Analytical Evaluation of QoS in the Downlink of OFDMA Wireless Cellular Networks Serving Streaming and Elastic Traffic

Mohamed Kadhem Karray

***Abstract*—The objective of the present paper is to build analytical methods for the evaluation of the *quality of service* (QoS) perceived by the users in the downlink of OFDMA wireless cellular networks serving streaming and elastic traffic.**

To do so, we first describe the *resource* (power and bandwidth) *allocation* problem and characterize its feasibility by some reference *feasibility condition* (FC). The QoS for FC may only by evaluated by simulations. To cope with this difficulty, we propose some *sufficient feasibility condition* (SFC) having the *multi-Erlang form* which permits analytical evaluation of the QoS. In particular, the *blocking probability* for streaming users can be calculated using *Kaufman-Roberts* algorithm. For elastic users, explicit expressions of the *throughputs* are obtained by using a multi-class *processor sharing* model. Moreover, we study the QoS in a network serving *simultaneously streaming and elastic* traffic.

We validate this approach by comparing SFC's blocking probabilities to these *simulated* for FC. Moreover, we illustrate the proposed approach by solving the *dimensioning* problem; i.e., evaluating the minimal density of base stations assuring acceptable QoS of a given traffic demand per surface unit.

***Index Terms*—Communication system performance, OFDMA, cellular, wireless, QoS, blocking probability, throughput.**

I. Introduction

OUR objective is to build analytical methods for the evaluation of the *quality of service* (QoS) perceived by the users in the downlink of Orthogonal Frequency-Division Multiple Access (OFDMA) wireless cellular networks.[1] In doing so, we aim to account for the *dynamics* induced by the call arrivals, service and departures.

Cellular networks can serve *streaming* (real-time) and *elastic* (non-real-time) traffic. Streaming calls require some given bit-rate for some given duration. Elastic connections aim to transmit some given volume of data at a rate that may be decided by the network. The QoS perceived by streaming users is typically the *blocking probability*, while the QoS perceived by elastic users is the mean *throughput*. Evaluation of these QoS parameters is crucial for the network *dimensioning*; i.e., evaluating the minimal number of base stations assuring some QoS (for some given traffic demand). This permits to minimize the network cost.

Manuscript received October 8, 2009; revised December 22, 2009; accepted February 19, 2010. The associate editor coordinating the review of this paper and approving it for publication was V. K. N. Lau.

M. K. Karray is with France Telecom, Research and Development Division, 38/40 rue du Général Leclerc, 92794 Issy-Moulineaux France (e-mail: mohamed.karray@orange-ftgroup.com).

Digital Object Identifier 10.1109/TWC.2010.05.091501

[1]Typical examples of such systems are the 3GPP Long Term Evolution (LTE) system and IEEE 802.16 WirelessMAN Air Interface standard (WiMAX).

A. Outline of our approach

The problem at hand may be decomposed into three subproblems: information theory, resource allocation and queueing theory. *Information theory* characterizes the performance of each single link from a base station to its mobile; i.e. expresses the bit-rate in the channel as function of the power and bandwidth allocated to this mobile. We shall assume such characterization given for example by Shannon's formula [12, Eq. (10.60)] in the most simple case of AWGN channel (other characterizations will be discussed in Section III-D).

We study the problem of allocating the resource (power and bandwidth) to the users respecting the information theoretic constraint. Besides, we consider the constraints on the maximal transmitted power and total bandwidth. We formulate this *resource allocation* problem and characterize its feasibility by some reference (necessary and sufficient) *feasibility condition* (FC).

Then we account for the dynamics induced by the call arrivals, service and departures. We consider a reference admission control which admits a new steaming user if and only if FC is satisfied. In this case, the blocking probability may be evaluated only by time-consuming simulations which are usually not appropriate for practical purposes such as network dimensioning. To cope with this problem, we propose some particular *sufficient feasibility condition* (SFC) which has the *multi-Erlang* form; i.e., it can be written as the weighted sum of the bit-rates of users not exceeding some constant.

This particular form permits to calculate the blocking probability using the *Kaufman-Roberts algorithm* [20], [26]. Moreover, assuming some appropriate separation of the time scale of the coexisting streaming and elastic traffic [13], one can also evaluate the mean throughput of the elastic traffic using a *multi-class processor sharing model*. These tools are in the field of *queueing theory*. Simulations show that the loss of capacity induced by SFC with respect to FC is acceptable. Thus SFC permits to build a rapid and accurate dimensioning method.

B. Paper organization

The remaining part of this paper is organized as follows. In the next subsection we discuss briefly the related work.

Our model is presented in Section II. In Section III we study the resource allocation problem and establish a reference and a sufficient condition for its feasibility denoted respectively by FC and SFC. This latter condition is used to construct an Erlang's loss model for streaming traffic in Section IV. SFC is also used to evaluate analytically the throughput of elastic users in Section V. The case of a network carrying both streaming and elastic traffic *on the same bandwidth* is studied in Section VI. The validation of SFC is described in Section VII. In Section VIII we illustrate the proposed approach by solving the *dimensioning problem*.

C. Related work

The problem of power allocation in Code-Division Multiple Access (CDMA) networks is addressed in many papers such as [27], [32], [33]. In [25] the power allocation problem is studied jointly with beamforming. More recently, an extensive literature addresses resource (power and bandwidth) allocation in OFDMA networks. Here are some examples [1], [21], [30]. In general it is however difficult to evaluate the QoS offered by the network with these methods implemented. Some studies consider the case of a single cell, e.g. [17], [31]. The multi-cell case is studied in [15] and [24]. In [14] different frequency reuse schemes are compared.

The present work adopts the approach proposed in [8] (with a background in [3], [5], [19]) that is implemented in the dimensioning tool of Orange. It consists in proposing some network control mechanism that is simple enough and can be studied by the classical tools of queueing theory. Moreover, we follow the ideas presented in [4] and in [10] for queueing models suitable for streaming and elastic traffics respectively. The present paper relies on and continues the work in [9]. Besides presenting in more detail the results there, we study the performance of a network serving *elastic traffic*, as well as a network serving *simultaneously* streaming and elastic traffic. Moreover, we illustrate the proposed approach by solving the dimensioning problem.

II. Basics

A. Model assumptions

We will consider a wireless network composed of a finite set of base stations (BSs). Each BS is equipped with a single antenna (no MIMO) and its total power is limited to some given maximal value. The same frequency spectrum is available to all BSs. There is no macro-diversity, i.e., each user is served by exactly one BS.

We assume that each user has a receiver with *single user detection* (as opposed to multiuser detection), thus the signals transmitted to the other users are considered as interference. Efficient codes are used to obtain bit-rates close to the information theoretic limit. Moreover, we neglect fading effects in a first analysis but extensions of the model to account for fading and real coding are possible as will be discussed in Section III-D.

Each BS allocates disjoint sub-carriers to its users. Thus, any given user receives other-BSs interference on the sub-carriers allocated to him by his BS.

Assumption 1: The number of interfering BSs is large and it is reasonable to assume that the interference power spectral density is constant in the whole spectrum. A suitable fast sub-carrier permutation (for a given configuration of users) may give a further justification of this assumption.

Assumption 2: We assume that each BS has a given geographic coverage region (*cell*) and each user in that region is assumed to be associated with that BS. We will not address in the present paper the more complex problem of the association of the BSs to the users independently of their geographic positions. In particular, we don't account for the effect of shadowing.

Despite the above simplifying assumptions, the problem at hand remains practically important and difficult to solve. Its solution will give a useful insight on key questions in cellular networks: resource allocation, QoS evaluation, dimensioning, cost optimization. It may be also a useful basis for future work studying more complex problems.

B. Notation

We present now the notations used in the paper. The reader may skip this section and go back to it when necessary.

1) Antenna locations and path loss:

- U is the set of base stations which is assumed finite.
- $u, v \in U$ are indexes for base stations.
- m, n are indexes for users (mobiles). The letter designating a base station (or a user) is sometimes used to designate its geographic position in $\mathbb{R}^2$. We denote $m \in u$ to say that user m is served by base station u.
- $L_{u,m}$ is the propagation-loss between base station u and user m (not including the fading).

2) Engineering parameters:

- W is the system bandwidth.
- N_0 is the power spectral density of external noise. We denote by $N = WN_0$ the noise power in the bandwidth W.
- w_m is the bandwidth allocated to user m.
- r_m is the bit-rate of user m.
- The powers are denoted as follows:
  - $\tilde{P}_u$ is the maximal total power emitted by a base station.
  - $\hat{P}_u$ is the power of common channels (not dedicated to a specific user) emitted by a base station; we assume that $\hat{P}_u = \epsilon \tilde{P}_u$ where ϵ is a given positive constant.
  - $P_{u,m}$ is the power emitted by base station u to user $m \in u$;
  - the total power emitted by base station u is denoted by

$$P_u = \hat{P}_u + \sum_{m \in u} P_{u,m} = \epsilon \tilde{P}_u + \sum_{m \in u} P_{u,m}. \quad (1)$$

III. Resource Allocation Feasibility Conditions

In OFDMA networks, each BS u allocates some number of sub-carriers of the total width w_m from the total spectrum of width W to each user $m \in u$, in such a way that two different users of the same BS have disjoint subsets of sub-carriers.

However, since the same frequency spectrum is allocated (assumed on average uniformly; see Assumption 1) by all BSs, user $m \in u$ receives interference from each BS $v \neq u$ of power $\frac{w_m}{W} P_v / L_{v,m}$. We assume that this interference acts as Gaussian noise[2], thus we assume that the SINR of user $m \in u$ is equal to

$$\text{SINR}_m = \frac{P_{u,m}/L_{u,m}}{w_m N_0 + \frac{w_m}{W} \sum_{v \neq u} P_v / L_{v,m}}$$

where N_0 is the power spectral density of the thermal noise. Information theory [12, Eq. (10.60)] implies that the bit-rate r_m of user m is bounded by

$$r_m \leq w_m \log_2 (1 + \text{SINR}_m), \quad m \in u.$$

The allocation problem in OFDMA may be formulated as follows. Find bandwidths (w_m), powers $(P_{u,m})$, and bit-rates (r_m) such that for all BS u and all user $m \in u$

$$\begin{cases} r_m \leq w_m \log_2 \left(1 + \frac{P_{u,m}/L_{u,m}}{w_m N_0 + \frac{w_m}{W} \sum_{v \neq u} P_v / L_{v,m}}\right) \\ P_u \leq \tilde{P}_u \\ \sum_{m \in u} w_m \leq W \end{cases} \tag{2}$$

which are the information theoretic, maximal-power and total-bandwidth constraints respectively. All powers, bit-rates and bandwidths should be nonnegative, but we will not write this explicitly in the formulations of our problems.

A. Reference feasibility condition (FC)

Definition 1: We will say that a vector of user bit-rates (r_m) is *feasible* if there exist powers $(P_{u,m})$ and bandwidths (w_m) such that the constraints in (2) are satisfied. In this case we will also say that (r_m) satisfies the *(reference) feasibility condition* (FC).

A natural and interesting idea is to use FC as admission control scheme. The network admits a new streaming call when its bit-rate associated to those of currently served users satisfy FC. Unfortunately, in this case the QoS (blocking probability, throughput) evaluation is intractable analytically. This is due to the fact that FC has not the multi-Erlang form; i.e., FC can not be written as the weighted sum of the bit-rates of users less than some constant (see Section IV-B for more discussion). To cope with this difficulty (analytical intractability), we will give a more explicit sufficient condition for the feasibility of the bit-rates. The particular form of this condition will permit an analytical evaluation of the QoS parameters in the subsequent sections.

B. First sufficient feasibility condition (SFC1)

In order to simplify the notation in the remaining part of the paper we introduce, for all BS u and all $m \in u$, the so-called *interference factor* (or *f-factor*)

$$f(m) = \sum_{v \neq u} \frac{L_{u,m}}{L_{v,m}} \frac{\tilde{P}_v}{\tilde{P}_u}$$

[2]Using [29, Theorem 18], we may show that the worst noise process distribution (not necessarily white nor Gaussian) for capacity with given second moment, is the AWGN.

(which is the interference to signal ratio when all the BSs transmit at their maximal powers). We introduce also the following slightly modified version of the interference factor

$$\hat{f}(m) = \frac{1}{1-\epsilon} \left(\frac{N L_{u,m}}{\tilde{P}_u} + f(m) \right). \tag{3}$$

Moreover, we introduce the notation

$$\xi_m = \frac{w_m}{W} \left(2^{r_m / w_m} - 1 \right). \tag{4}$$

Rewriting the above equation as follows $r_m = w_m \log_2 \left(1 + \frac{W}{w_m} \xi_m\right)$ shows that ξ_m is closely related to the SINR in Shannon's formula.

Proposition 1: Assume that there exist some user bit-rates (r_m) and some bandwidths (w_m) such that

$$\begin{cases} \sum_{m \in u} \hat{f}(m) \xi_m \leq 1, & u \in U \\ \sum_{m \in u} w_m \leq W, & u \in U \end{cases} \tag{5}$$

where $\hat{f}(m)$ and ξ_m are defined by (3) and (4) respectively. Then the bandwidth allocation (w_m) associated with the following power allocation

$$P_{u,m} = \hat{f}(m) \xi_m (1-\epsilon) \tilde{P}_u, \quad m \in u \in U \tag{6}$$

is solution of (2).

Proof: Assume that (5) holds true. The second inequality in (5) is precisely the total-bandwidth constraint. Let $(P_{u,m})$ be given by (6). Applying (1) we get $P_u = \epsilon \tilde{P}_u + (1-\epsilon) \tilde{P}_u \sum_{m \in u} \hat{f}(m) \xi_m$. Using the first inequality in (5) shows that $P_u \leq \tilde{P}_u$ which is the maximal-power constraint. It remains to show the information theoretic constraint of (2). To do so, note that (6) imply

$$\begin{aligned} \xi_m &= \frac{P_{u,m}}{\hat{f}(m)(1-\epsilon)\tilde{P}_u} \\ &= \frac{P_{u,m}}{N L_{u,m} + f(m) \tilde{P}_u} \\ &= \frac{P_{u,m}/L_{u,m}}{N + \sum_{v \neq u} \tilde{P}_v / L_{v,m}} \leq \frac{P_{u,m}/L_{u,m}}{N + \sum_{v \neq u} P_v / L_{v,m}} \end{aligned}$$

which implies the information theoretic constraint. ∎

We call the condition (5), the *first sufficient feasibility condition* (SFC1). Note that SFC1 is *decentralized*, i.e., it depends of the mobiles in each cell independently from those in the other cells. This simplifies the QoS evaluation, but simulations are always needed since SFC1 has not the multi-Erlang form (see Section IV-B).

C. Sufficient feasibility condition (SFC)

Proposition 2: Assume that the user bit-rates (r_m) satisfy the following condition

$$\sum_{m \in u} \frac{r_m}{\log_2 \left(1 + 1/\hat{f}(m)\right)} \leq W \tag{7}$$

where $\hat{f}(m)$ is given by (3). Then the bandwidth allocation

$$w_m = \frac{r_m}{\log_2 \left(1 + 1/\hat{f}(m)\right)} \tag{8}$$

associated with the power allocation given by (6) is solution of (2).

Proof: Assume that (7) holds true. Let (w_m) be defined by (8). Using (4), we get

$$\xi_m = \frac{w_m}{W}\left(2^{r_m/w_m} - 1\right) = \frac{w_m}{W}\frac{1}{\hat{f}(m)}$$

which may be rewritten as $\frac{w_m}{W} = \hat{f}(m)\xi_m$. On the other hand, note that (8) and (7) imply $\sum_{m\in u} w_m \leq W$. The previous two equations imply SFC1 (5). Using Proposition 1 finishes the proof. ∎

We call Condition (7) the *sufficient feasibility condition* (SFC). We shall compare SFC and SFC1 to the reference feasibility condition FC in Section VII.

D. Model extension: AWGN assumption

Till now we assumed AWGN channels for which the capacity is given by [12, Eq. (10.60)]

$$C_m = w_m \log_2(1 + \mathrm{SNR}_m) \tag{9}$$

where w_m is the bandwidth allocated to user m and SNR_m designates the signal to noise power ratio for user m. In [18] it is observed that for OFDMA systems implementing a family of M-QAM modulations (as those described in [16]) with some BER target, the AWGN capacity formula (9) should be replaced by $C_m = w_m \log_2\left(1 + \frac{\mathrm{SNR}_m}{a}\right)$ where $a = -\ln(5 \times \mathrm{BER})/1.5$. Thus accounting for real coding schemes may be taken into account in our approach by an appropriate modification of the AWGN capacity formula.

IV. Blocking Probabilities for Streaming Traffic

In order to evaluate the QoS in cellular networks, it is necessary to account not only for the geometry of interference (which leads to the resource allocation problem as explained in Section III), but also for the dynamics of call arrivals, service and departures. We consider here streaming traffic. The case of elastic traffic will be considered in Section V.

A. Traffic demand

Denote by $\mathbb{D} \subset \mathbb{R}^2$ the bounded region covered by the network; that is $\mathbb{D}$ is the union of all the cells. Consider only streaming calls whose inter-arrival times to $\mathbb{D}$ are independent and identically distributed (i.i.d.) exponential random variables with rate λ (mean $1/\lambda$). The position of each arrival is picked at random in $\mathbb{D}$ according to some distribution $Q(dm)$. We assume that users don't move during their calls. Each call requires to be served by the network at a given bit-rate during some service time. The durations of the different calls are assumed to be i.i.d. exponentially distributed with mean $1/\mu$. (This assumption may be relaxed due to the so-called insensitivity property [2, p.123], but this is not in the scope of the present paper.) The measure $\rho(dm) = \frac{\lambda}{\mu}Q(dm)$ is called *traffic demand density* (expressed in Erlangs per surface unit).

Elements $m \in \mathbb{D}$ denote *geographic locations of users* in the system. Configurations $\{m_i\} \subset \mathbb{D}$ of users in the system are identified by corresponding counting measures $\nu = \sum_i \varepsilon_{m_i}$; where the Dirac measure ε_m is defined by $\varepsilon_m(A) = 1$ if $m \in A$ and 0 otherwise, consequently $\nu(A)$ is the number of users in the set $A \subset \mathbb{D}$. We denote by $\mathbb{M}$ the set of *all finite configurations of users* (i.e., finite counting measures) on $\mathbb{D}$.

We denote by $\{N_t\}_{t\geq 0}$ the Markov process describing the evolution in time of the user configurations in $\mathbb{D}$ (due to arrivals and departures) in the absence of any admission control. It takes its values in $\mathbb{M}$. We call it the *free process*. By our previous assumptions $\{N_t\}_{t\geq 0}$ is ergodic and its stationary distribution, denoted Π, is the distribution of the Poisson process on $\mathbb{D}$ with mean measure $\rho(dm)$. In other words: the stationary free process (offered traffic) of positions of users is Poisson with mean measure equal to the traffic demand density. Moreover $\{N_t\}_{t\geq 0}$ is reversible with respect to Π.

B. Loss model

We assume that a given admission condition consists of verifying whether the current configuration of users with a new arrival belongs to some set of *feasible* configurations $\mathbb{M}^{\mathrm{f}}$. Denote the evolution of the free process modified (controlled) by the given admission condition by $\left\{N_t^{\mathrm{f}}\right\}_{t\geq 0}$. This process has the same dynamics as the free process except that the transitions (i.e., arrivals) that would lead outside $\mathbb{M}^{\mathrm{f}}$ are blocked. Such a modification of the Markov process is called *truncation* (of the free process) to $\mathbb{M}^{\mathrm{f}}$. Due to the reversibility of the free process, the truncated process $\left\{N_t^{\mathrm{f}}\right\}_{t\geq 0}$ admits as its invariant distribution the truncation of Π to $\mathbb{M}^{\mathrm{f}}$ (see [28, Proposition 3.14] for a proof).

The *blocking probability* is defined as the proportion of the blocked calls to the total number of arrivals in the long run of the system. One needs an efficient way to evaluate this probability. Such efficient method exists for some particular form of the admission condition as we explain in what follows. We say that the admission condition has the *multi-Erlang* form if the corresponding set of feasible configurations $\mathbb{M}^{\mathrm{f}}$ has the following form

$$\mathbb{M}^{\mathrm{f}} = \bigcap_{u\in U}\left\{\nu \in \mathbb{M} : \sum_{m\in u, m\in\nu} \varphi(m) \leq 1\right\} \tag{10}$$

where U is the set of all base stations, the sum of the values of some function $\varphi(m)$ is taken over of users m in configuration ν and served by BS u. Note that m is the user index which specifies its geographical location and also its bit-rate and serving base station. Note in particular that SFC (7) has the *multi-Erlang* form with

$$\varphi(m) = \frac{r_m}{W\log_2\left(1 + 1/\hat{f}(m)\right)}, \quad m \in u$$

whereas the reference feasibility condition FC doesn't have the multi-Erlang form.

The multi-Erlang form allows evaluation of the blocking probability by discretizing the cell u and using the Kaufman-Roberts algorithm described below. If the admission condition doesn't have the multi-Erlang form, as for example the reference feasibility condition FC, then time-consuming simulations are needed to calculate the blocking probability.

Algorithm 1: Kaufman-Roberts algorithm [20], [26]. Assume that the cell u is composed of a finite set of positions and that the set of feasible configurations has the form

$$\mathbb{M}^{\mathrm{f}} = \left\{\nu \in \mathbb{N}^u : \sum_{k\in u} \nu_k\varphi_k \leq C\right\}$$

where C and $(\varphi_k)_{k\in u}$ are given integers. Let $q(n)$ be the probability that the sum $\varphi(\nu) := \sum_{k\in u}\nu_k\varphi_k$ equals n, that is $q(n) = \sum_{\nu\in\mathbb{M}^{\mathrm{f}}:\varphi(\nu)=n}\Pi^{\mathrm{f}}(\nu)$. Then $q(\cdot)$ satisfies the following equations

$$\sum_{n=0}^{C} q(n) = 1, \quad \text{and } q(n) = \sum_{k\in u}\rho_k\varphi_k q(n-\varphi_k), \quad n = 0,\ldots,C$$

and the blocking probabilities are given by

$$b_k = 1 - \sum_{n=0}^{C-\varphi_k} q(n).$$

Moreover,

$$E[\varphi(\nu)] = \sum_{n=0}^{C} nq(n). \tag{11}$$

1) Erlang's approximation: Even though the Kaufman-Roberts algorithm 1 permits a precise and rapid evaluation of the blocking probability, we give here an Erlang's approximation which gives a more explicit expression (which will be validated in Section VII-B1). The idea is to average the admission condition in (10) over the positions of the users but not over their number. Thus we consider the approximate admission condition

$$M_u\bar{\varphi}_u \le 1$$

where M_u is the number of users in the cell u and $\bar{\varphi}_u$ is the average of φ over the cell u with respect to the traffic demand density, that is

$$\bar{\varphi}_u = \frac{1}{\rho(u)}\int_u \varphi(m)\,\rho(dm)$$

where the integral is over the cell u. Thus the blocking probability in cell u, denoted b_u, may be approximated by the classical Erlang's formula with traffic demand $\rho(u)$ for a queue with

$$\Gamma_u = \frac{1}{\bar{\varphi}_u}$$

servers, that is

$$b_u \simeq \frac{\rho(u)^{\Gamma_u}}{\Gamma_u!}\left(\sum_{n=0}^{\Gamma_u}\frac{\rho(u)^n}{n!}\right)^{-1}. \tag{12}$$

We call Γ_u the *equivalent number of servers*.

Proposition 3: For an OFDMA network operating the SFC (7) and serving a streaming traffic with bit-rate r, the equivalent number of servers equals

$$\Gamma_u = \frac{1}{r\bar{\gamma}_u} \tag{13}$$

where

$$\gamma(m) = \frac{1}{W\log_2\left(1+1/\hat{f}(m)\right)}. \tag{14}$$

and

$$\bar{\gamma}_u = \frac{1}{\rho(u)}\int_u \gamma(m)\,\rho(dm). \tag{15}$$

Proof: The expression of $\varphi(m)$ is deduced from (7), that is $\varphi(m) = r\gamma(m)$, thus $\bar{\varphi}_u = r\bar{\gamma}_u$ from which the desired result follows. ∎

V. Delay for Elastic Traffic

A. Traffic dynamics

Consider only elastic bit-rate calls whose inter-arrival times to the network $\mathbb{D}$ are i.i.d. exponential random variables with rate λ (mean $1/\lambda$). The position of each arrival is picked at random in $\mathbb{D}$ according to some distribution $Q(dm)$. Again we assume that users don't move during their calls. Each call requires to transmit a given volume of data (amount of bits that has to be sent or received), which is modeled by an exponential random variable with parameter μ that is independent of everything else. The *traffic demand density* defined by $\rho(dm) = \frac{\lambda}{\mu}Q(dm)$ is expressed in kbps [3] per surface unit. Users are served by the BS according to some bit-rate allocation policy.

The set of positions of all users served at a given time is called *configuration of users*. Let $\mathbb{M}$ be the set of all possible configurations. We denote by $\{N_t\}_{t\ge0}$ the process describing the evolution in time of the user configurations in $\mathbb{D}$ (due to arrivals and departures). It takes its values in $\mathbb{M}$. If the process $\{N_t\}_{t\ge0}$ isn't ergodic, then the mean number of users in the system grows unboundedly in the long run of the system. This situation has to be avoided; in which case we say that the system is *stable* (or equivalently ergodic).

B. Processor sharing model

Note that SFC (7) may be written as follows

$$\sum_{m\in u}\gamma(m)\,r_m \le 1$$

where $\gamma(m)$ is given by (14).

Proposition 4: Consider an OFDMA network operating the SFC (7) and serving elastic traffic. A given cell u is stable (whatever is the bit-rate allocation to users assumed work-conserving) when the traffic demand per cell satisfies

$$\rho(u) < \rho_{\mathrm{c}}(u)$$

where the *critical traffic demand* $\rho_{\mathrm{c}}(u)$ is defined by

$$\rho_{\mathrm{c}}(u) = \frac{1}{\bar{\gamma}_u}$$

where $\bar{\gamma}_u$ is given by (15). Consider now the following particular bit-rate allocation

$$r_m = \frac{1}{M_u\gamma(m)}$$

where M_u is the number of users in the cell u. At the steady state, the expected number of users in cell u, the *mean delay* and *throughput* per user are given respectively by

$$\bar{N} = \frac{\rho(u)}{\rho_c(u)-\rho(u)}, \quad \bar{T} = \frac{1}{\mu(\rho_c(u)-\rho(u))},$$
$$\bar{r} = \rho_c(u) - \rho(u).$$

Proof: By definition a bit-rate allocation is said to be *work-conserving* when

$$\sum_{m\in u} r_m\gamma(m) = 1_{\{u\neq\emptyset\}}$$

[3] The abbreviation kbps designates "Kilo-bit per second".

The results for stability and the number of users at the steady state follows from known results for multi-class processor sharing queues [11], [22]. Applying Little's formula [6] we get the desired result for the delay. Recalling that the throughput is the ratio of the date volume average and the delay, finishes the proof. ■

VI. Mixing Streaming and Elastic Traffic

We consider in the present section an OFDMA network carrying both streaming and elastic traffic *on the same bandwidth*. So interference between streaming and elastic users has to be taken into account. We aim to establish analytical formulae (or bounds) for the QoS of each type of service in this mixed scenario.

We assume that the network operates SFC. We also assume that streaming traffic has preemptive priority over elastic traffic, which has two important consequences. First, the evolution of the streaming users is *independent* of the elastic ones (in particular, the *blocking probability* of streaming calls is the same as if there were no elastic ones). Secondly, the elastic users are served with the capacity left free by the streaming users. Hence the novelty when we consider the integration is that elastic traffic observes a *random environment*.

The notations are the same as those of the previous two sections. Moreover, in order to distinguish the streaming and elastic traffic characteristics, we use the superscript $^{\mathrm{s}}$ for parameters specific to streaming traffic. In particular, we denote by ν^{s} and ν the measures representing the locations of streaming and elastic users respectively.

In this context, SFC (7) may be written as follows

$$\sum_{m\in u, m\in\nu} \gamma(m)\, r_m \leq 1 - \sum_{m\in u, m\in\nu^{\mathrm{s}}} \gamma(m)\, r_m \qquad (16)$$

where $\gamma(m)$ is given by (14). Let

$$\varphi(\nu^{\mathrm{s}}) = \sum_{m\in u, m\in\nu^{\mathrm{s}}} \gamma(m)\, r_m \qquad (17)$$

be the part of the total service capacity consumed by the streaming traffic.

In [13] the performance of the elastic traffic is bounded using the so-called *fluid* (fl) and *quasi-stationary* (qs) regimes. The *fluid regime* corresponds to the case where the elastic traffic observes a constant capacity equals to the average of the capacity left free by the streaming users. In other words, in the fluid regime we replace (16) by

$$\sum_{m\in u, m\in\nu} \gamma(m)\, r_m \leq 1 - E\left[\varphi(\nu^{\mathrm{s}})\right]$$

The *quasi-stationary regime* corresponds to the assumption that, for each given configuration ν^{s} of streaming users, the elastic traffic attains its stationary regime.

In the following three propositions we consider an OFDMA network operating SFC (7) and serving streaming and elastic traffic simultaneously. We are interested in the QoS of *elastic* traffic. First we give a stability condition.

Proposition 5: A given cell u is stable (whatever is the bit-rate allocation to elastic users assumed work-conserving) when the elastic traffic demand per cell satisfies

$$\rho(u) < \rho_{\mathrm{c}}^{\mathrm{fl}}(u)$$

where the *critical traffic demand* $\rho_{\mathrm{c}}^{\mathrm{fl}}(u)$ is defined by

$$\rho_{\mathrm{c}}^{\mathrm{fl}}(u) = \frac{1 - E\left[\varphi(\nu^{\mathrm{s}})\right]}{\bar{\gamma}_u}$$

where $\bar{\gamma}_u$ is given by (15) and $\varphi(\nu^{\mathrm{s}})$ is given by (17). The expectation in the above display may be evaluated using (11).

Proof: The cell may be viewed as a single server (serving elastic traffic) whose capacity varies randomly over time due to streaming users consuming the fraction of capacity given by (17); i.e. a $\mathrm{M/GI/1}$ queue in a random environment. The desired result then follows immediately from the properties of such queues given in [23, Proposition 1], [7, Theorem 1]. ■

We give now bounds of the QoS using the fluid (fl) and quasi-stationary (qs) regimes.

Proposition 6: We have the following inequalities between the delays

$$\bar{T}^{\mathrm{fl}} \leq \bar{T} \leq \bar{T}^{\mathrm{qs}}$$

and between the throughputs

$$\bar{r}^{\mathrm{qs}} \leq \bar{r} \leq \bar{r}^{\mathrm{fl}}$$

Proof: From [13] we deduce the following inequalities between the average number of elastic users

$$\bar{N}^{\mathrm{fl}} \leq_{\mathrm{icx}} \bar{N} \leq_{\mathrm{icx}} \bar{N}^{\mathrm{qs}}$$

where icx is the increasing convex ordering [6, p.272]. Applying Little's formula we get the desired inequalities for the delays. Recalling that the throughput is the ratio of the date volume average and the delay, finishes the proof. ■

The following proposition gives the QoS in the fluid and quasi-stationary regimes.

Proposition 7: For the fluid regime, the results of Proposition 4 apply when replacing the function $\gamma(m)$ there by $\gamma(m)\left(1 - E\left[\varphi(\nu^{\mathrm{s}})\right]\right)^{-1}$.

Consider now the quasi-stationary regime. For each given configuration ν^{s} of streaming users, the results of Proposition 4 apply by replacing the function $\gamma(m)$ there by $\gamma(m)\left(1 - \varphi(\nu^{\mathrm{s}})\right)^{-1}$.

Proof: Indeed the fluid regime may be viewed as a system serving only elastic traffic but where the function $\gamma(m)$ is replaced by $\gamma(m)\left(1 - E\left[\varphi(\nu^{\mathrm{s}})\right]\right)^{-1}$.

Recall that the quasi-stationary regime corresponds to the assumption that, for each given configuration ν^{s} of streaming users, the elastic traffic attain its stationary regime. The desired result then follows from the observation that for each given configuration ν^{s}, the system may be viewed as a system serving only elastic traffic but where the function $\gamma(m)$ is replaced by $\gamma(m)\left(1 - \varphi(\nu^{\mathrm{s}})\right)^{-1}$. ■

VII. Validation of the Sufficient Condition

We aim now to show that SFC is accurate enough, by showing that it induces an acceptable loss of capacity with respect to the reference feasibility condition FC for OFDMA networks.

A. Model specification

In order to obtain numerical values, we consider the most popular *hexagonal network model*, where the BSs are placed on a regular hexagonal grid. Let R be the radius of the disc whose area is equal to that of the hexagonal cell served by each BS, and call R the *cell radius*. In order to avoid the border effects we consider the network that is "wrapped around"; i.e., deployed on a torus comprising $4 \times 4 = 16$ cells. In order to get a discrete model, each cell is decomposed into 5 equally-thick rings around the BS. Users arrive (spatially) uniformly to the network and don't move during their calls.

We assume a propagation loss $L(r) = (Kr)^\eta$, with $\eta = 3.38$ and $K = 8667$ where r is the distance between the transmitter and the receiver. The system bandwidth equals $W = 5$MHz. BSs are equipped with omnidirectional antennas having a gain 9dBi and no loss. The BS maximal total power equals 43dBm; thus $\tilde{P} = 43 + 9 = 52$dBm when we account for antenna gain and loss. The common channel power $\hat{P}$ is the fraction $\epsilon = 0.12$ of $\tilde{P}$ and the ambient noise power $WN_0 = -103$dBm. We consider three values of the cell radius $R = 0.525$, 3 or 5km.

We consider streaming voice calls at 12.2Kbps. The corresponding SNR in real channels (including fading) is typically -16dB. Such SNR corresponds to a bit rate $W \log_2(1 + \text{SNR}) = 180$Kbps on AWGN channel. We shall consider that the voice calls require such high bit rate (in place of the usual 12.2Kbps), which permits to account for fading effect in an approximate way (for further discussion see §III-D).

We considered the above values of the system parameters since they are typical for real networks. But we made also other experiments (other propagation constants, bit rates, cell radii, powers, etc.) and observed that the basic conclusions from the numerical results presented below are sufficiently robust.

B. Results

Figure 1 shows the blocking probability per cell as a function of the traffic demand for three different admission conditions: FC, SFC1 and SFC. The curves for FC and SFC1 are obtained by long simulations (several days on a typical PC) while this of SFC is obtained either by simulations (about one day) or by Kaufman-Roberts algorithm 1, which takes only a few seconds. The following important observations can be made: *the blocking probability induced by SFC is close to this of SFC1 that in turn is sufficiently close to this of FC*.

We define the *capacity* as the traffic demand that can be served at the blocking probability equal to 0.02. It is important to bound the loss of capacity induced by the sufficient feasibility conditions relatively to the reference FC. In particular, one may consider the naive condition which consists of blocking all the users. This is clearly a sufficient condition for the feasibility of the resource allocation, nevertheless it is far from efficient.

The maximal loss of capacity of SFC compared to FC is about 10%. This loss of capacity seems to be acceptable for network operators looking for rapid network dimensioning tools. Note also that it is evaluated with respect to the reference

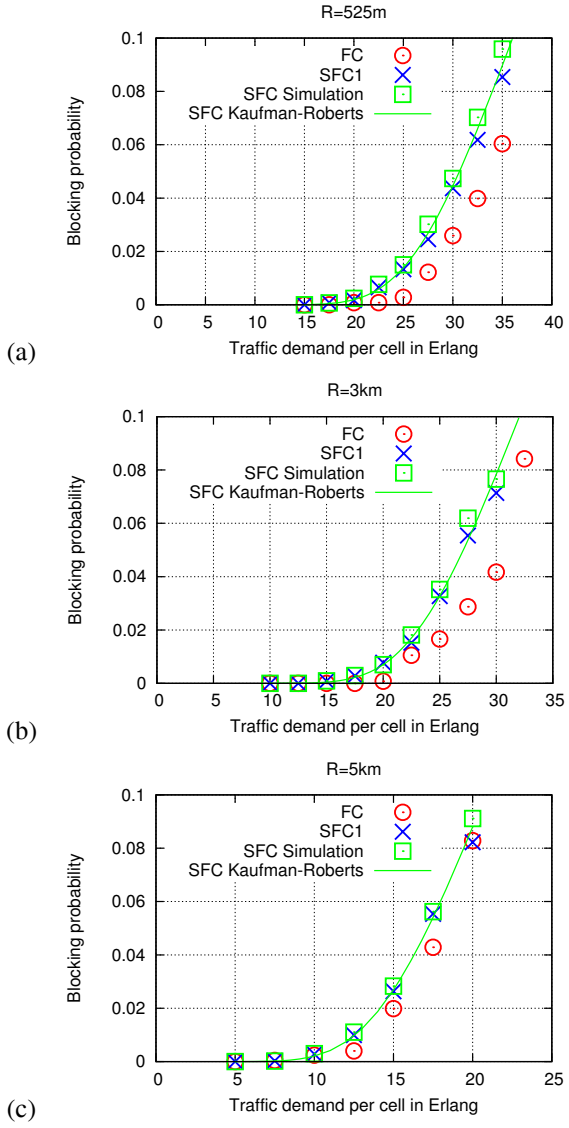

Fig. 1. Comparison of feasibility conditions performance for OFDMA downlink.

feasibility condition assuming some perfect control scheme. Observe moreover on Figure 1 that the Kaufman-Roberts algorithm result is very close to that of the simulations of SFC. We conclude that this algorithm combined with SFC gives a *rapid* and *accurate* method for the QoS evaluation of OFDMA cellular networks. Moreover note that there is *no longer need to separate coverage and capacity*. Finally, given the traffic per cell, the *blocking probability increases with cell radius*.

The Kaufman-Roberts algorithm permits to calculate the blocking probability for SFC even if the traffic demand is not spatially uniform. But in the case the loss of capacity of SFC compared to FC may depend on the traffic distribution. However, operators usually assume such distribution to be uniform in the dimensioning process.

1) Erlang's approximation: Recall that there is a single streaming class in our experiment. Figure 2 shows SFC's blocking probability obtained either by the Erlang's approximation (12) or by Kaufman-Roberts algorithm. This figure

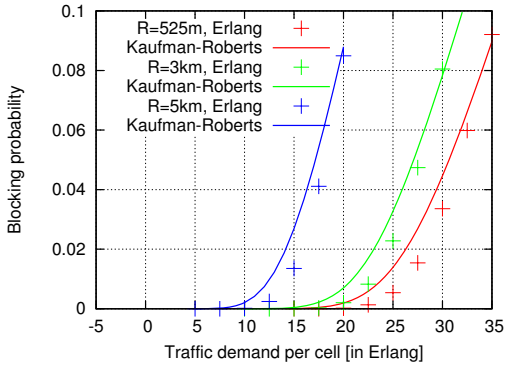

Fig. 2. Erlang's approximation versus Kaufman-Roberts algorithm for OFDMA downlink.

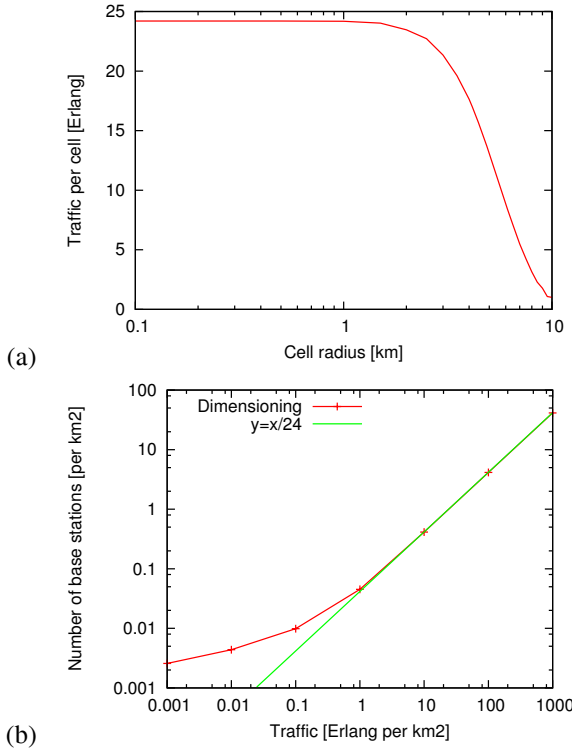

Fig. 3. Streaming traffic: (a) traffic per cell as function of cell radius; (b) dimensioning.

shows that Erlang's approximation is accurate in the considered experiment. Note however that it is necessary to use the Kaufman-Roberts algorithm (i) if the function $\varphi(m)$ in (10) is too varying with the location m or (ii) when there are multiple streaming classes. Thus the Erlang's approximation may only be used for a first rough estimate which should be refined with Kaufman-Roberts algorithm if one seeks accuracy.

VIII. Application to the Dimensioning Problem

We aim now to illustrate the proposed approach by solving the *dimensioning problem*, i.e., by evaluating what is the minimal density of BSs assuring a given quality of service (QoS) to a given traffic demand per surface unit. The model is identical to that described in Section VII-A.

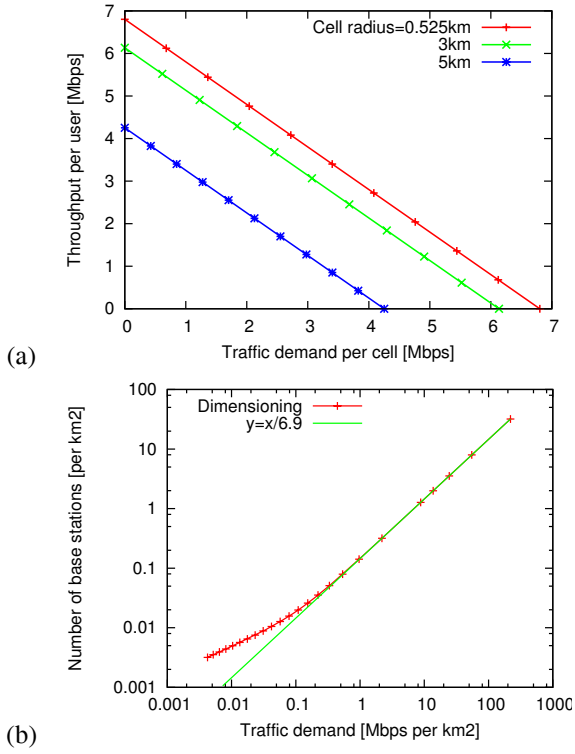

Fig. 4. Elastic traffic: (a) QoS evaluation; (b) dimensioning.

A. Streaming traffic

We consider a streaming traffic with the required bit rate 180Kbps. The blocking probability is calculated by using Kaufman-Roberts algorithm. We fix the blocking probability target to 2%. Fig. 3 (a) shows the traffic per cell as function of the cell radius R. We observe that, for $R \leq 1.5$km the traffic per cell is approximately constant (equal to 24Erlang). This corresponds to the *interference limited* regime where the noise is negligible compared to the interferences. Fig. 3 (b) shows the density of BSs $1/\left(\pi R^2\right)$ as function of the traffic demand per surface unit. As expected, we observe that in the interference limited regime (i.e. for a density of BS larger than $1/\left(\pi \times 1.5^2\right) \simeq 0.1$), *the dimensioning relation is linear* (corresponding to 24Erlang per cell).

B. Elastic traffic

We assume now that the network carries only elastic traffic. Fig. 4 (a) shows the throughput per user as function of the traffic demand per cell for different cell radii. The throughput is a decreasing linear function of the demanded traffic in the bounded stability region. Moreover, given the traffic per cell, the *throughput decreases with cell radius*. Since the delay $\bar{T}$ is simply the mean data volume $(1/\mu)$ divided by the throughput $\bar{r}$, we don't report the curves for the delay not to increase the number of figures unnecessarily. Fig. 4 (b) shows the density of BSs as function of the traffic demand per surface unit for a throughput target 384Kbps per user. Here also *the dimensioning relation is linear* when the density of BS is larger than 0.1 (corresponding to 6.9Mbps per cell).

IX. Conclusion

We have proposed an *analytical* method for the *evaluation* of QoS in the downlink of OFDMA cellular networks. Its is based on some sufficient condition for the feasibility of resource allocation. It is much faster than simulation for streaming traffic since it is based on a *multi-rate Erlang loss model*, whose blocking probabilities can be evaluated by means of the Kaufman-Roberts algorithm or more simply approximated by Erlang's formula. Our numerical experiments show that it is also accurate enough, since it induces only up to 10% loss of capacity with respect to a theoretical reference feasibility condition.

The proposed method permits also to evaluate analytically the QoS of elastic users (mean throughput and delay) by using a *multi-class processor sharing model*. Moreover, we study the performance of a network serving *simultaneously* streaming and elastic traffic. We illustrate the proposed approach by solving the *dimensioning problem*. An interesting question for future work is to evaluate the impact of the shadowing and the mobility of users.

Acknowledgement 1: The author thanks Prof. Bartłomiej Błaszczyszyn at INRIA for his encouragements and help.

References

[1] R. Agarwal, V. Majjigi, Z. Han, R. Vannithamby, and J. Cioffi, "Low complexity resource allocation with opportunistic feedback over downlink OFDMA networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1462-1472, Oct. 2008.

[2] S. Asmussen, *Applied Probability and Queues*. New York: Springer, 1987.

[3] F. Baccelli, B. Błaszczyszyn, and M. K. Karray, "Up and downlink admission/congestion control and maximal load in large homogeneous CDMA networks," *MONET*, vol. 9, no. 6, Dec. 2004.

[4] F. Baccelli, B. Błaszczyszyn, and M. K. Karray, "Blocking rates in large CDMA networks via spatial Erlang formula," in *Proc. IEEE Infocom*, 2005.

[5] F. Baccelli, B. Błaszczyszyn, and F. Tournois, "Downlink admission/congestion control and maximal load in CDMA networks," in *Proc. IEEE Infocom*, 2003.

[6] F. Baccelli and P. Brémaud, *Elements of Queueing Theory. Palm Martingale Calculus and Stochastic Recurrences*. Springer, 2003.

[7] F. Baccelli and A. A. Makowski, "Stability and bounds for single server queues in random environment," research report 536, INRIA, June 1986.

[8] B. Błaszczyszyn and M. K. Karray, "An efficient analytical method for dimensioning of CDMA cellular networks serving streaming calls," in *Proc. ValueTools*, 2008.

[9] B. Błaszczyszyn and M. K. Karray, "Dimensioning of the downlink in OFDMA cellular networks via an Erlang's loss model," in *Proc. European Wireless*, 2009.

[10] T. Bonald and A. Proutière, "Wireless downlink data channels: user performance and cell dimensioning," in *Proc. Mobicom*, Sep. 2003.

[11] J. W. Cohen, "The multiple phase service network with generalized processor sharing," *Acta Informatica*, vol. 12, pp. 245-284, 1979.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 1991.

[13] F. Delcoigne, A. Proutière, and G. Régnié, "Modelling integration of streaming and data traffic," in *Proc. ITC Specialist Seminar IP Traffic*, July 2002.

[14] S-E. Elayoubi, O. Ben Haddada, and B. Fourestie, "Performance evaluation of frequency planning schemes in OFDMA-based networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5-1, pp. 1623-1633, 2008.

[15] S. Gault, W. Hachem, and Ph. Ciblat, "Performance analysis of an OFDMA transmission system in a multi-cell environment," *IEEE Trans. Commun.*, Apr. 2007.

[16] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, pp. 1218-1230, 1997.

[17] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," in *Proc. CISS*, Mar. 2006.

[18] J. Jang and K. B. Lee. "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, Feb. 2003.

[19] M. K. Karray, "Analytic evaluation of wireless cellular networks performance by a spatial Markov process accounting for their geometry, dynamics and control schemes," Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, 2007.

[20] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, no. 10, pp. 1474-1481, 1981.

[21] K. Kim, Y. Han, and S.-L. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems," *IEEE Commun. Lett.*, vol. 9, pp. 526-528, 2005.

[22] L. Kleinrock, *Queueing Systems*. Wiley and Sons, 1976.

[23] J. Neveu, "Construction de files d'attente stationnaires," in *Lecture Notes Control Inf. Sciences*, vol. 60, pp. 31-41, 1983.

[24] M. Pischella and J.-C. Belfiore, "Achieving a frequency reuse factor of 1 in OFDMA cellular networks with cooperative communications," in *Proc. VTC Spring*, pp. 653-657, 2008.

[25] F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Trans. Commun.*, vol. 46, no. 10, Oct. 1998.

[26] J. W. Roberts, "A service system with heterogeneous user requirements," *Performance of Data Communications Systems and their Applications* (edited by G. Pujolle), 1981.

[27] A. Sampath, P. S. Kumar, and J. Holtzmann, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. IEEE PIMRC*, vol. 1, Sept. 1995.

[28] R. Serfozo, *Introduction to Stochastic Networks*. New York: Springer, 1999.

[29] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, 1948.

[30] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2726-2737, 2005.

[31] G. Wunder and C. Zhou, "Queueing analysis for the OFDMA downlink: throughput regions, delay and exponential backlog bounds," *IEEE Trans. Wireless Commun.*, Feb. 2009.

[32] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, Sep. 1995.

[33] J. Zander, "Distributed co-channel interference control in cellular radio systems," *IEEE Trans. Veh. Technol.*, vol. 41, 1992.

Mohamed Kadhem Karray received his diploma in engineering from Ecole Polytechnique and Ecole Nationale Supérieure des Télécommunications (ENST) in 1991 and 1993, respectively. He prepared a PhD thesis at ENST under the guidance of Eric Moulines and Bartek Blaszczyszyn within 2004-2007.

Since 1993 he works at France Telecom R&D (Orange Labs) in France. He co-authered together with François Baccelli and Bartek Blaszczyszyn publications in scientific journals and international conferences, and together with them he hold two patents on the load control in cellular networks. His research activities aim to evaluate the performance of communication networks. His principle tools are probability and stochastic processes, and more specifically information and queueing theories. In his recent research, he shows how to articulate the tools of these two theories to build global analytical performance evaluation methods for wireless cellular networks. The methods and tools he develops are used by Orange Operator for dimensioning its networks and for several practical studies. Dr Karray received the best paper award of the ComNet'09 conference.