPS}, DL_{nrtPS}, UL_{nrtPS}, DL_{BE}, andUL_{BE}. For DL and UL UGS connections, they have chosen to apply a fixed bandwidth allocation strategy. In second layer scheduling, three different algorithms were assigned to the other classes of services: EDF for rtPS, WFQ for nrtPS and RR for BE. nrtPS connections are scheduled based on weights corresponding to the ratio between the nrtPS connection's minimum reserved traffic rate and the sum of the minimum reserved traffic rates of all nrtPS connections. A basic admission control algorithm is also proposed in this work. It accepts the connections for which the minimum reserved traffic rate does not exceed the available channel capacity; all BE connections are nevertheless accepted.

Table 2.2 summarizes the hierarchical scheduling proposals described above. In this table, we precise either DL connections are concerned or not by the proposed scheduling mechanism. We also specify the different steps of the proposed solution: which scheduling services are considered in each level and which queuing disciplines are applied.

Scheduling proposal	Layer/Phase		DL	UL	UGS	rtPS	nrtPS	BE
[WON 03b, WON 03a]	1^{st} layer			•	Fixed Priority			
	2^{nd} layer				Fixed Bandwidth	EDF	WFQ	Equally distributed
[SUN 06]	BS Scheduler	1^{st}phase		•	Fixed Bandwidth	Grant Bandwidth Request Opportunities		
		2^{nd}phase				Guarantee the Minimum Reserved Rate		
		3^{rd}phase				WFQ to distribute residual bandwidth		
	SS Scheduler			•	Fixed Priority			
					Fixed bandwidth	EDF	EDF (Virtual Time)	FIFO
[PER 06]	BS Scheduler	1^{st}phase		•	Fixed Bandwidth	Unicast Polling		
		2^{nd}phase				Proportional Fair based on bandwidth Requests		
	SS Scheduler				EDD			
[SET 06]	1^{st} layer		•		Fixed Priority			
	2^{nd} layer		•		Fixed Bandwidth	WRR		RR
[CHE 05]	1^{st} layer		•		DFPQ			
	2^{nd} layer		•		Fixed Bandwidth	EDF	WFQ	RR
[LIU 05]	Scheduler 1				EDF (UGS + rtPS + Polling rtPS and nrtPS)			
	Scheduler 2						WFQ (weights are proportional to the bandwidth) request	
	Scheduler 3							WFQ (weights correspond to Traffic priority)

Table 2.2. *Hierarchical scheduling structures*

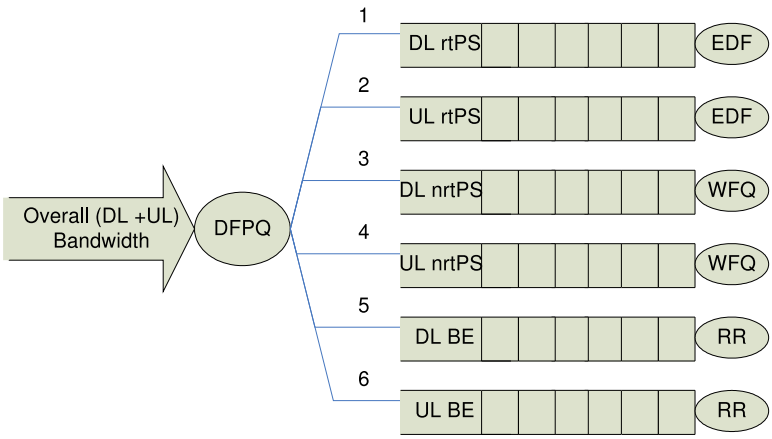

Figure 2.7. *Hierarchical structure of bandwidth allocation [CHE 05]*

In [CHA 06a], an original two-tier scheduling algorithm (2TSA) was proposed to avoid starvation problem and to provide fair allocation of residual bandwidth. UGS connection is not concerned by the "2TSA" algorithm since it is allocated a fixed amount of bandwidth per frame. Each connection is classified into either "unsatisfied", "satisfied", or "over-satisfied" category and is assigned a weight indicating its shortage or satisfaction degree—depending on its category. The connection is considered as:

– "unsatisfied" if the allocated bandwidth is less than its minimum requirement,

– a "satisfied" connection if the allocated bandwidth is between its minimum and maximum specified requirements,

– "over-satisfied" if it is granted more bandwidth than its maximum need,

The first-tier allocation algorithm is category-based and gives the highest priority to "unsatisfied" connections. For a specific category, the second-tier allocation algorithm is applied to share residual bandwidth based on weights. The flowchart of the proposed 2TSA is shown in Figure 2.8.

Compared to simple-structured scheduling solutions, the hierarchical scheduling mechanisms presented in this section combine in general an inter-service scheduling discipline with a specific queuing mechanism for each service class. Such structures

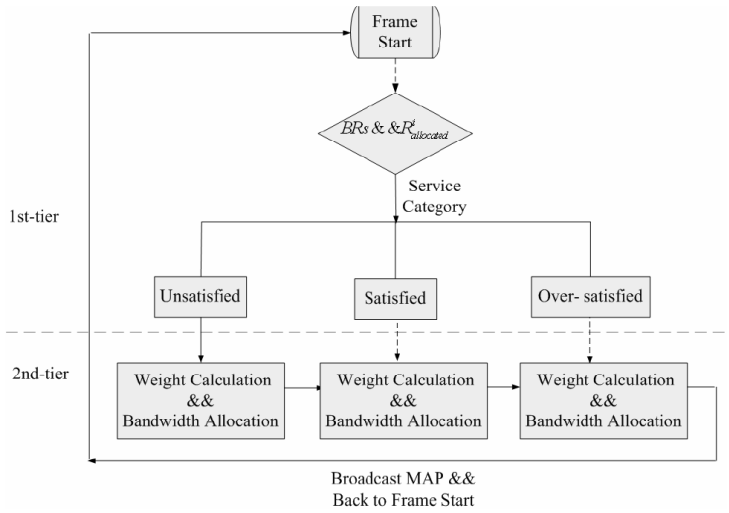

Figure 2.8. *Operation flowchart of 2TSA [CHA 06a]*

lead to a high computational complexity that may be prohibitive from an implementation point of view and that may not fit the delay constraints of real-time scheduling services.

Regardless of the proposed scheduling structure, some service-specific scheduling solutions are presented in literature. Lee *et al* for example focused in [LEE 05b] on VoIP services. They argued that both UGS and rtPS have some problems to support the VoIP services and proposed an enhanced scheduling algorithm to solve the mentioned problems. In fact, the fixed-size grants, assigned to UGS connections of voice users, cause a waste of uplink resources during silence periods. Moreover, the bandwidth request mechanism used by rtPS connections leads to MAC overhead and access delay which is not convenient for VoIP applications. Therefore the authors assumed that a voice activity detector (VAD) or silence detector (SD) is used by the SS in the higher layer and proposed an algorithm to be used by the SSs to inform the BS of their voice state transitions. In order to avoid MAC overhead, the proposed algorithm makes use of one of the reserved bits of the conventional generic MAC header of IEEE 802.16 [802 04] to do that. Simulations results showed that, compared to rtPS, the proposed algorithm decreases the MAC overhead and access delay; Also it can admit more voice users than UGS making more efficient use of uplink resources.

In a more recent work [LEE 06a], they demonstrated, using the analysis of resource utilization efficiency, that the ertPS service introduced by the IEEE 802.16e standard [802 05] is more suitable than UGS and rtPS for VoIP services with variable data rate and silence suppression. Indeed they proved that ertPS not only solves the

problems of resource wasting, delay, and overhead caused by the use of UGS and rtPS, respectively but also increases the number of number of voice users that can be supported by the network.

2.1.3.2. *Optimization-based strategy*

This second category of scheduling strategies consist in formulating the scheduling problem, in 802.16 environment, as an optimization problem aiming at optimizing the allocation of resources to different SSs. Table 2.3 presents the formulation of some examples of optimization problems proposed in literature.

To get an optimal solution to the optimization problem formulated in [SIN 06] (see Table 2.3), the authors need to use an NP-complete Integer Programming because the number of slots allocated per SS on a given channel should have an integer value. Relaxing this constraint, the authors proposed a second solution based on a linear programming approach that exhibits a complexity of $O(n^3.m^3.N)$ where n, m, and N denote the number of SSs, the number of subchannels and the total number of slots, respectively. However, because it is still a computationally demanding problem, the authors suggested the use of a heuristic algorithm whose computational complexity is $O(n.m.N)$. The authors then proved that the proposed algorithms optimize the overall system performance but may not be fair to different SSs. Therefore they modified them using the proportional-fair concept.
Based on the developed algorithms, they defined a scheduling algorithm for the BS and another one for the SS. The authors agree that considering a joint scheduling for uplink and downlink, at the BS, is more efficient. They nevertheless argue that it is not possible to do that when considering the context of OFDMA/TDD. Therefore they adopted a scheduling mechanism in which downlink and uplink are scheduled separately for all the classes. The priorities are assigned ass follows. Allocations are made first for UGS, then rtPS, then for nrtPS just to guarantee the minimum requirements, and finally to satisfy the remaining demands. The choice of one of the proposed algorithms depends on the availability of resources and on the channel conditions.
As for the SS, the authors took into account the overall system performance and fairness to different users. They proposed the same sequence followed by the BS but with two different models: a packet model, in which fragmentation is prohibited, for both UGS and rtPS and a byte model—fragmentation is possible—that may be used by nrtPS and BE services.

In [NIY 06e], Niyato and Hossain considered systems operating in a TDMA/TDD access mode and using WirelessMAN-SC air interface. They defined a utility function that depends on the amount of allocated bandwidth, the average delay, the throughput, and the admission control decision for UGS, rtPS, nrtPS, and BE, respectively. Using these utility functions, they formulated the optimization problem illustrated in Table 2.3. The authors set a limit of the allocated bandwidth between b_{min} and b_{max} for each connection. They also defined a threshold for each service class since the total

available bandwidth is shared using a threshold-based complete partitioning approach. To obtain the optimal threshold setting, an optimization-based scheme is proposed. To solve the proposed optimization problem, Niyato and Hossain suggested two solutions using an optimal approach and an iterative approach, respectively. The first solution has a complexity of $O(2^{M(\triangle b)})$ where M denotes the number of ongoing and incoming connections and $\triangle b = b_{max} - b_{min} + 1$. Since the complexity of the optimal algorithm may be prohibitive from an implementation point of view, the authors proposed an iterative approach based the water-filling mechanism. This solution is more implementation-friendly—its complexity is $O(C)$—while providing similar system performances.
To analyze the connection-level (such as the blocking probability) and packet-level (e.g. transmission rate) performance measures, the authors developed a queuing and a queuing analytical model, respectively. The proposed connection-level model [NIY 06e, NIY 06c] defines the connection blocking probability and the number of ongoing connections via a Continuous Time Markov Chain (CTMC) model. These parameters are then used to formulate an optimization problem (see Table 2.3) aiming at maximizing the system revenue while maintaining the blocking probability at the target level.

2.1.4. *Connection Admission Control Proposals*

Connection Admission Control (CAC) strategy is essential to provide Quality of Service (QoS) in mobile networks. Before a decision, CAC should confirm that the new call does not degrade the QoS of current connections and the system can provide the QoS requirements for the new call.

In the special case of WiMAX, four classes of services have been defined: Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS), Non-Real-Time Polling Service (nrtPS) and Best Effort (BE) [NIY 06e]. CAC is thus crucial for Supporting Quality of Service (QoS) guarantees for these services. In the following, we present first some CAC algorithms, and then discuss the different analytical methods that has been proposed to evaluate them.

2.1.4.1. *CAC proposals*

In [WAN 07], an uplink CAC algorithm has been proposed: A connection is admitted if: (1) there is enough bandwidth to accommodate the new connection, (2) the newly admitted connection will receive QoS guarantees in terms of both bandwidth and delay and (3) QoS of existing connections is maintained. The proposed CAC scheme is based on a token bucket: It reserves adequate bandwidth for every admitted flow. However, it may be considered as conservative in case of video transmissions leading to much of bandwidth being reserved unnecessarily.

In [YAN 06], the authors propose an admission control algorithm for real-time video applications, that takes into account the periodicity of arrival of the frames, in

Proposed Solution	**Cost Function (Minimize/Maximize)**	**Constraints (subject to)**
Joint Bandwidth Allocation and admission control [NIY 06c]	Minimize The average delay	* The average delay should meets the delay requirements of rtPS connections. * The transmission rate meet the transmission rate requirements of connections. * The amount of allocated bandwidth for each connection is between b_{min} and b_{max}. * The total amount of allocated bandwidth does not exceed the total available bandwidth.
Queuing theoretic and optimization-based model for resource management [NIY 06e]	Maximize level of users' satisfaction <=> Maximize Utility function	* The allocated bandwidth for UGS connections is equal to the required bandwidth * The delay requirements for rtPS connections (depending on the arrival rate, the average SNR and the allocated bandwidth) are met. * The transmission rate requirements of nrtPS connections (depending on the arrival rate, the average SNR and the allocated bandwidth) are met. * BE connections are admitted. * The amount of allocated bandwidth for a given connection is between b_{min} and b_{max}. * The total amount of allocated bandwidth does not exceed the total available bandwidth. * The thresholds (corresponding to the amount of reserved bandwidth for each service class) are respected.
Queuing model for connection-level performance analysis [NIY 06e]	Maximize The system revenue <=> Maximize the number of ongoing connections	* The connection blocking probabilities [3]for UGS, rtPS, nrtPS and BE connections do not exceed the target blocking probabilities.
Efficient and fair Scheduling of Uplink and Downlink in OFDMA Networks [SIN 06]	Minimize the unsatisfied demands	* The number granted slots on a given subchannel do not exceed the number of slots of this subchannel * The amount of bandwidth (slots) allocated per connection do not exceed the whole demand of that connection.

Table 2.3. *Optimization approach: Cost function and constraints*

order to overcome the excessive delay caused by simultaneous arrivals of connection demands. This scheme sets up a pending period for each new arrival. The flow is not admitted until the CAC finds the earliest proper time within its pending period to establish a connection. If there is no such an appropriate access time, the CAC ultimately rejects the incoming flow after the pending period expires. It has been shown that the profile of aggregated traffic is relatively smoother, leading to lower

delay violation. However, the performance of the algorithm has not been assessed in the presence of other types of traffic.

The paper [WAN 07] focuses on the non-preprovisioned service flow, for which the MS initiates the connection creation. The BS has to decide whether to admit or reject each new connection, and how much bandwidth should be reserved for the admitted connection during its dwell time in the cell. The proposed algorithm is based on a guard channel scheme as this introduced in [RAM 97] and its performance is assessed using simulations. A guard channel CAC is also proposed in [LEE 06b].

In [CHA 06b], the authors define a so-called QoS-CAC, where the new connection request is classified into a particular queue depending on the associated Service Class type. QoS-CAC serves the UGS connection queue first, followed by RTPS and then by NRTPS queues. Thus, it provides highest priority to UGS connections requests followed by RTPS and NRTPS connection requests.

In [JIA 06], a token-bucket CAC scheme is proposed. Each connection is controlled by two token bucket parameters: token rate and bucket size. When a traffic flow wants to establish a connection with BS, it sends these two parameters to BS and waits for response from BS. An extra parameter, delay requirement, is sent by rtPS flow. A threshold on capacity is fixed for each class and, at each arrival, the remainder uplink capacity is calculated and compared to the bandwidth requirement of the new connection. If there is enough capacity it is accepted. If not, CAC looks at the connections that belong to lower classes than this new connection. If there is a class that uses more capacity than its threshold, it calculates how much capacity can be stolen from it to satisfy the new arrival. A connection can steal capacity from connections of a higher class only if the class it belongs to occupied less capacity than its threshold and the higher class use more capacity than its threshold.

The paper [RON 07] studies the CAC from two different points of view. From the perspective of service provider, the admission control policy that produces optimal revenue is desired. Service provider charges different revenue rates (revenue per bandwidth unit and time unit) from different service types. The admission control policy is thus likely to give preference to the traffic load of high revenue rate. As of the perspective of WiMAX subscribers, the admission control policy of optimal utility is expected, since it can produce the maximum access bandwidth. This policy will allocate more bandwidth resources to the traffic load that can yield high utility. As a compromise has to be made between the service provider and the WiMAX subscriber, the authors define the concept of utility-constrained optimal revenue policy.

In [NIY 05a], two CAC strategies are compared for fixed OFDM networks. The first scheme is threshold-based, in which the concept of guard channel is used to limit the number of admitted connections to a certain threshold. The second scheme is based on the information on queue status and also it inherits the concept of fractional

guard channel in which an arriving connection is admitted with certain connection acceptance probability. It is shown that the queue-aware CAC scheme offers more adaptability to the traffic load.

Paper [GE 06] presents an adaptive admission control scheme for adaptive multimedia services in IEEE 802.16e, and compare it also to threshold-based CAC. First of all, UGS, RT-VR/ERT-VR, NRT-VR and BE are prioritized from the highest to the lowest. Furthermore, different traffics belonging to the same data delivery service may also have different levels, following jitter and delay requirements. A degrading/upgrading policy is then defined following these priorities, and blocking occurs when no more degrading is possible. A similar algorithm has been proposed in [CHO 05a].

The paper [NIY 06a] presents a fuzzy logic-based CAC algorithm for OFDMA WiMAX. The proposed admission control algorithm considers various traffic source parameters (i.e., normal rate, peak rate and probability of peak rate) and packet-level delay requirements for the traffic to decide whether an incoming connection can be accepted or not. The inference rules for resource allocation in the proposed fuzzy logic admission control are defined based on the following scheme: When a new connection is initiated, the corresponding mobile node informs the base station with approximate traffic source parameters (i.e., normal rate, peak rate and probability of peak rate) and target delay requirement. These inputs are fuzzified into fuzzy sets and the traffic source estimator estimates traffic intensity as the output. Next, the base station measures and fuzzifies average SNR of the new connection. This traffic intensity and channel quality information are used by the resource allocation processor together with the user-specified delay requirement to obtain the number of subchannels to be assigned.

A more complex CAC scheme is proposed in [NIY 06b] for fixed WiMAX (Single Carrier WirelessMAN), where adaptive bandwidth allocation (BA) and connection admission control mechanisms are developed for polling services based on game theory. A non-cooperative two-person general-sum game is formulated where the base station and a new connection are the players of this game. The objective of the proposed game-theoretic model is to find the equilibrium point between the base station and a new connection. The conflict in this game arises due to the fact that constrained by limited radio resources (i.e., bandwidth), the base station wants to maximize its utility (e.g., revenue) from the ongoing connections by providing higher level of QoS to these connections, while a new connection wants to achieve the highest possible QoS performance as well. Among the available strategies of both base station and new connection, the Nash equilibrium is determined by using the best response function and the decision on admission control is made based on admissible strategy pair from the Nash equilibrium.

In [NIY 06e], two CAC approaches are analyzed, namely, the optimal and the iterative approaches. For the optimal approach, an assignment problem is formulated and solved by using binary integer linear programming. However, this optimal approach is shown to incur a huge computational complexity, and therefore, may not be suitable for online execution. On the other hand, the iterative approach, which is based on the water-filling method, is shown to be an implementation-friendly one, with comparable performance.

Again, the CAC in [WAN 06] is limited to the case of fixed WiMAX, where on BS serves several subscriber stations, each having several users connected to it. The problem of CAC is thus a problem of finding, for each new arrival, if the aggregated bandwidth is less than a limit, insuring a given blocking probability.

In [ELA 06c], an adaptive admission control scheme is presented, and its performance is assessed in the presence of different adaptive modulation and coding schemes, namely based on the received power or based on the interference. A cross-layer approach is followed, considering the impacts of the physical layer conditions (modulation and path loss), the MAC layer techniques (radio resource management algorithms) and the traffic characteristics.

2.1.4.2. *Performance evaluation of CAC algorithms*

Several CAC proposals have been evaluated using simulations [WAN 07, YAN 06, JIA 06, GE 06, WON 03b]. Other papers propose analytical evaluation using Markov analysis.

In [GE 06], a Markov chain is constructed to describe the evolution of the state of the system with adaptive rate control. Two classes of calls are differentiated, and the blocking probability, in addition to the dropping rate of handoff calls are calculated. Simulation results show good matching between analytical and simulation results.

In [WAN 06], realistic assumptions have been made on the traffic models (Poisson arrivals and exponential durations for voice, Poisson Pareto burst process for rtPS and nrtPS services, and heavy-tailed traffic with Pareto distribution for best effort. The Gaussian approximation is used to derive chernoff bounds for the blocking probability. However, the model were limited to fixed WiMAX, in a Wireless MAN setting.

In [NIY 05a, NIY 06b], the performance of the WirelessMAN system is also assessed using a queuing analysis. However, while in [NIY 05a], a Poisson arrival of packets is considered, the burstiness in traffic arrival is modeled in [NIY 06a, NIY 06b] and [NIY 06e] using Markov modulated Poisson process. The performance of the system is analyzed at the packet level: Each connection has its own queue were PDUs are queued, and served with a throughput that depends on the state of the channel. The admission control strategy reserves separate bandwidth for each type of connections. The PDU dropping probability, the queue throughput and the average

delay are calculated for different admission control strategies (static, adaptive, and game theory based).

A classical Markov model is used in [LEE 06b] to evaluate the performance of guard-bandwidth CAC, with the difference that a two-states AMC was considered. However, a simplistic model with no interference is considered, even if a multi-cell setting is considered.

In [ELA 06c], a more realistic model is considered for AMC, taking into account the impact of inter-cell interference. The state of a call is then modeled as hyper-exponential with several states corresponding to the different modulations, and some performance measures are calculated.

2.2. Capacity at the MAC layer

The MAC layer of IEEE 802.16 supports a primarily Point-to-Point (PMP) architecture. The communication path between a subscriber stations (SSs) and Base station (BS) has two direction : uplink (from the mobile to BS) and downlink (from the BS to the mobile). The dowlink is generally broadcast, but the uplink is shared by the SSs. IEEE 802.16 has defined the MAC layer as connection-oriented, is designed to support different QoS for different services. In the following subsection, we specify some basic characteristics of the common IEEE 802.16 MAC protocol to create a framework for designing the QoS architecture.

2.2.1. *QoS architecture for IEEE 802.16 MAC protocol*

In order to support the QoS for different services by scheduling the uplink access opportunity, four QoS service are defined in the standard : Unsolicited Grant Services (UGS); Real-Time Polling Service (rtPS), Non-Real-Time Polling Service (nrtPS) and Best effort (BE). UGS is designed to support real-time flows. The (BS) must provide fixed size data grants at periodic intervals to the UGS flow. UGS can be used for constant bit-rate (CBR) for CBR-like service flow as Voice over IP and and T1/E1. The rtPS is designed to support real-time service flows that generate variable size data packets on periodic basis. The rtPS can be used by rt-VBR-like service flows such as MPEG video. The SS is allowed to use only the unicast request issued by BS for connection, moreover, rtPS flows prohibed from using any contention requests. The nrtPS is designed to support delay-tolerant flows and require a minimum data rate such as FTP. However, the nrtPS flow receive few request polling opportunities during network congestion and frequency enough to meet the delay requirement. The BE service is designed to support a flow for which no minimum transmission rate such as HTTP. The SS is allowed to use contention request opportunities as well as unicast request opportunities for BE service flow. The BE flows receive few request polling opportunities comparing to nrtPS flow.

2.2.2. *Contention mode : Binary Exponential Backoff*

On the donwlink, the transmission is relatively simple because the BS is the only one that transmits during the donwlink subframe. The data packet are broadcasted to all SSs and an SS only picks up the packet destined to it.

We focus only on the uplink subframe. Itself partitioning in four TDMA subframes. The first three are reserved for the CAC : initial and maintenance connection. The last one carry the data transmission through numerous time slots. The whole capacity is greatly improved by using, for all these subframe, several OFDMA frequency. Moreover, the ranging intervals can manage a large number of contending connection because each of the three time slots use CDMA technique. This allows to share the channel resources through all contending nodes as well as minimize the collision probability. The figure 1.2, shows all these specificities.

On the uplink, the BS determines the number of slots that each SS will be allowed to transmit. This information is broadcasted by the BS through the uplink map message (UL-MAP) at the beginning of each frame. After receiving the UL-MAP which containing information element IE, each SS will transmit data in the predefined time slots which indicated in IE. The information element is obtained by using the bandwidth request sent from SS to BS.

An SS which has a packet to send is called active. The bandwidth request procedure depends on the node state: if the node is silent, it uses the contention time slot in the Bandwidth Request Ranging Interval. Else, when the node is still transmitting, the request is achieved by using an aggregate or incremental bandwidth request in its data reserved time slots. The incremental one is required when a node needs more resources. The other one allows to reevaluate, often periodically, the node needs.

As soon as a node want to send data, it chooses one of the N codes composing the dedicated bandwidth request code family, and proceed to its demand by transmitting its coded request through the bandwidth request ranging interval. These requests follow a backoff process in case of collision in the selected code. A collision occurs if two or more nodes have chosen the same code in the same ranging interval.

Before entering its contention resolution process, an active SS first gets the initial backoff W and the maximum backoff W_{max} from BS. The SS randomly selects a backoff value within the initial backoff. Backoff value decreases by one on every transmission and when this value reach zero, the SS sends its bandwidth request. After transmission, the SS waits the message (UL-MAP) which contains the information element. For this, the SS waits its bandwidth response until a timeout threshold. The IEEE802.16e standard version defines the timer T_3 as the maximum MAC frame number that a contending node can wait before consider that its request has been lost on the wireless channel, or in the BS request queue.

2.2.3. *Literature on MAC*

There has been few research activity on modeling of IEEE 802.16 medium access standards. In the literature, the performance evaluation of 802.16 has been carried by means of simulation. No much has been done for analytical model. The capacity of the OFDMA-CDMA ranging subsystem in 802.16 has been studied in a few papers. In [RYU 03], the authors analyzed the performance about random access protocol which use ranging subchannel in OFDMA-CDMA environment with respect to mean delay time (MDT) and first exit time (FET). In [HWA 04], authors designed and analyzed the performance analysis model to control adaptively the size of each ranging code for IR, PR, and BR ranging in order to do efficiently random access. In [KIM 05b], they evaluate the capacity of a ranging subchannel in terms of the ranging code error probability versus the number of active users to attempt ranging. Recently, several works addressing QoS in general and call admission control (CAC) in particular have been produced. For instance, in Reference [LI 05], an admission control scheme is proposed. It ensures highest priority to UGS flows while maximizing overall bandwidth by means of bandwidth borrowing. Recently, there are amount of research works published on the QoS service provisioning in the WiMAX networks. Most of the proposals focus on the enhancement of the QoS service architecture [CHU 02, CHO 05b, MA 06]. In [LEE 05a], an enhacement have been suggested to support Voice ove IP traffic in both UGS service and the rtPS with aim to increase the utilization of the uplink bandwidth and reduce overhead. In Reference [WAN 05b], QoS is treated based on classical intserv and diffserv paradigms as well as their mapping to IEEE 802.16 MAC layer.

IEEE 802.16 has defined the MAC protocol stack for BS to assign the uplink channel to SSs. But during initial maintenance and bandwidth contention periods, all SSs still need to contend the uplink channel. As mentioned before, the contention resolution that be supported by 802.16 is based on a truncated Binary Exponential Backoff algorithm. This algorithm has been wildly investigate in IEEE 802.11 networks [MIO 05, BIA 00, XIA 04, XIA 03, LI 03]. The first contribution in this part, focus on the BEB in 802.16. The purpose of this part is to analyze IEEE 802.16e medium access control (MAC) sublayer and provides a simple analytical model to compute the 802.16e MAC throughput. Our approach is to begin with a key approximation made by [MIO 05] in 802.11. This lead a fixed point equation, which can be expected to characterize the operating points of system. This fixed point equation allow us to compute the probability rate of a mobile in saturated case, the throughput formulas for the overall network and the throughput of a ranging code. Moreover, our performance model deals with a recently released IEEE802.16 criterium, the T_3 timer (tr), which have a main impact on the MAC performance.

2.2.4. *Problem formulation*

Here, we concentrate on the saturated case, i.e, each mobile has a packet to send. We consider a single IEEE802.16e cell in which there are n mobiles. We assume that the retransmission processes are engaged after the T_3 timeout (t_r parameter). As described in the previous section, each mobile will engage a CDMA+OFDMA request through the bandwidth request ranging interval (N is the number of code dedicated for the bandwidth request). Then, the nodes listen to the following downlink frames until the T_3 timeout expiration. If the node receives a ranging response, the transmission starts. Else, the node enters into the backoff mode : before proceeding to a new transmission attempt, the node waits several frames. Their number is randomly chosen in the window $[0, b_k]$. Where b_k is defined in equation (2.2). Note the backoff is also called truncated exponential backoff because from a determined number of retransmission (m parameter) the backoff windows is no more increased.

In the IEEE802.16e standard, the channel occupation from the trunked binary exponential process and the transmission time are completely independents. This is due to the fact that the ranging requests are achieved on a different channel from the data transmissions. Also, we can simplify the backoff time representation as the concatenation of the different backoff stages for a specific node. Based on these properties, we can develop an accurate throughput model with a Fixed Point Analysis. The FPA uses these properties in order to model the collision process. It reveals a recursive function, based on the attempt rate per slot and per node. This relation converges to the collision rate. Moreover, an other FPA development leads to the effective throughput model.

The following figure describes the evolution of the back-off of a node. R_j is the number of attempts until success for the jth packet.

The following figure (2.9) shows the back-off chronogram for the transmission of three ranging requests. The first one undergoes one collision, two collisions occur for the second one, and three for the last one. Note that we represent here the t_r time between the sending and the response instant.

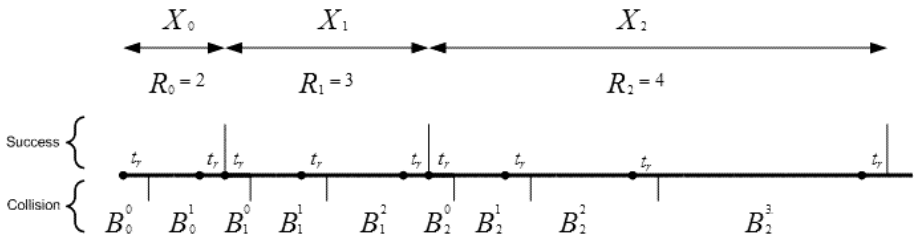

Figure 2.9. *IEEE802.16e backoff process chronogram*

Hence, the total slot number required to transmit the jth packet is given by

$$X_j = R_j.t_r + \sum_{i=0}^{R_j-1} B_j^i$$

Let γ be the collision probability seen by a node and k the maximum retry number. Now, we can define $G(\gamma)$, as the average attempt rate per slot.

$$G(\gamma) = \frac{E(R)}{E(X)}$$

where

$$E(X) = E(R).t_r + E\Big(\sum_{i=0}^{R_j-1} B_j^i\Big)$$

Since the back-off behavior of all nodes is the same, the collision probability is the same for all nodes. Hence, from the previous assumptions, it is easy to obtain

$$E(R) = 1 + \gamma + \gamma^2 + ... + \gamma^k$$

$$E(\sum_{i=0}^{R_j-1} B_j^i) = b_0 + \gamma b_1 + \gamma^2 b_2 + ... + \gamma^k b_k$$

However, after some calculations we have

$$\begin{aligned} G(\gamma) &= \frac{E(R)}{E(R).t_r + E\Big(\sum_{i=0}^{R_j-1} B_j^i\Big)} \\ &= \frac{1}{t_r + \frac{E(\sum_{i=0}^{R_j-1} B_j^i)}{E(R)}} \end{aligned}$$

Thus,

$$\frac{E\Big(\sum_{i=0}^{R_j-1} B_j^i\Big)}{E(R)} = \frac{b_0 + \gamma b_1 + \gamma^2 b_2 + ... + \gamma^k b_k}{1 + \gamma + \gamma^2 + ... + \gamma^k} \tag{2.1}$$

Note that we have k maximum retries for the backoff process. We define the following back-off parameters for a node as ($m < k$)

$$b_k = \frac{2^k b_0 - 1}{2} \text{ for } 0 \leq k \leq m-1$$
$$\text{and} \tag{2.2}$$
$$b_k = \frac{2^m b_0 - 1}{2} \text{ for } k \geq m$$

By substituting these into the expression, $G(\gamma)$ yields

$$\frac{E\Big(\sum_{i=0}^{R_j-1} B_j^i\Big)}{E(R)} = \frac{(1-2\gamma)(b_0-1) + \gamma b_0(1-(2\gamma)^m)}{2(1-2\gamma)} \tag{2.3}$$

Finally we compute the slot attempt rate as :

$$G(\gamma) = \frac{1}{t_r + \frac{(1-2\gamma)(b_0-1)+\gamma b_0(1-(2\gamma)^m)}{2(1-2\gamma)}}$$

Now if all nodes have the same back-off process, they will all share the collision probability and the same attempt rate $G(\gamma)$. We assume that the number of attempts made by the other nodes is binomially distributed with parameters $G(\gamma)$, $n-1$ and N. In fact, the probability of collision of an attempt by a node i is given by

$$P_{coll}(G(\gamma)) = 1 - \Gamma(G(\gamma)) \tag{2.4}$$

where $\Gamma(G(\gamma))$ is the probability that the other nodes that attempt in the same slot, do not use the same channel used by node i. This probability is given by :

$$\Gamma(G(\gamma)) = \sum_{i=0}^{n-1} G(\gamma)^i (1-G(\gamma))^{n-i-1} (1-\frac{1}{N})^i$$

Now we expect that the equilibrium behavior of the system will be characterized by the solution of this following fixed point equation.

$$\gamma = P_{coll}(G(\gamma))$$

For the existence and the uniqueness fixed point, the results in [MIO 05] can be easily extended in our case.

2.2.5. *Performance Analysis*

At each slot, the arrival ranging requests are buffered in a queue with infinite buffer size. Let H_1, H_2,... be the number of the ranging requests served during a time slot with the generating function $A(z) = \sum_{i=1} a_i z^i$ and finite mean batch size μ.

In the sequel, we determine the arrival batch sizes during a time slot. The attempt number engaged in a single ranging request interval leads to several aggregate requests entering into the request queue at the base station. A request incomes in the system only if its ranging code is used only by a single node at the same time. Let $P(Z_t = j|N)$ be the probability that the base station receives successfully j ranging requests over N codes at time slot t, where $j \in \{0, 1, .., N\}$. $P(X_t = i)$ is probability that i nodes simultaneously transmit their ranging request at time slot t, which is given by

$$P(X_t = i) = \binom{n}{i} G(\gamma)^i (1 - G(\gamma))^{n-i}$$

Thus,

$$P(Z_t = j|N) = \sum_{i=j}^{N} P(Z_t = j|X_t = i, N) P(X_t = i) \tag{2.5}$$

The conditional probability $P(Z_t = j|X_t = i, N)$ can be evaluated by a recursive expression as follows :
$P(Z_t = j|X_t = i, N) =$

$$\begin{cases} \sum_{k=0, k\neq 1}^{i} \binom{i}{k} \left(1 - \frac{1}{N}\right)^{i-k} \left(\frac{1}{N}\right)^k P(Z_t = j|X_t = i - k, N - 1) \\ + \binom{i}{1} \left(1 - \frac{1}{N}\right)^{i-1} \frac{1}{N} P(Z_t = j - 1|X_t = i - 1, N - 1) \end{cases} \tag{2.6}$$

The initial conditions for $P(Z_t = j|X_t = i, N)$ is given by :

$$P(Z_t = j|X_t = i, 0) = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{otherwise} \end{cases}$$

Hence, the average arrival rate λ is given by

$$\lambda = \sum_{k=1}^{N} k P(Z = k|N)$$

Note that the stability condition needs the average arrival rate to be less than the average service time slot, i.e.,

$$\lambda = \sum_{k=1}^{N} k P(Z = k|N) < \mu$$

From the previous analysis and assumptions, it is possible to model the number of packets (ranging) M_t with the discrete-time Markov chain in $\mathcal{N}$. The one step transition probability that the state of Markov chain M_t from $M_t = i$ at time t to M_{t+1} at time slot $t+1$ is given by $Q_{ij} =$

$$Q_{ij} = \begin{cases} P(Z = j) \text{ if } i = 0 \\ \sum_{k=0}^{i} a_k P(Z = j - i + k) \text{ otherwise} \end{cases}$$

Hence the stationary distribution π of this Markov chain is given by the solution of the following linear equations : $\pi = \pi.Q$, and using the conservation relationship

$$\sum_{i=0}^{\infty} \pi_i = 1$$

The average number of ranging request in the buffer is given by

$$S(\gamma) = \sum_{k=0}^{\infty} k\pi(k)$$

From Little's formula, the average time which each ranging spends in the queue is given by the ratio between the average number of ranging $G(\gamma)$ and the arrival rate. Therefore, we have for the average delay suffered by a ranging :

$$D(\gamma) = 1 + \frac{S(\gamma)}{\mu}$$

In our case, the delay is one slot larger, since a ranging is assumed to join the system only after the slot in which the message is generated.

2.2.6. *Numerical analysis*

In this section, we present several simulation results obtained with the Matlab software. We discuss the three main study topics : attempt rate per slot, average request incoming and the Fixed Point Equation results. For each of them, we provide figure sets. Note here that the bolded elements in the figure legends correspond to the default parameter used in the IEEE802.16e standard : the initial mean back-off b_0 is 16 slots. The backoff windows reaches its maximum values after 16 retries. So, m corresponds to 16. We use a ranging response reception timeout (timer T_3), t_r equivalent to 50 MAC frames : the mean MAC frame duration is 1ms, and the default value for T_3 timeout is 50ms. Moreover, knowing that a MAC frame duration varies between 0.5ms. and 2ms, the T_3 timer may represent a waiting time up to 100

slots. This singularity justifies the low range where the attempt rate function belongs, and we think that the first technology improvements will come from this parameter reduction. Moreover, we would like to discuss the number of ranging codes : the standard defines a large code spectrum composed by 256 orthogonal codes. These codes are split into four families. But, to the best of our knowledge, only one work [HWA 04] has proposed the respective size of each families. So, we feel that it is accurate to reserve at least the half of the code spectrum to the periodic ranging code family. Hence, we use here N equal to 128 as default value.

First we deal with the attempt rate performance in order to highlight how the communication parameters (ie. b_0, m and t_r) affect the attempt rate performance.

The figures 2.10, 2.11 and 2.12 show the attempt rates reached respectively in function b_0, m and t_r parameters. Concerning the b_0 parameter impact, using few slots for the initial backoff window permits to engage more attempts per slot. However, we will see later the drawbacks on the collision probability. Although, the figure 2.10 testifies to a better robustness against collisions for a fewer b_0 value.

Now we focus on the other main communication parameter : m. the figure 2.11 shows the impact of this parameter on the attempt rate. We remark that for a collision probability lower than 0.3, the backoff window expansion limit have no impact on the attempt rate per slot experienced by each node. However, from a collision probability equals to 0.45, a small expansion limit allows to keep a relatively high rate while the attempt rate performance would collapse with higher values.

So, we discuss later, through the FPA results if the standard could use efficiently a lower default value for the b_0 and m communication parameters.

The figure 2.12 shows the T_3 timer influence. We note here that if nodes have to wait between 50 and 100 slots to obtain a ranging response, the attempt rate undergoes a low performance variation. But, since the timer limit is lower than 30 slots, the attempt performances takes off. The figure proves clearly the strong impact of the t_r parameter. For a lack of place, we can not show the results obtained for different b_0 and m, but we have observed the same attempt rate behavior whether the t_r parameter is.

2.2.7. *Fixed Point Analysis*

The next topic of the numerical analysis deals with the Fixed Point Equation results. We have seen the impact of the main communication parameters on the connectivity performance. The FPA allows us to deepen our criticism. With the following figures we can appreciate how b_0, m, t_r and N modify the collision probability. Please keep in mind that the Fixed Point Equation solution corresponds to the intersection between the studying function with the $y = x$ function.

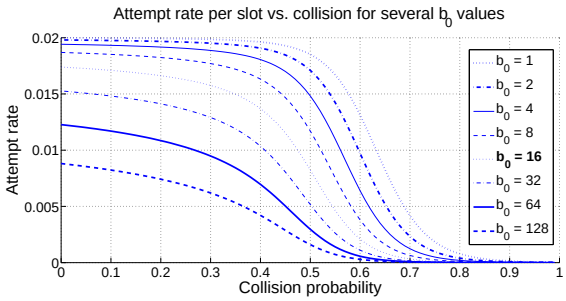

Figure 2.10. *Plots of $G(\gamma)$ vs. γ : Attempt rate values depending the collision probability for different values of b_0. IEEE802.16e default values : m=16, t_r=50.*

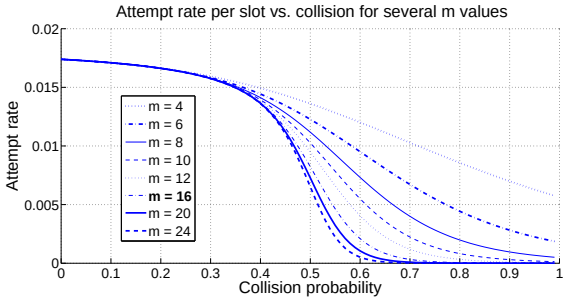

Figure 2.11. *Plots of $G(\gamma)$ vs. γ : Attempt rate values depending the collision probability for different values of m. IEEE802.16e default values : b_0=16, t_r=50.*

So, the figure 2.13 shows that by increasing the b_0 parameter, we increase the collision probability as well. This result was expected after the observation of the figure 2.10. Here, manufacturers have to find a tradeoff between the individual attempt rate, and the global collision probability.

Now, we see through the figure 2.14 a key result topic. Indeed, the figure shows that the collision probability is almost independent of the m parameter : from a $m = 4$ to a $m = 16$ range the collision probability increases only by 0.05 points. So, we think that it would be a great enhancement to reduce the m parameter in the IEEE802.16e standard. In addition, the figure 2.11 shows that this keeps a relatively high attempt rate while the collision probability slightly increases.

Concerning the t_r parameter, the figure 2.15, as the figure 2.13, we increase the collision probability by lowering the t_r parameter. Here, we have a 0.07 collision probability range for t_r including in 5 and 50 slots. We also observe that for a t_r lower than the standard value, we never exceed 0.1 point increasing, but we multiply in the same time the attempt rate. Finally we pinpoint on the fact that the performance

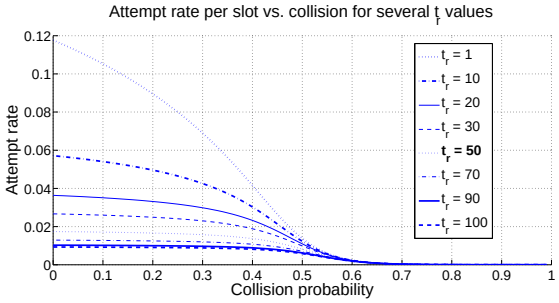

Figure 2.12. *Plots of $G(\gamma)$ vs. γ : Attempt rate values depending the collision probability for different values of t_r. IEEE802.16e default values : b_0=16, m=16.*

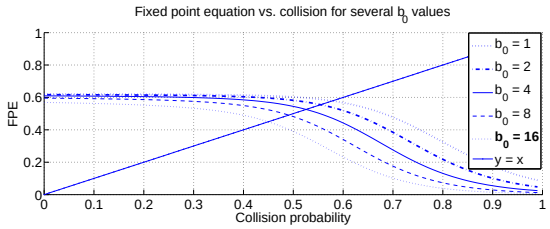

Figure 2.13. *Plots of $1-\Gamma(\beta)$ vs. γ : Fixed Point Equation, function of the collision probability for different values of b_0. IEEE802.16e default values : m=16, t_r=50 and N=128, n=50*

enhancement achieved by the t_r decreasing is greatly linked to the CDMA capacity (ie. figure 2.12).

For the last figure (2.16), it confirms clearly the top rank impact of n on the connectivity performance. The collision probability is rapidly increased with the number of users. We think that the CDMA partitioning, as well as the code range enlargement, will be the main topic leading to an actual performance enhancement.

2.2.8. *Request queuing*

Here, we deal with the average number of request that income to the base station. The figure 2.17 represents the arrival rate performance in function of the collision probability. As we expect, using numerous codes allows to reach a higher arrival rate. But, the figure also shows that this does not affect the robustness against the collision.

The figure 2.18 shows the income enhancement obtained by using a large number of ranging code. Obviously, the average arrival rate increases with the number of users, but this increase can be linear only if the CDMA capacity fits the attempt rate

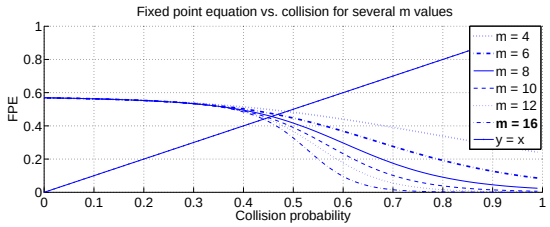

Figure 2.14. *Plots of* $1 - \Gamma(\beta)$ *vs.* γ *: Fixed Point Equation, function of the collision probability for different values of m. IEEE802.16e default values :* b_0*=16,* t_r*=50 and N=128, n=50*

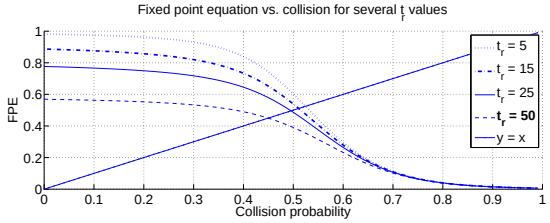

Figure 2.15. *Plots of* $1 - \Gamma(\beta)$ *vs.* γ *: Fixed Point Equation, function of the collision probability for different values of* t_r*. IEEE802.16e default values :* b_0*=16, m=16, and N=128, n=50*

behavior. We observe on the figure, that the lowest values of N induce a slower arrival rate increasing. For instance, 50 users perform an average arrival rate equals to 0.5 through 2 ranging code, compare to a 0.8 arrival rate with 32 ranging code. In fact, it seems that the arrival rate converges, function of the ranging code number, and the convergence values rapidly increases with this parameter. So, it proves that the N parameter is one of the most important factors for the connectivity performance.

The parameter N corresponds to the main factor of robustness against the collision : with an insufficient ranging code diversity, these numerous attempt will easily undergo collisions. The figure 2.18 testifies that it must be evaluated in function of the number of user. We use here 50 users because of some simulator limitation. But the IEEE802.16e is designed to assume far more customer through a large ranging code variety.

We finish our analysis by presenting the average queuing delay encountered by the incoming ranging request. The figure 2.19 shows the results for an unbuffered queue, while the figure 2.20 presents the buffered case results.

First we remark that the average delay for both cases are far lower than the authorized delay defined in the standard. The T_3 timer allows the sender to wait up to 50 MAC frames before acknowledging a possible loss. From a pool of 50 users, we observe that the average delay does not exceed 3 MAC frames for the unbuffered case.

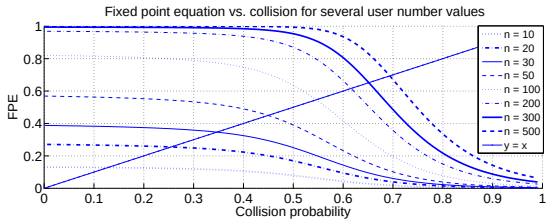

Figure 2.16. *Plots of $1-\Gamma(\beta)$ vs. γ : Fixed Point Equation function of the collision probability for different values of N. IEEE802.16e default values : b_0=16, m=16, t_r=50 and N=128*

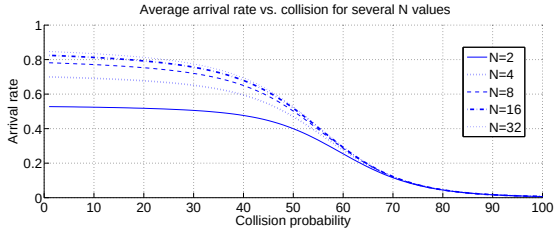

Figure 2.17. *Plots of λ vs. γ : Average request incoming, function of the collision probability for different values of N. IEEE802.16e default values : b_0=16, m=16, t_r=50 and n=50*

Obviously we note that this maximum values will increase with the number of users. But it also seems that it would be relevant to adapt the t_r parameter in accordance with the pool of customers. Moreover, the figure 2.19 proves clearly that this parameter, by inducing a lower arrival rate, also decreases the queuing delay. We could find a trade-off between these two impacts to design the best fitted value for the T_3 timer.

Concerning the buffered case in the figure 2.20, we observe that for a threshold value of the buffer size, ranging requests undergo a decreasing queuing delay. In fact, these results testify that some overflows occur, i.e. many incoming requests are dropped because of there are no remaining slot in the queue. So, we remark the key topic which has to be observed to design the base station buffer size : the t_r parameter, which has a major impact on the ranging requests incoming in the queue, can achieve better queuing performance in function of the collision probability. The figure 2.20 shows that a T_3 timer equal to 5 achieves better performance that the others for a collision probability equal to 0.5, whereas for this value, the performances fall below the 0.4 collision probability. We encourage to consider this remark in order to develop a more efficient designing algorithm which includes an adapted approach of the communication parameter depending on the collision parameter and the arrival rate factors (i.e. mainly, t_r and N).

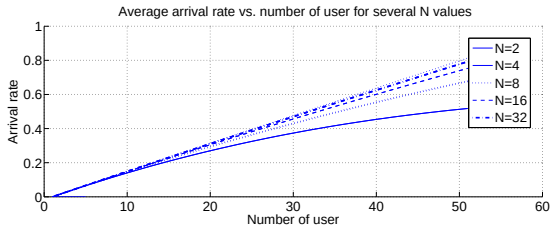

Figure 2.18. *Plots of λ vs. n : Average request incoming, function of the number of user for different values of N. IEEE802.16e default values : b_0=16, m=16, and t_r=50*

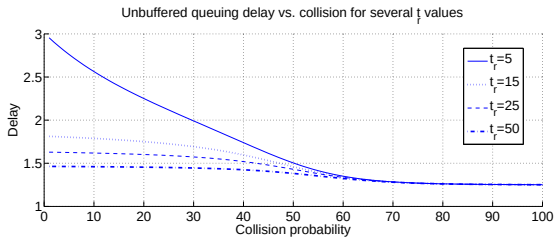

Figure 2.19. *Plots of $D(\gamma)$ vs. γ : Average delay, function of the collision probability for different values of t_r. IEEE802.16e default values : b_0=16, m=16, and n=50*

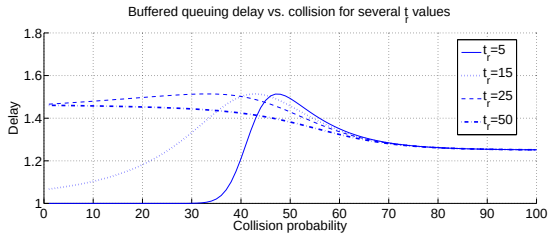

Figure 2.20. *Plots of $D(\gamma)$ vs. γ : Average delay, function of the collision probability for different values of t_r. IEEE802.16e default values : b_0=16, m=16, n=50 and buffer=6*

With this contribution we provide a complete analytical model for the MAC layer performance. Due to the Fixed Point Analysis, we provide the attempt rate behavior and its impact of the collision probability in function of the communication parameters. First the study reveals that the t_r parameter is the main performance factor, and an accurate tuning study could lead to a great performance enhancement. Second, b_0 and particularly m can also be tuned to increase the attempt rate without necessary impacts on the collision probability. Finally, we provide the collision statistics for a

large range of the user number. This testifies to the needs of a fitted ranging code partitioning to manage the pool of users.

The next step of our study consisted in defining the queuing performances relative to the requests incoming into the base station. First, we observe the impact of the code range N on the arrival rate in function of the collision probability and then of the number of users. The first one shows that the number of code does not enhance specifically the robustness of the system against the collision. But the second one testifies that the code range have to be designed in function of the pool of users : the performance can drastically fall with an unadapted number of ranging codes.

The last topic of the queuing analysis dealt with the delay experienced by the incoming requests. The unbuffered case leads to the observation which has motivated originally our study. The average queuing delay is largely smaller than the one expected by the standard. The nodes could wait up to 50 MAC frames to acknowledge a request loss, but our simulations show that this delay does not exceed 3 frames, for 50 user using a t_r parameter equal to 5. Please note that this experienced delay decrease with the t_r increasing. In addition, this last remark leads to an other observation : by increasing the t_r parameter, we also decrease the arrival rate as well as the queuing delay. So, future works are required to define an adapted way to tune the t_r parameter in function of the number of users, and the actual queuing delay at the base station. The buffered case shows, as expected that for the smallest values of t_r, some overflow drops occur. But it also reveals that the delay performances evolve in function of the collision probability. The t_r parameter can achieve better performance with an increasing collision probability. So here too, a tuning study could be completed.

2.3. Erlangian approach

2.3.1. *Problem formulation*

The Erlang capacity of a given system refers to the amount of traffic that can be handled by the system for a given target blocking probability while achieving a certain QoS requirement.

Traffic, both offered and accepted, is given in terms of a mean arrival rate, assuming arrivals follow a Poisson process, divided by a mean service rate, assuming service obeys to an Exponential distribution. Users in this case follow then a dynamic configuration, i.e., they come and leave the system after a finite duration, as opposed to static users that come to the system at time 0 and remain for the whole duration of analysis, such as long-lived flows.

Resources are in our case formulated in terms of (dedicated) subcarriers allocated to users. The latter can be of two types: streaming and elastic. The former are characterized by a constant-bit-rate requirement and are thus allocated subcarriers for the

whole duration of their call, which is independent of the quantity of resources they receive. Elastic flows are on the contrary driven by a maximum throughput; their transfer time is proportional to the amount of resources they get. When using TCP at the transport layer, they have the ability to share resources fairly among themselves; such a behavior can be modelled using Processor Sharing (PS). We next detail how subcarriers are allocated in OFDMA systems.

2.3.2. *Sub-carrier allocations*

OFDMA is a multiple access technique which divides the total Fast Fourier Transform (FFT) space into a number of sub-channels (set of sub-carriers that are assigned for data exchange) whereas the time resource is divided into time slots (i.e. in WiMAX OFDMA PHY [YAG 04], the minimum frequency-time unit of sub-channelization is one slot, which is equivalent to 48 sub-carriers) and a frame is constructed by a number of slots.

As stated earlier, WiMAX standards [802 04] specify two different distributed allocation modes which impact greatly the capacity of the system. This actually refers to the way the pilot allocation is performed in an OFDM symbol which specifies the type of sub-channelization: Fully Used Sub-Channelization (FUSC) occurs if the pilot sub-carriers are allocated first and the remaining sub-carriers are divided into data sub-channels. In the other sub-channelization method, called Partially Used Sub-Channelization (PUSC), data and pilot sub-carriers are partitioned into sub-channels, and then within each sub-channel, pilot sub-carriers are allocated. All UL sub-frames use PUSC mode, while DL sub-frames could use FUSC or PUSC. For example, in the downlink of a WiMAX system with FFT size of 1024 and after reserving the pilot and guard sub-carriers, a FUSC allocation corresponds to $L = 16$ sub-channels of $K = 48$ data sub-carriers each, while a PUSC allocation corresponds to $L = 30$ sub-channels, each containing $K = 24$ data sub-carriers. Please note that we hereafter use FUSC and consider one burst per frame.

In OFDMA-based WiMAX system, resource allocation is done in time-frequency domain: a call may share a sub-channel with other users. This is illustrated in Figure 2.21 where users $2, 3, 4$ and 5 occupy each one sub-channel half of the time while user 1 occupies one sub-channel all the time. With OFDMA, the user device could choose sub-channels based on geographical location with the potential of eliminating the impact of deep fades.

2.3.3. *Interference*

When cellular networks are designed using OFDMA technology, inter-cell interference appears as the limiting problem. In the downlink for instance, inter-cell interference occurs at a mobile station when a nearby base station transmits data over a

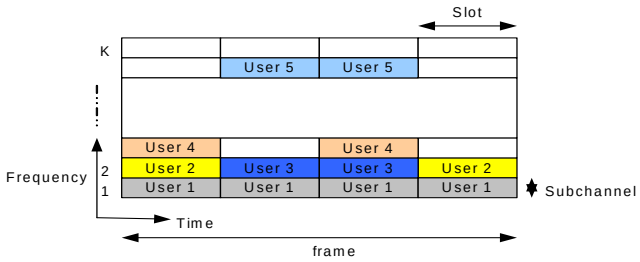

Figure 2.21. *Time-frequency resource allocation in OFDMA WiMax system*

subcarrier used by its serving base station, as illustrated in Figure 2.22. This is called collision and, depending on the number of interfering base stations, we can have more than one collision at the same subcarrier[4]. As the frequency is allocated in WiMAX on the basis of subchannels, each consisting of several subcarriers, different scenarios are possible:

– In the case of adjacent allocation, when a collision occurs, all the subcarriers of the subchannel are involved. Frequency hopping is then necessary in order to distribute the interference between users.

– For distributed allocation, frequency diversity is ensured when constructing the subchannels, thus leading to an averaged interference between calls.

However, authors in [ELA 06d] showed that the number of collisions is independent of the allocation mode, and is always distributed following a hyper-geometric distribution when the system is homogeneous.

When a frequency reuse of 1 is supported, i.e., all cells/sectors operate on the same frequency channel to maximize spectral efficiency, the inter-cell interference is a major concern due to heavy cochannel interference (CCI). Users at the cell edge may thus suffer degradation in connection quality. This cell edge interference problem has been addressed by appropriately configuring frequency usage without resorting to traditional frequency planning. Indeed, the classical interference avoidance scheme is obtained by dividing the frequency band into 3 equal subbands and allocate the subbands to the cells so that adjacent cells always use different frequencies. This scheme, called reuse 3 scheme and illustrated in Figure 2.23, is possible using the PUSC mode. The underlying idea is to allow interference only from cells located in ring 2, leading thus to low interference.

4. Please note that a collision does not necessarily mean loss of a subcarrier but merely a probability on that event, as quantified in [ELA 06d].

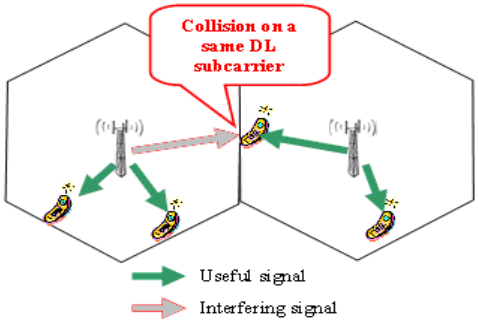

Figure 2.22. *Inter-cell interference in WiMAX.*

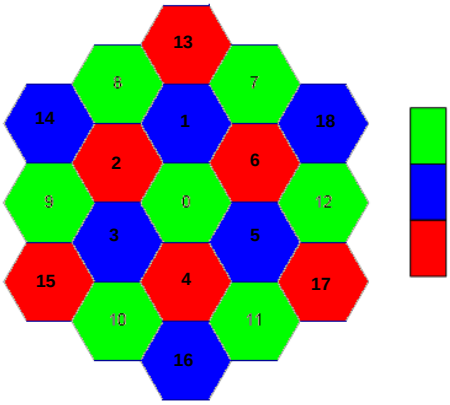

Figure 2.23. *Reuse 3 scheme: interfering cells are in ring 2.*

A hybrid solution between reuse 1 and reuse 3 schemes, also called reuse partitioning, has been proposed [WiMb]. The idea is to use a frequency reuse of 1 at the cell centers where interference is low, and a frequency reuse of 3 at the cell edges where users are more subject to interference. This is illustrated in Figure 2.24 and called fractional frequency reuse. This frequency allocation mode is possible in WiMAX using the PUSC mode. In fact, each segment in PUSC is decomposed into two groups, resulting in six different groups: 3 even groups of 6 subchannels each and 3 odd groups

of 4 subchannels each. All even groups can thus be allocated to the cell centers, while only one odd group is allocated to cell-edge users. This results in the loss of 2 odd groups (8 subchannels); to compare with the loss of two segments, equivalent to 20 subchannels when reuse 3 is used.

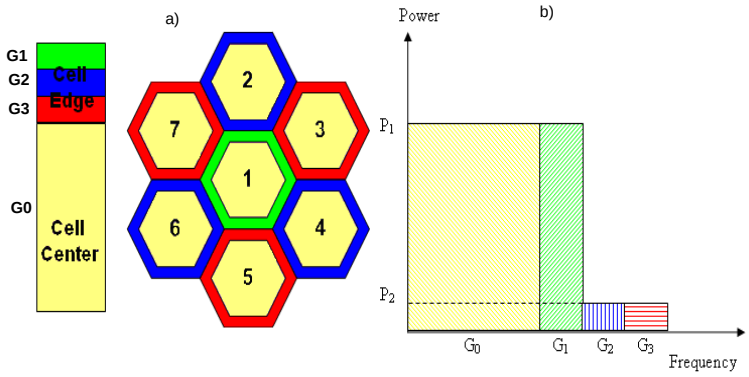

Figure 2.24. *a) Fractional frequency allocation scheme where a reuse 3 scheme is used at cell edges and b) power/frequency scheduling with reduced power in cell 1 at the frequencies used for cell edge users in the cells 2-7.*

When using this fractional reuse scheme, upon the arrival of a user, it is allocated a subchannel within the frequency band that corresponds to its position in the cell. As the location of the mobile cannot be precisely known, the choice is based on the path loss: A threshold on the path loss is fixed and terminal equipments with a path loss larger than this threshold are assigned a subchannel within the frequency reuse 3 bandwidth.

Even if the overall cell throughput is large in the hybrid frequency allocation scheme, there is still a loss of subchannels compared with the reuse 1 scenario. To overcome this problem, a proposed solution is to use a power control on some frequency bands to limit interference at the cell edges. In this context and referring to Figure 2.24-a, only cell 1 is allowed to transmit with full power using the "G_1" part of the spectrum while cells 2-7 are allowed to transmit in this part of the spectrum using only a reduced power. This is illustrated in Figure 2.24-b. This will reduce the downlink interference seen by cell-edge users served by cell 1 compared to a classical reuse 1 scheme. The radio resources used for transmission to UEs in a cell are controlled by the scheduler in the base station and fractional reuse can therefore be implemented as part of the scheduling decision. Fractional reuse can thus simply be seen as constraints to the scheduler.

This scheme has been proposed for 3G LTE systems [3GP 05]. This power/frequency scheduling is possible in WiMAX [ELA 06b], as for fractional reuse, using the even and odd groups in the PUSC mode, with the difference that all groups are used in each cell with different powers. Only 22 subchannels are used with full power (18 subchannels for cell center users are 4 for cell-edge ones). The remaining 8 subchannels are allowed to be used within the cell center with a reduced power ($P_2 = P_1/R$, with $R > 1$), only when the 18 subchannels assigned for cell center are occupied.

2.3.4. *AMC and cell decomposition*

AMC, in the presence of path loss only, denoted by ξ, yields high efficiency modulation is used for users where $\xi_i \ll \xi$, corresponding to a large SNR[5]. This results in the division of the cell into r regions, $i = 1...r$ (see Figure 2.25), which we assume to be concentric circles of radius R_i for simplicity, but might be of different topology if we take into account other phenomena, such as fast-fading. In each region, users have the same modulation scheme and experience thus a corresponding bit rate which decreases as users get further from the base station.

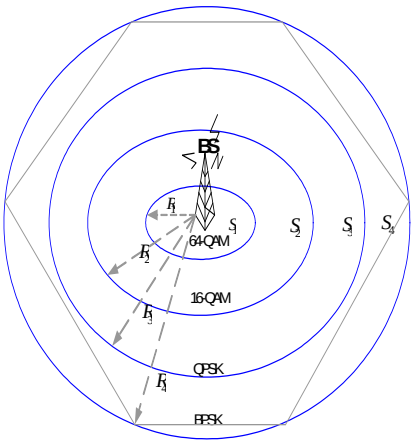

Figure 2.25. *Cell decomposition into regions*

To calculate the area covered by each modulation scheme, we must determine the maximal distance R_i between Base Station (BS) and users using a corresponding modulation. This distance is determined using the maximal SNR a user should receive

5. Please note that, for the time being, only the SNR matters, and not SINR, the Signal to Interference plus Noise ratio, as we now talk about the case of one cell in isolation. In the next subsection, the multiple-cell setting will arise along with underlying interference and SINR.

without data loss. Different values of received SNR for different modulation/coding schemes have been calculated in Reference [802 04] and are shown in Table I (first three columns). We use them to calculate R_i [TAR 07].

The path loss for the free space model is given by [STU 01]:

$$\begin{aligned} PLi[dB] &= -10\log\left[G_E G_R(\frac{\lambda}{4\pi R_i})2\right] \\ &= -10\log G_E - 10\log G_R + 20\log(\frac{4\pi R_i}{\lambda}) \end{aligned}$$

where G_E is the emitter antenna gain, G_R is the receiver antenna gain, R_i is the distance between the emitter and the receiver and λ is the wavelength. This path loss is also equal to

$$PLi[dB] = P_E[dBm] - SNR[dB] - N[dBm]$$

where P_E is the emitted power and N is the thermal noise (in units of decibels) which is equal to:

$$N[dBm] = 10\log(\tau TW) \tag{2.7}$$

$\tau = 1.38 \cdot 10^{-23} watt/K - Hz$ is the Boltzmann constant, T is the temperature in Kelvin ($T = 290$) and W is the transmission bandwidth in Hz.

Using the above equations, we can calculate the relationship between the distance and the SNR as follows:

$$R_i = \frac{\lambda * 10^{\frac{P_E[dBm]+10\log(G_E)[dB]+10\log(G_R)[dB]-SNR[dB]-N[dBm]}{20}}}{4\pi} \tag{2.8}$$

The area of each region S_i is given by:

$$S_i = \pi \cdot (R_i^2 - R_{i-1}^2)$$

where $R_0 = 0$.

For the sake of illustration, let us consider the following example based on the licensed band for WiMAX to outdoor use in France which starts at a frequency of $3.4GHz$ and which has system bandwidth equal to $20MHz$. At this bandwidth, the thermal noise is equal to $-100.97dBm$. According to the maximum allowed *Effective Isotropic Radiated Power (EIRP)* of $1W$, where the emitters are assumed to have an emission power P_E of $1W$ for users. We consider the case of antennas in BS and user equipment without gain. In Figure 2.26, we represent the distance assigned to SNR for switching points. The proportion of each surface area per PHY assumption is determined and shown in in Table 2.4.

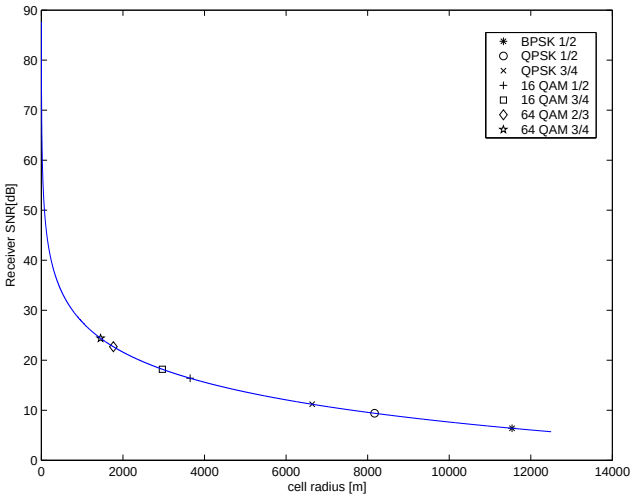

Figure 2.26. *Received SNR function of the distance*

Modulation	Coding rate	Receiver SNR(dB)	Surface [%]
BPSK	1/2	6.4	39.4
QPSK	1/2	9.4	20.75
	3/4	11.2	28.0
16 QAM	1/2	16.4	4.07
	3/4	18.2	5.14
64 QAM	2/3	22.7	0.9
	3/4	24.4	1.74

Table 2.4. *IEEE802.16 PHY assumptions*

2.3.5. *Flow throughput*

The instantaneous physical bit rate $\bar{R}_i^{s,e}$ of streaming or elastic users in region S_i is given by:

$$\begin{aligned} \bar{R}_i^{s,e} &= \frac{L_i^{s,e} \times K \times C \times \log_2(M)}{T_s \times S_c} \times (1 - BLER) \\ &= L_i^{s,e} \times K \times B \times E_i \times (1 - BLER) \end{aligned} \tag{2.9}$$

where $L_i^{s,e}$ is the number of sub-channels to be assigned to streaming/elastic users in region S_i, K is the number of data sub-carriers assigned to each sub-channel, C is the

coding rate of the M-ary modulation, T_s is the OFDMA symbol duration given by:

$$T_s = T_b + T_g$$

with T_b the useful symbol period (in units of microseconds) given by $\frac{N}{W \times n}$ and T_g the guard period equal to $G \times T_b$, W is the bandwidth (MHz), n is the sampling factor, G is the ratio of cyclic prefix (CP) to useful time, S_c is the sector coefficient (S_c is equal to 1 in FUSC and 3 in PUSC 3 sectors), B is the baud rate (symbols/sec), E_i is the efficiency of the modulation (bits/symbol) in each region S_i and $BLER$ is the perceived Block Error Rate[6].

2.3.6. *Capacity evaluation*

References [TAR 06] and [TAR 07] consider the issue of capacity with dynamic arrival and departure of streaming and elastic uses to the system, with a priority to the former over the latter which share the left-over capacity on a processor sharing basis.

In [TAR 06], emphasis is set on one region only. Using the Quasi-Stationary (QS) assumption [BEN 01], wherein streaming calls are assumed to arrive and leave the system in a manner slower than that of data ones, i.e., the ratio of the mean arrival rate of streaming flows to that of elastic ones, λ^s/λ^e, is very small, and so, in the presence of n^s streaming calls, the n^e data flows may be studied as if they were in a stationary regime, following an M/G/1 Processor Sharing (PS) queue.

Let $\bar{n^e}$ denote the mean number of data flows in the system[7]. It is given by:

$$\bar{n^e} = \sum_{k=0}^{\infty} kPr(n^e = k) \tag{2.10}$$

where

$$Pr(n^e = k) = \sum_{j} Pr(n^e = k | n^s = j) Pr(n^s = j) \tag{2.11}$$

and

$$Pr(n^e = k | n^s = j) = \frac{1}{G} \Pi_{i=1}^{k} \frac{\lambda^e}{\mu^e(j,i)} \tag{2.12}$$

where G is the normalizing constant.

The distribution of n^s is given by an M/M/m/m queue :

---

6. Note that for each value of SINR, we can determine a couple of values $(E, BLER)$ and these values are determined by link level curves $E = f(SINR)$ and $BLER = g(SINR)$

7. Note that in the absence of any admission control, this number can go to infinity

$$Pr(n^s = j) = Pr(n^s = 0)\Pi_{k=0}^{j-1} \frac{\lambda^s}{(k+1)\mu^s} \tag{2.13}$$

for all $j \leq n^s_{max}$ and where $Pr(n^s = 0)$ is determined by the normalization condition $\sum_i Pr(n^s = i) = 1$.

In [TAR 07], the whole cell is considered, first in isolation, where mostly AMC is considered, and then in a multiple-cell setting, i.e., taking into account interference too. The analysis follows in this case an exact Markovian model. Steaming calls are assumed to arrive to region S_i according to a Poisson process with intensity λ^s_i and use L^s_i sub-channels for an exponentially distributed time with mean $1/\mu^s$ independent of the share of the resources they get. Elastic flows are assumed to arrive to the system according to a Poisson process with intensity λ^e_i and assumed for tractability to have a service exponentially distributed with mean $\mu^e_i = \frac{R^e_i}{E[Z]}$ where $E[Z]$ is the mean file size[8].

The system can be modelled as a Continuous Time Markov Chain (CTMC) by taking into account the proposed priorities for the integration of streaming and elastic flows as well as the way they share resources.

The state is characterized by the following row vector

$$\overrightarrow{n} := (n^s_1, n^s_2, ..., n^s_r, n^e_1, n^e_2, ..., n^e_r)$$

where n^s_i and n^e_i, for $i = 1...r$, represent the number of streaming and elastic calls in region S_i, respectively.

The state space of the system is given by

$$\mathfrak{S} := \{\overrightarrow{n} \in \mathbb{N}^{\mathbf{2r}} | \sum_{i=1}^{r} (L^s_i n^s_i + L^e_i n^e_i) \leq L\} \tag{2.14}$$

where L^s_i and L^e_i denote the number of sub-channels allocated to streaming and elastic calls in region S_i respectively and L is the maximum number of sub-channels in the cell.

The steady-state probability vector is given by $\overrightarrow{\Pi} = \{\pi(\overrightarrow{n})_{\overrightarrow{n} \in \mathfrak{S}}\}$. Note that the corresponding system is non homogeneous as the departure rate of elastic calls depends on the overall number of calls in the system whereas streaming calls do not.

8. In fact, the total length of an elastic flow in units of packets is found to follow a log normal distribution, according to the measurement-based modelling [DOW 01]

The solution of the steady-state distribution is obtained by solving the set of linearly independent equations given by:

$$\begin{array}{l} \overrightarrow{\Pi} \cdot Q = 0 \\ \sum_{\overrightarrow{n} \in \mathfrak{S}} \pi(\overrightarrow{n}) = 1 \end{array} \tag{2.15}$$

To construct the transition matrix Q, all possible transitions between neighboring states should be considered. Let $q(\overrightarrow{n} \rightarrow \overrightarrow{n'})$ denote the transition probability from state $\overrightarrow{n}$ to neighboring states $\overrightarrow{n'}$. Note that when a new call is accepted in region S_i, $1 \leq i \leq r$ the state is noted by $\vec{n}_{i^+}^{s,e}$ and when a call terminates the service the next state is $\vec{n}_{i^-}^{s,e}$. We thus have the following transition rates:

$$\begin{array}{l} q(\overrightarrow{n} \rightarrow \overrightarrow{n_{i^+}^s}) = \lambda_i^s \\ q(\overrightarrow{n} \rightarrow \overrightarrow{n_{i^-}^s}) = n_i^s \mu^s \\ q(\overrightarrow{n} \rightarrow \overrightarrow{n_{i^+}^e}) = \lambda_i^e \\ q(\overrightarrow{n} \rightarrow \overrightarrow{n_{i^-}^e}) = \frac{n_i^e \mu_i^e(\overrightarrow{n})}{E[Z]} \end{array} \tag{2.16}$$

and the values $q(\overrightarrow{n} \rightarrow \overrightarrow{n})$ must be obtained as the sum of all terms in each line in matrix Q is equal to zero for $1 \leq i \leq r$.

This analysis enables to quantify several performance measures, namely the blocking probabilities for both types of traffic, streaming and elastic, as well as mean transfer time of elastic flows. Results in [TAR 07] show that in terms of blocking there is only one class for streaming flows in the inner and outer regions. Data flows however are elastic and share capacity among themselves in a fair manner on the basis of processor sharing. This makes them obtain the same blocking rate. They however obtain different mean transfer times in each region corresponding to the bit rate they achieve therein.

In a multiple-cell setting, with reuse partitioning, for streaming flows, the blocking probability in the inner region decreases as flows in this region have now access to all sub-channels whereas flows in the outer region do not. The latter have thus higher blocking. For data flows however, both blocking rates, inner and outer, increase with respect to the case with no frequency reuse, as more streaming flows are now accepted; with a higher increase in the outer ring as less sub-channels are now available.

Reference [ELA 06a] is also on capacity in OFDMA. Frequency reuse is considered as the major concern because the channel can only be reused when interference is low but its reuse increases the number of collisions and hence interference. Solutions are sectorization or special allocation of subcarriers in the edge of the cell. The authors note that collision does not necessarily mean loss of symbol, it just makes it more probable. Authors in this work first calculate the mean number of collisions in

a multicell setting, and then calculate the corresponding probability of SNR degradation, and this for a given load. They numerically consider the Erlang capacity of such a system, including the trade-off between dimensioning and modulation, in the presence of both streaming and elastic flows.

As of the frequency planning schemes, an analytical model has been proposed in [ELA 06b] to evaluate their performance based on a queuing analysis. It has been shown that a reuse 1 gives a high throughput to the cell, although the performances at the edge of cells are very bad. A reuse 3 scheme decreases severely all the throughput because only a third of the capacity is used in each cell. A reuse 1 in the center of cells combined with a reuse 3 in the edges of cells can realize an acceptable compromise between the total throughput and the performance of edge of cells [this is partitioning!]. Finally, power/frequency scheduling scheme realizes a high cellular throughput with an acceptable performance at cell edges and can then be considered as the best compromise.

Reference [NUA 06] is on OFDM, with streaming and elastic traffic, including inter-cell interference with frequency reuse 1/1 and 1/3, with Erceg propagation model and slow fading. The work is based on simulations, using Matlab. They calculate the outage time, the time during which the received signal is below a certain threshold, the mean download time and the cell capacity (in Mbps), all as a function of number of users in the cell.

This work contains two other references [BAL 05b] [BAL 05a] on works on performance of 802.16 using simulations only. Their criticism is that the other works use complex simulators, whereas they don't.

Eventually, Reference [PAN 07] is on intercell interference in wireless broadband access in general, not only 802.16 (no mention of OFDMA for instance), as capacity is mostly dependent on intercell interference. The analytical models for intercell interference are developed both in the uplink and downlink and take into account especially the effect of rain fading in the expression of the path loss.

Bibliography

[3GP 05] 3GPP, "Soft Frequency Reuse Scheme for UTRAN LTE", *R1-050507, Huawei*, 2005.

[802 04] 802.16-2004 I. S., "Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, *IEEE Standard for local and Metropolitan Area Networks*", 2004.

[802 05] 802.16-2005 I. S., "Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, *IEEE Standard for local and Metropolitan Area Networks*", 2005.

[AND 06] ANDREWS J., *OFDMA*, Prentice Hall, 2006.

[BAC 02] BACCARELLI E., FASANO A., BIAGI M., "Novel Efficient Bit-Loading Algorithms for Peak-Energy-Limited ADSL-Type Multicarrier Systems", *IEEE Trans. on Signal Processing*, May 2002.

[BAL 05a] BALL C., HUMBURG E., IVANOV K., TREML F., "Performance Analysis of. IEEE802.16 based cellular MAN with OFDM-256 in mobile Scenarios", *Proceeding of VTC Spring*, 2005.

[BAL 05b] BALL C., HUMBURG E., IVANOV K., TREML F., "Performance evaluation of IEEE802.16 WiMax with fixed and mobile subscribers in tight reuse", *Proceeding of PIMRC'2005*, 2005.

[BEN 01] BENAMEUR N., FREDJ S. B., DELCOIGNE F., OUESLATI-BOULAHIA S., ROBERTS J., "Integrated Admission Control for Streaming and Elastic Traffic", *Proceeding of QofIS 2001, Coimbra*, 2001.

[BIA 00] BIANCHI G., "Performance Analysis of the IEEE 802.11 distributed coordination function", *IEEE Journal on Selected Areas in Communications (JSAC)*, March 2000.

[CAM 98] CAMPELLO J., CIOFFI J., "Optimal Discrete Loading, ANSI Contribution T1E1.4/98-341", 1998.

[CHA 06a] CHAN L.-F., CHAO H.-L., CHOU Z.-T., "Two-Tier Scheduling Algorithm for Uplink Transmissions in IEEE 802.16 Broadband Wireless Access Systems", *Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, p. 1-4, 2006.

[CHA 06b] CHANDRA S., SAHOO A., An Efficient Call Admission Control for IEEE 802.16 Networks, Technical report, IITB/KReSIT, 2006.

[CHE 93] CHENG R. S., VERDU S., "Gaussian Multiaccess Channels with ISI: Capacity Region and Multiuser Water-filling", *IEEE Trans. on Information Theory*, May 1993.

[CHE 05] CHEN J., JIAO W., WANG H., "A Service Flow management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode", *Proceedings of IEEE International Conference on Communications (ICC2005)*, Seoul, Korea, 2005.

[CHO 95] CHOW P., CIOFFI J., BINGHAM J., "A Practical Discrete Multitone Transceiver Loading Algorithm for Data Transmission over Spectrally Shaped Channels", *IEEE Trans. on Communications*, February 1995.

[CHO 05a] CHO D.-H., SONG J.-H., KIM M.-S., HAN K.-J., "Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network", *Proceedings of the First International Conference on Distributed Frameworks for Multimedia Applications (DFMA'05)*, p. 130–137, 2005.

[CHO 05b] CHO D., SONG J., KIM M., HAN K., "Performance analysis of the 802.16 Wireless Metropolitan Area Network", *Proc. of the First International Conference on Distributed Framework for multimedia applications*, p. 130-137, 2005.

[CHU 02] CHU G., WANG D., MCI S., "A QoS architecutre for the MAC protocol of IEEE 802.16 BWA Systems", *Proceeding of IEE International Conference on Communications, Circiut and System and west Sino Exposition 2002*, p. 435-439, 2002.

[CIC 06] CICCONETTI C., LENZINI L., MINGOZZI E., EKLUND C., "Quality of service support in IEEE 802.16 networks", *IEEE Network*, March 2006.

[CIC 07] CICCONETTI C., ERTA A., LENZINI L., MINGOZZI E., "Performance Evaluation of the IEEE 802.16 MAC for QoS Support", *IEEE Trans. on Mobile Computing*, January 2007.

[COV 79] COVER T., GAMAL A. E., "Capacity theorems for the relay channel", *IEEE Trans. on Information Theory*, September 1979.

[DOW 01] DOWNEY A., "The structural cause of file size distributions", *ACM SIGMETRICS Performance Eval. Rev.*, June 2001.

[ELA 06a] ELAYOUBI S., FOURESTIÉ B., AUFFRET X., "On the capacity of OFDMA 802.16 systems", *Proceedings of IEEE International Conference on Communications (ICC)*, 2006.

[ELA 06b] ELAYOUBI S., HADDADA O. B., FOURESTIÉ B., *On the best frequency reuse scheme in WiMAX*, World Scientific Review, 2006.

[ELA 06c] ELAYOUBI S.-E., FOURESTIE B., "NXG02-1: On Inter-Cell Interference and Adaptive Modulation in OFDMA WiMAX Systems", *Proceedings of IEEE Global Telecommunications Conference*, p. 1–5, 2006.

[ELA 06d] ELAYOUBI S.-E., FOURESTIÉ B., "On frequency allocation schemes in 3G LTE systems", *Proceedings of IEEE PIMRC 2006*, 2006.

[FAS 03] FASANO A., "On the Optimal Discrete Bit Loading for Multicarrier Systems with Constraints", *Proceedings of the IEEE Vehicular Technology Conference*, 2003.

[FAT 02] FATTAH H., LEUNG C., "An overview of scheduling algorithms in wireless multimedia networks", *IEEE Wireless Communications*, October 2002.

[GAM 05] GAMAL A. E., ZAHEDI S., "Capacity of a class of relay channels with orthogonal components", *IEEE Trans. on Information Theory*, May 2005.

[GAM 06] GAMAL A. E., MOHSENI M., , ZAHEDI S., "Bounds on capacity and minimum energy-per-bit for AWGN relay channels", *IEEE Trans. on Information Theory*, April 2006.

[GE 06] GE Y., KUO G.-S., "An Efficient Admission Control Scheme for Adaptive Multimedia Services in IEEE 802.16e Networks", *Proceedings of IEEE 64th Vehicular Technology Conference, Fall*, 2006.

[GHO 05] GHOSH A., WOLTER D. R., "Broadband wireless access with WiMAX/802.16 : current performance benchmarks and future potential", *IEEE Communications Magazine*, February 2005.

[GOL 97] GOLDSMITH A., VARAIYA P., "Capacity of Fading Channels with Channel Side Information", *IEEE Trans. on Information Theory*, November 1997.

[GOL 99] GOLDEN G., FOSCHINI C., VALENZUELA R., , WOLNIANSKY P., "Detection Algorithm and Initial Laboratory Results using V-BLAST Space-Time Communication Architecture", *Electronics Letters*, January 1999.

[GUO 05] GUO D., SHAMAI S., VERDÚ S., "Mutual Information and Minimum Mean-Square Error in Gaussian Channels", *IEEE Trans. on Information Theory*, April 2005.

[HAN 05] HAN Z., JI Z., LIU K. J. R., "Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining and Coalitions", *IEEE Transactions on Communications*, August 2005.

[HØS 05] HØST-MADSEN A., ZHANG J., "Capacity bounds and power allocation for wireless relay channels", *IEEE Trans. on Information Theory*, June 2005.

[HUG 87] HUGHES-HARTOGS D., "Ensemble Modem Structure for Imperfect Transmission Media", *U.S. Patent 4 679 227*, July 1987.

[HWA 04] HWANG E. S., CHO C. H., SEO H. H., RYU B., LEE W., "A study of code partioning scheme of efficient random access in OFDMA-CDMA ranging subsystem", *Proceedings of JCCI*, Page262, 2004.

[JAN 03] JANG J., LEE K., "Transmit Power Adaptation for Multiuser OFDM Systems", *IEEE Journal on Selected Areas in Communications (JSAC)*, February 2003.

[JIA 06] JIANG C.-H., TSAI T.-C., "Token bucket based CAC and packet scheduling for IEEE 802.16 broadband wireless access networks", *Proceedings of 3rd IEEE Consumer Communications and Networking Conference*, p. 183–87, 2006.

[JIA 07] JIA H. ZHANG Z. Y. G. C. P. L. S. O., "The Performance of IEEE 802.16 OFDMA System Under Different Frequency Reuse and Subcarrier Permutation Patterns", *Proc. of IEEE International Conference on Communications (ICC)*, 2007.

[JIN 04] JINDAL N., VISHWANATH S., GOLDSMITH A., "On the Duality of Gaussian Multiple-Access and Broadcast Channels", *IEEE Trans. on Information Theory*, May 2004.

[JOO 07] JOO HEO INSUK CHA K. C., "Effective adaptive transmit power allocation algorithm considering dynamic channel allocation in reuse partitioning-based OFDMA system", *Springer Journal on Wireless Personal Communications*, vol. 43, num. 2, p. 362–370, October 2007.

[KAL 89] KALET I., "The Multitone Channel", *IEEE Trans. on Communications*, February 1989.

[KIM 05a] KIM K., HAN Y., KIM S., "Joint subcarrier and power allocation in uplink OFDMA systems", *IEEE Commun. Lett.*, 2005.

[KIM 05b] KIM K., YOU J., KIM K., "Capacity evaluation of the ofdma-cdma ranging subsystem in IEEE 802.16-2004", *Proceedings of WiMob*, Montreal, Canada, 2005.

[KIV 03] KIVANC D., LI G., LIU H., "Computationally efficient bandwidth allocation and power control for an OFDMA system", *IEEE Journal on Selected Areas in Communications (JSAC)*, November 2003.

[KOH 04] KOHJASTEPOUR M. A., SABHARWAL A., AAZHANG B., "Lower bounds on the capacity of Gaussian relay channel", *Proc. Conf. Information Sciences and Systems (CISS)*, Princeton, NJ, p. 597-602, 2004.

[KRA 04] KRAMER G., "Models and theory for relay channels with receive constraints", *Proc. 42nd Annu. Allerton Conf. Communication, Control, Computing*, Monticello, IL, p. 1312-1321, 2004.

[KRA 05] KRAMER G., GASTPAR M., GUPTA P., "Cooperative strategies and capacity theorems for relay networks", *IEEE Trans. on Information Theory*, September 2005.

[LAN 00] LANEMAN J., WORNELL G., "Exploiting distributed spatial diversity in wireless networks", *Proc. Annu. Allerton Conf. Communication, Control, Computing*, Monticello, IL, 2000.

[LAN 01] LANEMAN J., WORNELL G., TSE D., "An efficient protocol for realizing cooperative diversity in wireless networks", *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Washington DC, 2001.

[LAN 03] LANEMAN J., WORNELL G., "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks", *IEEE Trans. on Information Theory*, Oct 2003.

[LAN 04] LANEMAN J., TSE D., WORNELL G., "Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behaviour", *IEEE Trans. on Information Theory*, December 2004.

[LEE 05a] LEE H., KNWON T., CHO D., "An enhacement Uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e System", *IEEE Communication Letters*, August 2005.

[LEE 05b] LEE H., KWON T., CHO D.-H., "An Enhanced Uplink Scheduling Algorithm Based on Voice Activity for VoIP Services in IEEE 802.16d/e System", *IEEE Communications Letters*, August 2005.

[LEE 06a] LEE H., KWON T., CHO D.-H., LIMT G., CHANGT Y., "Performance Analysis of Scheduling Algorithms for VoIP Services in IEEE 802.16e Systems", *Proceedings of IEEE*

63rd Vehicular Technology Conference, Spring, p. 1231–1235, 2006.

[LEE 06b] LEE J., KWON E., YEON H.-J., JUNG K., "Markov Model for Admission Control in the Wireless AMC Networks", *IEICE Transactions on Communications*, 2006.

[LI 01] LI L., GOLDSMITH A. J., "Capacity and Optimum Resource Allocation for Fading Broadcast Channels – Part I: Ergodic Capacity", *IEEE Transactions on Information Theory*, March 2001.

[LI 03] LI N., BATTITI R., "Performance analysis of an enhanced IEEE 802.11 distributed coordination function supporting service differentiation", *Proceedings of QofIS'03*, p. 152-161, 2003.

[LI 05] LI W., WANG H., AGRAWAL D., "Dynamic admission control and QoS for 802.16 wireless MAN", *Proceedings of Wireless Telecommunications Symposium*, p. 60-66, 2005.

[LIA 05] LIANG Y., VEERAVALLI V., "Gaussian othogonal relay channels: optimal resource allocation and capacity", *IEEE Trans. on Information Theory*, September 2005.

[LIU 05] LIU N., LI X., PEI C., YANG B., "Delay Character of a Novel Architecture for IEEE 802.16 Systems", *Proceedings of Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies*, p. 293–296, 2005.

[LO 07] LO E., CHANG P., LAU V., CHENG R., LETAIEF K., R.D.MURCH, MOW W., "Adaptive resource allocation and capacity comparisons of downlink multiuser MIMO-MC-CDMA and MIMO-OFDMA", *IEEE Trans. on Wireless Communications*, March 2007.

[LOZ 06] LOZANO A., TULINO A. M., VERDÚ S., "Optimum Power Allocation for Parallel Gaussian Channels with Arbitrary Input Distributions", *IEEE Trans. on Information Theory*, July 2006.

[MA 06] MA M., NG B., "Supporting Differentiated Services in Wireless Access Networks Mode", *Proceedings of 10th IEEE International Conference on Communication systems*, Singapore, p. 1-5, 2006.

[MEU 71] VAN DER MEULEN E., "Three terminals communication channels", *Ad. Appl. Probab.*, 1971.

[MIO 05] MIORANDI D., KUMAR A., ALTMAN E., GOYAL M., "New insights from a fixed point analysis of single cell IEEE 802.11", *Proceedings of IEEE Infocom*, Miami, 2005.

[MOR 06] MORRIS P., ATHAUDAGE C., "Fairness based resource allocation for multiuser MIMO-OFDM systems", *Proceedings of IEEE Vehicular Technology Conference (VTC), Spring*, p. 314-318, 2006.

[MOR 07] MORETTI M., TODINI A., "A resource allocator for the uplink of multi-cell OFDMA systems", *IEEE Transactions on Wireless Communications*, vol. 6, num. 8, p. 2807–2812, August 2007.

[NAS 04] NASSER N., HASSANEIN H., "Prioritized multi-class adaptive framework for multimedia wireless networks", *Proceedings of IEEE International Conference on Communications*, p. 4295–4300, 2004.

[NG 07] NG T. C.-Y., YU W., "Joint optimization of relay strategies and resource allocations in cooperative cellular networks", *IEEE Journal on Selected Areas in Communications*

(JSAC), February 2007.

[NIY 05a] NIYATO D., HOSSAIN E., "Connection admission control algorithms for OFDM wireless networks", *Proceedings of IEEE Global Telecommunications Conference*, Page 5, 2005.

[NIY 05b] NIYATO D., HOSSAIN E., "Queue-Aware Uplink Bandwidth Allocation for Polling Services in 802.16 Broadband Wireless Networks", *Proceedings of IEEE GLOBECOM*, 2005.

[NIY 06a] NIYATO D., HOSSAIN E., "Delay-Based Admission Control Using Fuzzy Logic for OFDMA Broadband Wireless Networks", *Proceedings of IEEE International Conference on Communications*, p. 5511–5516, 2006.

[NIY 06b] NIYATO D., HOSSAIN E., "A Game-Theoretic Approach to Bandwidth Allocation and Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks", *Proceedings of 3rd international conference on Quality of service in heterogeneous wired/wireless networks*, 2006.

[NIY 06c] NIYATO D., HOSSAIN E., "Joint Bandwidth Allocation and Connection Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks", *Proceedings of IEEE International Conference on Communications*, p. 5540–5545, 2006.

[NIY 06d] NIYATO D., HOSSAIN E., "Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks", *IEEE Trans. on Mobile Computing*, June 2006.

[NIY 06e] NIYATO D., HOSSAIN E., "A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks", *IEEE Trans. on Computers*, November 2006.

[NUA 06] NUAYMI L., NOUN Z., "Simple capacity estimations in WiMAX/802.16 systems", *Proceeding of PIMRC'2006*, 2006.

[PAN 04] PAN C., CAI Y., XU Y., "Adaptive Subcarrier and Power Allocation for Multiuser MIMO-OFDMA Systems", *Proceedings of IEEE International Conference on Communications (ICC)*, p. 2631-2634, 2004.

[PAN 07] PANAGOPOULOS A., ARAPOGLOU P.-D. M., KANELLOPOULOS J., COTTIS P., "Intercell radio interference studies in broadband wireless access networks", *IEEE Journal on Selected Areas in Communications (JSAC)*, January 2007.

[PER 06] PERUMALRAJA R., ROY J., RADHA S., "Multimedia Supported Uplink Scheduling for IEEE 802.16d OFDMA Network", *Proceedings of Annual India Conference*, p. 1–5, 2006.

[RAM 97] RAMJEE R., NAGARAJAN R., TOWSLEY D., "On optimal call admission control in cellular networks", *Wireless Networks*, March 1997.

[REZ 04] REZNIK A., KULKARNI S., VERDU S., "Degraded Gaussian multirelay channels: capacity and optimal power allocation", *IEEE Trans. on Information Theory*, December 2004.

[RON 07] RONG B., QIAN Y., CHEN H.-H., "Adaptive power allocation and call admission control in multiservice WiMAX access networks", *IEEE Wireless Communications*, February 2007.

[RYU 03] RYU B. H., CHO C., WON J. J., SEO H., LEE H. W., "Perfomance analysis of random access protocol in OFDMA-CDMA", *Proceedings of KICS Fall Conference*, 2003.

[SAA] SAAD KIANI G. Ø., GESBERT D., "Maximizing Multicell Capacity Using Distributed Power Allocation and Scheduling".

[SAY 06] SAYENKO A., ALANEN O., KARHULA J., HAMALAINEN T., "Ensuring the QoS requirements in 802.16 Scheduling", *Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems*, Torremolinos, Spain, p. 108–117, 2006.

[SEN 03a] SENDONARIS A., ERKIP E., AAZHANG B., "User cooperation diversity, Part I: System description", *IEEE Trans. on Information Theory*, November 2003.

[SEN 03b] SENDONARIS A., ERKIP E., AAZHANG B., "User cooperation diversity, Part II: Implementation aspects and performance analysis", *IEEE Trans. on Information Theory*, November 2003.

[SEO 06] SEONG K., MOHSENI M., CIOFFI J., "Optimal resource allocation for OFDMA downlink systems", *Proc. IEEE International Symposium on Information Theory*, Seattle, WA, 2006.

[SET 06] SETTEMBRE M., PULERI M., GARRITANO S., TESTA P., ALBANESE R., MANCINI M., CURTO V. L., "Performance analysis of an efficient packet-based IEEE 802.16 MAC supporting adaptive modulation and coding", *Proceedings of International Symposium on Computer Networks*, p. 11–16, 2006.

[SHE 05] SHEN Z., ANDREWS J., EVANS B., "Adaptive Resource Allocation in Multiuser OFDM Systems with Proportional Fairness", *IEEE Trans. on Wireless Communications*, December 2005.

[SIN 06] SINGH V., SHARMA V., "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks", *Proceedings of IEEE Wireless Communications and Networking Conference*, p. 984–990, 2006.

[SON 05a] SONG G., LI Y., "Cross-layer optimization for OFDM wireless networks Part I: theoretical framework", *IEEE Trans. on Wireless Communications*, March 2005.

[SON 05b] SONG G., LI Y., "Cross-layer optimization for OFDM wireless networks Part II: algorithm development", *IEEE Trans. on Wireless Communications*, March 2005.

[STI 98] STILIADIS D., VARMA A., "Latency-rate servers: a general model for analysis of traffic scheduling algorithms", *IEEE/ACM Transactions on Networking*, October 1998.

[STU 01] STUBER G. L., *Principles of Mobile Communication*, 2nd ed. Norwell, MA:Kluwer, 2001.

[SUN 06] SUN J., YAO Y., ZHU H., "Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems", *Proceedings of IEEE 63rd Vehicular Technology Conference*, p. 1221–1225, 2006.

[TAR 06] TARHINI C., CHAHED T., "System capacity in OFDMA-based WiMAX", *Proceedings of ICSNC 2006, Tahiti*, 2006.

[TAR 07] TARHINI C., CHAHED T., "System capacity in OFDMA-based WiMAX including AMC", *accepted, LANMAN'2007, Princeton NJ*, 2007.

[TEL 95] TELATAR I., Capacity of Multi-Antenna Gaussian Channels, Technical report, AT & T Bell Labs, 1995.

[TOU 06] TOUFIK I., KNOPP R., "Channel allocation algorithms for multi-carrier multiple-antenna systems", *Signal Processing*, August 2006.

[TSE 00] TSE D., "Optimal Power Allocation over Parallel Gaussian Broadcast Channels", *unpublished*, 2000.

[TSE 04] TSE D., VISWANATH P., ZHENG L., "Diversity-multiplexing tradeoff in multiple access channels", *IEEE Trans. on Information Theory*, September 2004.

[WAN 05a] WANG B., ZHANG J., HØST-MADSEN A., "On the capacity of MIMO relay channels", *IEEE Trans. on Information Theory*, January 2005.

[WAN 05b] WANG H., CHEN J., JIAO W., "A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode", *Proceedings of IEEE International Conference on Communications (ICC)*, p. 3422-3426, 2005.

[WAN 06] WANG H., HE B., AGRAWAL D., "Admission control and bandwidth allocation above packet level for IEEE 802.16 wireless MAN", *Proceedings of 12th International Conference on Parallel and Distributed Systems*, Page 6, 2006.

[WAN 07] WANG L., LIU F., JI Y., RUANGCHAIJATUPON N., "Admission Control for Non-preprovisioned Service Flow in Wireless Metropolitan Area Networks", *Proceedings of Fourth European Conference on Universal Multiservice Networks*, p. 243–249, 2007.

[wima] "Mobile WiMAX–Part I: A Technical Overview and Performance Evaluation".

[WiMb] WiMAX Forum, `http://www.wimaxforum.org/`.

[WON 99] WONG C., CHENG R., LETAIEF K. B., MURCH R., "Multiuser OFDM with adaptive sub-carrier, bit, and power allocation", *IEEE Journal on Selected Areas in Communications (JSAC)*, 1999.

[WON 03a] WONGTHAVARAWAT K., GANZ A., "IEEE 802.16 based last mile broadband wireless military networks with quality of service support", *Proceedings of IEEE Military Communications Conference*, p. 779–784, 2003.

[WON 03b] WONGTHAVARAWAT K., GANZ A., "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems", *International Journal of Communication Systems*, February 2003.

[XIA 03] XIAO Y., "Backoff-based priority schemes for IEEE 802.11", *Proc. of IEEE ICC*, 2003.

[XIA 04] XIAO Y., "An analysis for differentiated services in IEEE 802.11e wireless LANs", *Proc. of IEEE IGDCS'04*, 2004.

[XU 06] XU J., KIM J., PAIK W., SEO J.-S., “Adaptive resource allocation algorithm with fairness for MIMO-OFDMA systems”, *Proceedings of IEEE Vehicular Technology Conference (VTC), Spring*, p. 1585-1589, 2006.

[YAG 04] YAGHOOBI H., “Scalalable OFDMA Physical Layer in IEEE 802.16 Wireless MAN”, *Intel Technology Journal*, August 2004.

[YAN 06] YANG O., LU J., “Call Admission Control and Scheduling Schemes with QoS Support for Real-time Video Applications in IEEE 802.16 Networks”, *Journal of Multimedia (JMM)*, May 2006.

[YAO 05] YAO Y., CAI X., GIANNAKIS G. B., “On energy efficiency and optimum resource allocation for relay transmissions on in the low power regime”, *IEEE Trans. on Wireless Communications*, November 2005.

[YU 02] YU W., CIOFFI J., “FDMA Capacity of Gaussian Multiple-Access Channels with ISI”, *IEEE Trans. on Communications*, January 2002.

[YU 04] YU W., RHEE W., BOYD S., CIOFFI J., “Iterative Water-filling for Gaussian Vector Multiple Access Channels”, *IEEE Trans. on Information Theory*, January 2004.

[YU 06] YU W., CIOFFI J., “Constant-power Waterfilling: Performance Bound and Low-Complexity Implementation”, *IEEE Trans. on Communications*, January 2006.

[ZHA 05] ZHANG Y., LETAIEF K., “Adaptive resource allocation for multiaccess MIMO/OFDM systems with matched filtering”, *IEEE Trans. on Communications*, November 2005.

[ZHE 03] ZHENG L., TSE D., “Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels”, *IEEE Trans. on Information Theory*, May 2003.