# Controlling impatience in cellular networks using QoE-aware radio resource allocation

Fabrice Guillemin, Salah Eddine Elayoubi
Orange Labs, France
{fabrice.guillemin,salah.elayoubi}@orange.com

Philippe Robert, Christine Fricker, Bruno Sericola
INRIA, France
{philippe.robert,christine.fricker,bruno.sericola}@inria.fr

***Abstract*—We consider in this paper an important Quality of Experience (QoE) indicator in cellular networks that is reneging of users due to impatience. We specifically consider a cell under heavy load conditions, modeled as a multiclass Processor Sharing system, and compute the reneging probability by using a fluid limit analysis. In order to enhance the user QoE, we propose a radio resource allocation control scheme that minimizes the global reneging rates. This control scheme is based on the $\alpha$-fair scheduling framework and adapts the scheduler parameter depending on the traffic load. While the proposed scheme is simple, our results show that it achieves important performance gains.**

## I. Introduction

Impatience of users when using an Internet service has a major impact on the quality of experience, especially via an access through a cellular network with scarce radio resources. In fact, impatience is negative for the user, who is not satisfied by the quality of service offered by the network, as well as the network itself, since radio resources are consumed for transporting data which will eventually not be exploited by the end user. Impatience is due to many factors related to the performance of servers, customer devices, etc., but also to bandwidth sharing in the network. In this paper, we focus on impatience caused by the network.

From a mathematical point of view, impatience has been investigated for many years in the queuing literature. In Stanford [1], see also [2], a new version of the Erlang loss formula has been derived taking into account user impatience, resulting in the so-called Erlang-I formula, where I stands for impatience; other works call it the Erlang-A formula, A standing for abandonment [3] (see also [4] for FIFO queues with abandonment and references therein). This formula is applicable for the case of streaming-like flows where the service duration is independent of the quantity of resources obtained by the user, unlike our present case of data traffic where service duration depends on the quantity of resources obtained by the user. In [5], [6], impatience has been studied for flows with heavy tailed distributions sharing a common resource according to the processor sharing discipline.

Impatience has been modeled in [7] by using the deterministic service curves approach. The authors of that paper also reported on empirical works showing evidence for user impatience; this phenomenon may cause up to 20% of aborted traffic with respect to the total amount of carried data traffic. Data traffic at the flow level has been modeled in [8], where impatience of users is studied under overload conditions.

While impatience can be seen as a negative phenomenon, it can also be used as a lever to discourage customers when the system becomes too much overloaded. This can be achieved in cellular networks by modulating the capacity available to customers being at a certain distance of the antenna. This general idea can be applied in several manners and can be viewed as a network optimization mechanism. In this paper, we reuse the general framework of $\alpha$-fair scheduler in order to perform this control. This has the advantage of being easy to implement in realistic settings as $\alpha$-fair schedulers (and especially Proportional Fair (PF)) are widely adopted in cellular networks. This also reduces the dimension of our problem as it narrows the optimization problem to the tuning of a single parameter $\alpha$.

In order to achieve this goal, we first derive a model for reneging probabilities under a general $\alpha$-fair scheduler. In particular, we consider a heavy load regime and develop a fluid flow analysis of impatience in cellular networks. This choice is motivated by the fact that impatience is a notable phenomenon under heavy traffic conditions. In this framework, we notably establish a fixed point formulation for the computation of the reneging probability and introduce a new metric, namely QoE perturbation, expressing how much a particular flow impacts the reneging probability in the system. We then use this QoE perturbation metric to design a new radio resource management scheme that controls the parameter of the scheduler in order to reduce the global reneging in the system. For instance, recognizing that customers far from the base station degrade the global performance of the system, impatience and $\alpha$-fair scheduling can be used to discourage those customers and in some sense to perform an implicit admission control in order to optimize the use of radio resources.

The organization of this paper is as follows: In Section II, we introduce the model for describing a cell of a cellular network accounting for impatience of customers. The reneging probability is computed in Section III. A scaled version of the model is developed in Section IV. The new controlling scheme is introduced in Section V. Some concluding remarks are presented in Section VI. Some technical results are proved in Appendices.

## II. Model description

### A. Model without impatience

We consider one cell of a 3G or 4G cellular network. Because of wave propagation and interference, the capacity decreases when one moves away from the base station. The cell can be represented as a concatenation of concentric rings so that the cell can be viewed as a multi-class system where users at different positions belong to the different classes. The number of classes corresponds to the number of rings (radio conditions) $K$; class $k \in \{1, \ldots, K\}$ is characterized by a service capacity $c_k$ ($c_1 > c_2 > \ldots > c_K$) and has a weight in the total traffic demand equal to $p_k$ such that $\sum_{k=1}^{K} p_k = 1$. The state of the system is defined by $\mathbf{n} = (n_1, \ldots, n_K)$, where $n_k$ is the number of active users in region $k$. The quantity $|\mathbf{n}| = n_1 + \ldots + n_K$ is the overall number of active users in the cell.

A Round Robin scheduling scheme shares the bandwidth equally between customers: If there are $n_k$ active customers in class $k$ for $k = 1, \ldots, K$, then the server rate for a customer of class $k$ is equal to $c_k/|\mathbf{n}|$. Beyond Round Robin, certain scheduling algorithms maximize the so-called $\alpha$-fair utility function defined by

$$U = \begin{cases} \sum_{i=1}^{|\mathbf{n}|} \log(d + r_i), & \alpha = 1, \\ \sum_{i=1}^{|\mathbf{n}|} \frac{(r_i+d)^{1-\alpha}}{1-\alpha} & \alpha \neq 1. \end{cases}$$

for some $d > 0$, where $r_i$ is the mean rate for user $i$ estimated over a given time interval. The scheduling algorithm maximizing this utility function is referred to as $\alpha$-fair scheduling algorithm. The case $\alpha = 1$ is known as Proportional Fair sharing algorithm that is often implemented in base stations.

To compare the $\alpha$-fair scheduling algorithm with the Round Robin algorithm, we can introduce the concept of scheduling gain defined for class $k$ customers as the ratio of the rate allocation $r_{k,\alpha}$ under the $\alpha$-fair scheduling algorithm to the rate allocation under Round Robin. This gain depends on the number of customers in the system and is denoted by $G_k(\mathbf{n}, \alpha)$; this means that the throughput achieved by users in ring $k$ when there are $\mathbf{n}$ active users in the cell is given by $c_k G_k(\mathbf{n}, \alpha)/|\mathbf{n}|$.

Taking into account this gain, we develop in this paper a performance model for general values of the gain, i.e., when the gain depends on the detailed number of users in the cell. However, it has been shown in various situations that the gain for each zone largely depends on the total number of users $|\mathbf{n}|$ and not on the detailed positions of users and that this becomes more obvious asymptotically, i.e. when $|\mathbf{n}|$ becomes very large [9], [10]. We thus develop in Section IV a scaled version of the model in overload situations taking into consideration this asymptotic value of the gain.

We assume that users present in the cell download data, thus giving rise to data flows (typically TCP connections). In the following, we assume that flows appear according to a Poisson process with rate $\lambda$ and we set $\lambda_k = \lambda p_k$. We further assume that the volume $\sigma$ of flows is exponentially distributed with mean $\mathbb{E}(\sigma)$. In the following we set

$$\mu_k(\mathbf{n}, \alpha) = G_k(\mathbf{n}, \alpha) c_k / \mathbb{E}(\sigma),$$

which is the service rate of a class $k$ flow for a gain $G_k(\mathbf{n}, \alpha)$.

Owing to the Markovian assumptions, the row vector $\mathbf{n}(t) = (n_1(t), \ldots, n_K(t))$ describing the number of the customers in the various rings of the cell at time $t$ is a Markov process. The generator $R$ of this process is an infinite matrix with non null components given by

$$r(\mathbf{n}, \mathbf{n} + \mathbf{e}_k) = \lambda p_k \quad \text{and} \quad r(\mathbf{n}, \mathbf{n} - \mathbf{e}_k) = \frac{n_k \mu_k(\mathbf{n}, \alpha)}{|\mathbf{n}|},$$

where $\mathbf{e}_k$ is the vector with all components equal to 0 except the $k$th one equal to 1. Furthermore, we set

$$r(\mathbf{n}, \mathbf{n}) = -\lambda - \sum_{k=1}^{K} \frac{n_k \mu_k(\mathbf{n}, \alpha)}{|\mathbf{n}|}.$$

Unless all the parameters $\mu_k(\mathbf{n}, \alpha)$ do not depend on $\mathbf{n}$ and $k$, there is no product form for the steady state probability of the system. The invariant probability $\Pi$ when it exists satisfies $\Pi R = 0$ together with the normalizing condition $\sum_{\mathbf{n} \in \mathbb{N}^K} \Pi(\mathbf{n}) = 1$.

### B. Modeling impatience

We assume that users may abort their ongoing transmission if their download time is too long. We specifically assume that users in class $k$ renege if their download is not completed after a period of time exponentially distributed with parameter $\gamma_k$. In the following, we assume that $0 < \gamma_k < \underline{\mu_k}(\alpha)$ for $k = 1, \ldots, K$ and for all $\alpha > 0$, where $\underline{\mu_k}(\alpha)$ is the minimal service rate over all values of $\mathbf{n}$. This assumption means that a user is ready to wait more than if he were alone in the cell and served according to the worst scheduling discipline.

The process $(\mathbf{n}(t))$ describing the number of customers in the system is still a Markov process thanks to the exponential assumptions. The corresponding generator $Q$ is an infinite matrix whose non null transition rates are given for $k = 1, \ldots, K$ by

$$q(\mathbf{n}, \mathbf{n}+\mathbf{e}_k) = \lambda p_k, \quad \text{and} \quad q(\mathbf{n}, \mathbf{n}-\mathbf{e}_k) = \frac{n_k \mu_k(\mathbf{n}, \alpha)}{|\mathbf{n}|} + n_k \gamma_k,$$

Furthermore, we set

$$q(\mathbf{n}, \mathbf{n}) = -\lambda - \sum_{k=1}^{K} \frac{n_k \mu_k(\mathbf{n}, \alpha)}{|\mathbf{n}|} - \sum_{k=1}^{K} n_k \gamma_k.$$

We introduce, for every $n \in \mathbb{N}$, the set $\mathcal{S}_n = \{\mathbf{n} \in \mathbb{N}^K : |\mathbf{n}| = n\}$. The cardinality of set $\mathcal{S}_n$ is $|\mathcal{S}_n| = \begin{pmatrix} n+k-1 \\ n \end{pmatrix}$.

The matrix $Q$ can be decomposed as

$$Q = \begin{pmatrix} C_0 & A_0 & & & \\ B_1 & C_1 & A_1 & & \\ & B_2 & C_2 & A_2 & \\ & & \cdot & \cdot & \cdot \end{pmatrix} \tag{1}$$

with matrices $A_n$, $B_n$, and $C_n$ being defined as follows:

- The non-null entries of matrix $A_n$ are defined for $n \geq 0$ and $\mathbf{n} = (n_1, \ldots, n_K) \in \mathcal{S}_n$ by

$$a_n(\mathbf{n}, \mathbf{n} + \mathbf{e}_k) = \lambda p_k = \lambda_k$$

for $k = 1, \ldots, K$;
- The non-null entries of matrix $B_n$ are defined for $n \geq 1$ and $\mathbf{n} = (n_1, \ldots, n_K) \in \mathcal{S}_n$ by

$$b_n(\mathbf{n}, \mathbf{n} - \mathbf{e}_k) = \frac{n_k \mu_k(\mathbf{n}, \alpha)}{|\mathbf{n}|} + n_k \gamma_k$$

for $k = 1, \ldots, K$;
- The non-null entries of matrix $C_n$ are defined for $n \geq 0$ and $\mathbf{n} = (n_1, \ldots, n_K) \in \mathcal{S}_n$ by

$$c_n(\mathbf{n}, \mathbf{n}) = -\lambda - \sum_{k=1}^{K} n_k \left( \frac{\mu_k(\mathbf{n}, \alpha)}{|\mathbf{n}|} + \gamma_k \right).$$

We assume that there exist a positive constants $\mu_*$ and $\mu^*$ such that

$$\mu_* \leq \mu_k(\mathbf{n}, \alpha) \leq \mu^* \tag{2}$$

for all $k = 1, \ldots, K$, $\alpha > 0$ and $\mathbf{n} \in \mathbb{N}^K$, so that the service rate for customers does not become arbitrarily small or large.

Under the above assumption, the number of customers in the system is less than the number of customers in an $M/M/\infty$ with arrival rate $\lambda$ and service rate $\gamma = \min_{1 \leq k \leq K} \gamma_k$. This implies that the system under consideration is stable even if $\rho > 1$. There hence exists a unique invariant probability distribution given by the row vector $(\pi(\mathbf{n}), \mathbf{n} \in \mathbb{N}^K)$ which satisfies

$$\pi Q = 0 \quad \text{and} \quad \sum_{\mathbf{n} \in \mathbb{N}^K} \pi(\mathbf{n}) = 1. \tag{3}$$

## III. Probability of reneging

Let $P_k(\mathbf{n})$ be the probability that a customer of class $k$ reneges while there are $n_\ell$ customers of class $\ell = 1, \ldots, K$ in the system upon its arrival so that $\mathbf{n} = (n_1, \ldots, n_K)$. By using the memoryless property of the exponential distribution, we can easily prove the following result.

*Lemma 1:* The probabilities $P_k(\mathbf{n})$ for $\mathbf{n} \in \mathbb{N}^K$ satisfy the recurrence relations

$$P_k(\mathbf{n}) = \frac{\gamma_k}{\Lambda_k(\mathbf{n})} + \sum_{\ell=1}^{K} \frac{\lambda p_\ell}{\Lambda_k(\mathbf{n})} P_k(\mathbf{n} + \mathbf{e}_\ell)$$
$$+ \sum_{\ell=1}^{K} \frac{1}{\Lambda_k(\mathbf{n})} \left( \frac{n_\ell \mu_\ell(\mathbf{n} + \mathbf{e}_k, \alpha)}{|\mathbf{n}| + 1} + n_\ell \gamma_\ell \right) P_k(\mathbf{n} - \mathbf{e}_\ell), \tag{4}$$

where

$$\Lambda_k(\mathbf{n}) = \lambda + \sum_{\ell=1}^{K} \frac{(n_\ell + \delta_{k,\ell}) \mu_\ell(\mathbf{n} + \mathbf{e}_k, \alpha)}{|\mathbf{n}| + 1} + \sum_{\ell=1}^{K} (n_\ell + \delta_{k,\ell}) \gamma_\ell$$

with $\delta_{k,\ell}$ denoting the Kronecker symbol.

Recurrence relation (4) can be rewritten in matrix form as

$$(\mathbb{I} - M_k) \mathbf{P}_k = \gamma_k \mathbf{u}_k, \tag{5}$$

where $\mathbb{I}$ is the identity matrix, $\mathbf{P}_k$ (resp. $\mathbf{u}_k$) is the column vector with components equal to $P_k(\mathbf{n})$ (resp. $1/\Lambda_k(\mathbf{n})$), $\mathbf{n} \in \mathbb{N}^K$, and the matrix $M_k$ is given by

$$M_k = \begin{pmatrix} 0 & A_{k,0} & & & \\ B_{k,1} & 0 & A_{k,1} & & \\ & B_{k,2} & 0 & A_{k,2} & \\ & & . & . & . \end{pmatrix} \tag{6}$$

with matrices $A_{k,n}$ and $B_{k,n}$ being defined as follows:

- The non-null entries of matrix $A_{k,n}$ are defined for $n \geq 0$ and $\mathbf{n} = (n_1, \ldots, n_K) \in \mathcal{S}_n$ by

$$a_{k,n}(\mathbf{n}, \mathbf{n} + \mathbf{e}_j) = \frac{\lambda p_j}{\Lambda_k(\mathbf{n})};$$

- The non-null entries of matrix $B_{k,n}$ are defined for $n \geq 0$ and $\mathbf{n} = (n_1, \ldots, n_K) \in \mathcal{S}_n$ by

$$b_{k,n}(\mathbf{n}, \mathbf{n} - \mathbf{e}_j) = \frac{1}{\Lambda_k(\mathbf{n})} \left( \frac{n_j \mu_j(\mathbf{n} + \mathbf{e}_k, \alpha)}{|\mathbf{n}| + 1} + n_j \gamma_j \right).$$

The reneging probability $\mathcal{P}_k$ that a class $k$ customer reneges is eventually given by

$$\mathcal{P}_k = \pi . \mathbf{P}_k = \sum_{\mathbf{n} \in \mathbb{N}^K} \pi(\mathbf{n}) P_k(\mathbf{n}), \tag{7}$$

where $\pi$ is the row vector satisfying Equation (3).

In spite of the fact that the infinite norm of matrix $M_k$ is equal to 1, we can show that under Assumption (2) the solution to Equation (5) is given by

$$\mathbf{P}_k = \gamma_k \sum_{r=0}^{\infty} (M_k)^r \mathbf{u}_k. \tag{8}$$

In spite of the above explicit representation of the reneging probability, Equation (5) may be very difficult to solve since we have to deal with the infinite block matrix $M_k$, we develop in the next section an approximating model when the load $\rho > 1$. For impatience, this case is the most interesting as many customers may renege; impatience has less impact when $\rho < 1$. Note that the case $\rho \sim 1$ requires a more detailed analysis as it involves intricate heavy traffic limit theorems.

## IV. Scaling model

Throughout this section, we assume that $\rho > 1$. This implies that the number of customers in the system becomes very large. In that case, the gain $G_k(\mathbf{n}, \alpha)$ tends to a limit $G_k(\alpha)$ for all $k = 1, \ldots, K$ [9], [10]. (Some numerical examples are given in the next sections to support this assumption.) Hence, under the assumption $\rho > 1$, the service rate of customers of class $k$ is denoted by $\bar{\mu}_k(\alpha)$ which depends only on the parameter $\alpha$ of the $\alpha$-fair scheduling algorithm.

Let us introduce the notation

$$C_k(\mathbf{n}(t)) \stackrel{\text{def.}}{=} \frac{n_k(t)}{n_1(t) + \cdots + n_K(t)},$$

where $\mathbf{n}(t) = (n_1(t), \ldots, n_K(t))$ is the number of customers in the various rings of the cell at time $t$. Additionally, let us recall that due to impatience, a class $k$ customer leaves the system at rate $\gamma_k > 0$.

For $\xi > 0$, let $\mathcal{N}_\xi(\mathrm{d}t)$ denote a Poisson process on $\mathbb{R}_+$ with rate $\xi$ and $(\mathcal{N}_{\xi,i}(\mathrm{d}t))$ is an i.i.d. sequence of such processes. All Poisson processes are assumed to be independent.

Provided that $\mathbf{n}(t)$ is not 0, the process $(\mathbf{n}(t))$ can be expressed as the solution of the following SDE (Stochastic Differential Equation)

$$\mathrm{d}n_k(t) = \mathcal{N}_{\lambda_k}(\mathrm{d}t) - \mathcal{N}_{\bar{\mu}_k(\alpha)C_k(\mathbf{n}(t-))}(\mathrm{d}t) - \sum_{i=1}^{n_k(t-)} \mathcal{N}_{\gamma_k,i}(\mathrm{d}t) \quad (9)$$

with initial condition $(n_k(0))$ for $k = 1, \ldots, K$.

### A. Scaled Version

In order to qualitatively analyze the system, we consider a scaling similar to the one used in Gromoll *et al.* [11] for a processor-sharing queue with impatience with a single class of jobs but with general assumptions on the distribution of service duration and impatience. The average impatience of jobs is assumed to be of the order of a large factor $N$ as follows: the parameters $\gamma_k$ is replaced with $\gamma_k/N$. The corresponding Markov process will be denoted by $(\mathbf{n}^N(t))$. The SDE (9) then reads

$$\mathrm{d}n_k^N(t) = \mathcal{N}_{\lambda_k}(\mathrm{d}t) - \mathcal{N}_{\bar{\mu}_k(\alpha)C_k(\mathbf{n}^N(t-))}(\mathrm{d}t) - \sum_{i=1}^{n_k^N(t-)} \mathcal{N}_{\frac{\gamma_k}{N},i}(\mathrm{d}t).$$

Denote by $\bar{n}_k^N(t) = \frac{1}{N} n_k^N(Nt)$ the corresponding fluid scaling of the process. By integrating the above SDE, $\bar{n}_k^N(t)$ can be expressed as, for $1 \le k \le K$,

$$\bar{n}_k^N(t) = \bar{n}_k^N(0) + \lambda_k t - \int_0^t \frac{\bar{\mu}_k(\alpha)\bar{n}_k^N(u)}{\bar{n}_1^N(u) + \cdots + \bar{n}_K^N(u)}\,\mathrm{d}u - \gamma_k \int_0^t \bar{n}_k^N(u)\,\mathrm{d}u + M_k^N(t), \quad (10)$$

for $t < T_0^N \stackrel{\text{def.}}{=} \inf\{s \ge 0 : \bar{\mathbf{n}}^N(s) = 0\}$, where $\mathbf{M}^N(t) = (M_k^N(t))$ is a martingale whose predictable increasing process is given by

$$\left\langle M_k^N \right\rangle(t) = \frac{1}{N}\left(\lambda_k t + \int_0^t \frac{\bar{\mu}_k(\alpha)\bar{n}_k^N(u)}{\bar{n}_1^N(u) + \cdots + \bar{n}_K^N(u)}\,\mathrm{d}u + \gamma_k \int_0^t \bar{n}_k^N(u)\,\mathrm{d}u\right).$$

See Ethier and Kurtz [12] for example.

### B. Convergence results

Assume that the initial conditions are such that $(\bar{\mathbf{n}}^N(0))$ converges to a non-zero vector $\ell(0) = (\ell_k(0))$.

We first state two technical lemmas, their proofs are given in Appendix.

*Lemma 2:* The martingale $(\mathbf{M}^N(t))$ converges in distribution to 0 for the uniform convergence on compact sets when $N$ tends to infinity.

*Lemma 3:* The process $(\mathbf{n}^N(t))$ is tight.

By using the above lemmas, we can now show that the system does not empty.

*Lemma 4:* Under the condition $\rho > 1$, the hitting time of 0 by process $(\bar{\mathbf{n}}^N(t))$ converges in distribution to infinity, i.e. for any $t > 0$, $\lim_{N\to+\infty} P\left(T_0^N < t\right) = 0$.

*Proof:* Introduce the process $(\tilde{n}^N(t))$ such that $\tilde{n}^N(t) = \sum_{k=1}^K n_k^N(t)/\bar{\mu}_k(\alpha)$. We have

$$\mathrm{d}\tilde{n}^N(t) = (\rho - 1)\,\mathrm{d}t - \sum_{k=1}^N \gamma_k \frac{n_k^N(t)}{\bar{\mu}_k(\alpha)}\,\mathrm{d}t + \sum_{k=1}^K \frac{\mathrm{d}M_k^N(t)}{\bar{\mu}_k(\alpha)}.$$

Hence,

$$\mathrm{d}\tilde{n}^N(t) \ge (\rho - 1)\,\mathrm{d}t - \gamma^* \sum_{k=1}^N \frac{n_k^N(t)}{\bar{\mu}_k(\alpha)}\,\mathrm{d}t + \sum_{k=1}^K \frac{\mathrm{d}M_k^N(t)}{\bar{\mu}_k(\alpha)}.$$

where $\gamma^* = \max_{1\le k\le K} \gamma_k$. The process $(\mathbf{n}^N(t))$ is tight and the martingale term in the right hand side of the above equation vanishes when $N$ tends to infinity. It follows that for any limiting process $(\tilde{n}(t))$ of the sequence $(\tilde{n}^N(t))$ as $N \to \infty$

$$\mathrm{d}\tilde{n}(t) \ge (\rho - 1)\,\mathrm{d}t - \gamma^* \tilde{n}(t)dt.$$

This implies that $\tilde{n}(t) \ge \check{n}(t)$ for all $t \ge 0$ where $\check{n}(t)$ satisfies $\mathrm{d}\check{n}(t) = (\rho - 1)\,\mathrm{d}t - \gamma^*\check{n}(t)\,\mathrm{d}t$ with $\check{n}(0) = \tilde{n}(0)$. Since $\check{n}(t) > 0$ for all $t \ge 0$, the result follows. ∎

We can now state the main result of this section.

*Theorem 1 (Fluid Limits):* If $\mathbf{n}^N(0)/N$ converges to a non-zero limit $(\ell_k(0))$, then the process $(\mathbf{n}^N(Nt)/N)$ converges in distribution to $(\ell_k(t))$, the solution to the differential equation

$$\dot{\ell}_k(t) = \lambda_k - \frac{\bar{\mu}_k(\alpha)\ell_k(t)}{\ell_1(t) + \cdots + \ell_K(t)} - \gamma_k \ell_k(t) \quad (11)$$

with the prescribed initial condition $(\ell_k(0))$.

*Proof:* We know from Lemma 3 that the process $(\mathbf{n}^N(t))$ is tight. By using again Relation (19), it is not difficult to show that if $(\ell(t)) = (\ell_k(t))$ is a limiting point of $(\bar{n}^N(t))$ then necessarily

$$\ell_k(t) = \ell_k(0) + \lambda_k t - \int_0^t \frac{\bar{\mu}_k(\alpha)\ell_k(u)}{\ell_1(u) + \cdots + \ell_K(u)}\,\mathrm{d}u - \gamma_k \int_0^t \ell_k(u)\,\mathrm{d}u, \quad (12)$$

consequently such a limit is unique. ∎

As an easy consequence of the above result, we have the convergence of distributions.

*Corollary 1:* If $\rho > 1$ and $(n_k^N(\infty))$ denotes a random variable with the same distribution as the invariant probability of $(n_k^N(t))$ then for the convergence in distribution

$$\lim_{N\to+\infty} \left(\frac{n_k^N(\infty)}{N}\right) = (\ell_k),$$

from Equation (11), one gets that, for $1 \le k \le K$,

$$\ell_k = \frac{\lambda_k S}{\bar{\mu}_k(\alpha) + \gamma_k S}, \quad (13)$$

where $S = \ell_1 + \cdots + \ell_K$ is the unique non-negative solution to the equation

$$\sum_{k=1}^{K} \frac{\lambda_k}{\bar{\mu}_k(\alpha) + \gamma_k S} = 1. \tag{14}$$

The reneging probability for class $k$ customers is then approximated by the quantity

$$\widetilde{P}_k = \frac{\gamma_k S}{\bar{\mu}_k(\alpha) + \gamma_k S}. \tag{15}$$

The global reneging probability is thus computed by:

$$\widetilde{P} = \sum_{k=1}^{K} \frac{\lambda_k}{\Lambda} \widetilde{P}_k = \frac{S}{\Lambda} \sum_{k=1}^{K} \frac{\lambda_k \gamma_k}{\bar{\mu}_k(\alpha) + \gamma_k S} \tag{16}$$

where $\Lambda = \sum_{k=1}^{K} \lambda_k$. It is worth noting that when $\gamma_k$ does not depend on $k$ (say, $\gamma_k = \gamma_0$ for $k = 1, \ldots, K$ with some constant $\gamma_0 > 0$), the reneging probabilities do not depend on $\gamma_0$ but only on the quantity $\gamma_0 S$, which itself does not depend on $\gamma_0$ (see Equation (14)).

We can further prove the following result for second order asymptotics, whose proof is given in Appendix A.

*Theorem 2:* If the sequence $\left((n_k^N(0) - N\ell_k(0))/\sqrt{N}\right)$ converges as $N$ tends to infinity to $(\delta_k)$, then the sequence $\left((n_k^N(Nt) - N\ell_k(t))/\sqrt{N}\right)$ converges in distribution to $(Z_k(t))$, the solution to the following SDE

$$\mathrm{d}Z_k(t) = \sqrt{2\lambda_k - \dot{\ell}_k(t)}\,\mathrm{d}B_k(t)$$
$$- Z_k(t)\left(\gamma_k + \frac{\bar{\mu}_k(\alpha)}{\sum_{i=1}^{K} \ell_i(t)}\right)\mathrm{d}t + \frac{\bar{\mu}_k(\alpha)\ell_k(t)}{\sum_{i=1}^{K} \ell_i(t)} \sum_{i=1}^{K} Z_i(t)\,\mathrm{d}t,$$

such that $(Z_k(0)) = (\delta_k)$, where $(B_k(t))$ is a standard $K$-dimensional Brownian motion and $(\ell_k(t))$ is the solution to the differential equation (11) with initial condition $(\ell_k(0))$.

A numerical validation of the above results for $K = 2$ can be found in [13], where the fluid limits (15) and the Gaussian behavior described in Theorem 2 are compared against numerical results obtained by solving Equation (7) and simulation results. These numerical results show good agreement with the above theoretical approximation.

In the following, we exploit the fluid limits to investigate how the utilization of the cell can be optimized in the presence of impatient customers, those at the border of the cell being much more impatient than those at the core, thus causing more waste of resources.

## V. Controlling the scheduler parameter

### A. Illustration of the scheduler gains

We begin by illustrating the scheduler gains for different values of $\alpha$. Several papers in the literature modeled these scheduler gains. For instance, [14] proposed a general framework for computing the scheduler gain for $\alpha = 1$ and this framework has been generalized in [9] for $\alpha > 0$. In [10], the authors proposed a fast statistical method based on kernel density estimation to evaluate scheduling gains for realistic physical layer models, including the impact of advanced receivers. In particular, it has been shown in these papers that the scheduler gain mainly depends on the total number of active users in the cell, and not on their exact distribution over the different radio conditions. In this paper, we use the framework proposed in [9] and compute the scheduler gains for different values of $\alpha$.

Figure 1 illustrates the average gain of the proportional fair scheduler for a 3G system where the radio conditions of users are classified into two classes: Cell edge users characterized by an average Signal to Interference Ratio (SINR) equal to 16 dB and cell center users characterized by an average SINR equal to 24 dB. It can be observed that the scheduler gain increases with the number of users, until reaching a saturation value when the number of users is large. Furthermore, the gain in cell edge is larger than the gain in cell center; this is due to the fact that the physical throughput is saturated (limited by a maximal modulation and coding rate) and that cell center users are close to this saturation value.

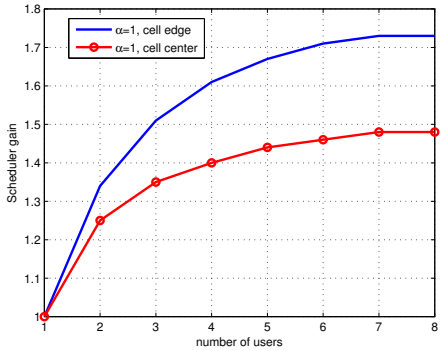



Fig. 1. Scheduler gain for cell edge and cell center users for a proportional fair scheduler.

We now move to the illustration of the impact of $\alpha$ on the scheduler gain. We illustrate in Figure 2 the case of two active users and show the scheduler gain for cell edge and cell center users for different values of $\alpha$. It can be observed that a small $\alpha$ gives a large gain ($> 1$) to cell center users and a negative gain ($< 1$) to cell edge users; as the scheduler tends to be more opportunistic[1]. For $\alpha > 1$, cell edge users are privileged as the tendency is to enhance fairness.

### B. Impact of the scheduler parameter on the reneging probabilities

In order to illustrate the impact of the scheduler on the reneging probabilities, we consider the cases of the Round Robin scheduler, the Proportional Fair scheduler ($\alpha$=1), a scheduler with $\alpha$=0.5 and a scheduler with $\alpha$=2. We consider a

[1]It is worth noting that the case $\alpha = 0$ does not correspond to a blocking of cell edge users even if their throughput is largely reduced. Indeed, the presence of fast fading randomizes the radio conditions and makes cell edge users sometimes in relatively good conditions.

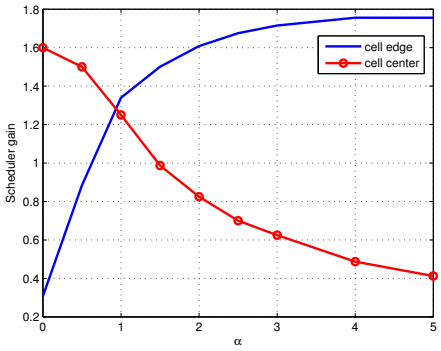



Fig. 2. Scheduler gain for cell edge and cell center users for different values of $\alpha$ when there are two users in the cell.

system with $\lambda_1 = \lambda_2$ and that is characterized by round robin service rates given by $\mu_1 = 1$ and $\mu_2 = 0.5$. The parameters $\gamma_1$ and $\gamma_2$ are taken equal so that the reneging probability does not depend on their value.

Figure 3 shows the impact of the scheduler configuration on the global reneging probability. It can be observed that all the $\alpha$-fair schedulers achieve considerable gains over round robin. On the other hand, the scheduler parameter $\alpha$ has a significant impact on the performance and has to be considered in network control. Note that, for this specific configuration, a proportional fair scheduler gives the best performance; this is not however the case in general as we will show in the next sections.

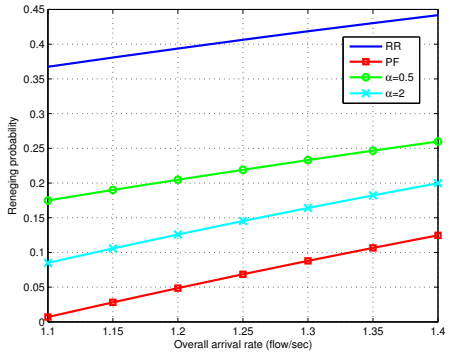



Fig. 3. Average reneging probabilities for different scheduler configurations (Round Robin, Proportional fair, $\alpha = 0.5$ and $\alpha = 2$.

### C. QoE perturbation metric and the impact of the scheduler parameter

Before moving to the optimization of $\alpha$ and to give more insights about the impact of the scheduler on the QoE, we introduce a new metric that is relative to the perturbation in QoE caused by a communication.

QoE perturbation is related to the impact of the presence of a given flow on the QoE of other users. In this paper, we define the QoE perturbation induced by a customer of class $k$ as the derivative of the global reneging rate with respect to $\lambda_k$ (the shadow cost of a class $k$ user). By using the approximation of the previous section, we then define the QoE perturbation function for class $k$ users as follows:

$$\tilde{\Gamma}_k(\lambda_1, ..., \lambda_K) = \partial \tilde{R} / \partial \lambda_k,$$

where the global impatience rate is defined by the following formula when all the parameters $\gamma_k$ are equal to some $\gamma_0$

$$\tilde{R}(\lambda_1, ..., \lambda_K) = \sum_{j=1}^{K} \lambda_j \widetilde{\mathcal{P}}_j(\lambda_1, ..., \lambda_K) = \gamma_0 S(\lambda_1, ..., \lambda_K).$$

In the general case of $K$ radio conditions, it is sufficient to derive the fixed point equation (14) in order to obtain the QoE perturbation metric:

$$\begin{aligned} \tilde{\Gamma}_k(\lambda_1, ..., \lambda_K) &= \frac{\partial \tilde{R}}{\partial \lambda_k} \\ &= \left( (\bar{\mu}_k + \gamma_0 S) \sum_{j=1}^{K} \frac{\lambda_j}{(\bar{\mu}_j + \gamma_0 S)^2} \right)^{-1} \end{aligned} \quad (17)$$

Note that this QoE perturbation metric depends only on the value $\gamma_0 S$ that is independent from the impatience rate $\gamma_0$ (see the fixed point equation (14)).

Figures 4 show the QoE perturbation metrics for different schedulers ($\alpha$ =0.5, 1 and 2) for cell edge and cell center users, respectively. By privileging cell edge users, a scheduler with $\alpha < 1$ increases the impact of cell edge users on the overall QoE, while a more opportunistic scheduler ($\alpha > 1$) makes cell center users contribute more to QoE perturbation.

### D. Optimal scheduler configuration

We now propose a framework for controlling the impatience of users. The objective is to tune $\alpha$ in order to minimize the global reneging rate $\tilde{R}(\lambda_1, ..., \lambda_K, \alpha) = \gamma_0 S(\lambda_1, ..., \lambda_K, \alpha)$.

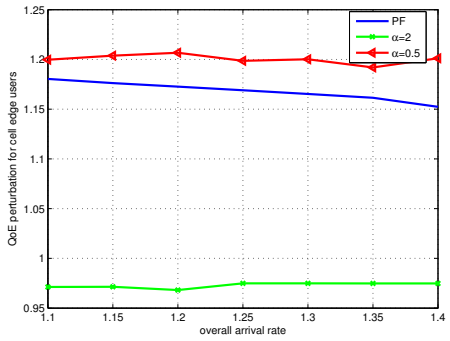
For each network and traffic configuration, the objective is to find $\alpha^*$ that satisfies:

$$\alpha^* = \text{argmin}_\alpha S(\lambda_1, ..., \lambda_K, \alpha) \quad (18)$$

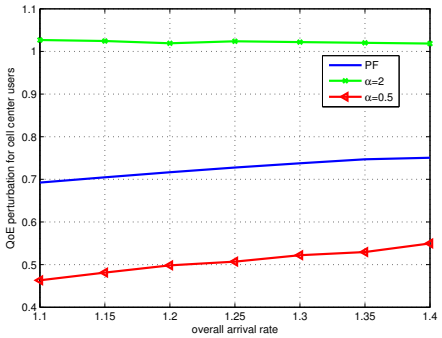
Figure 5 shows the optimal scheduler parameter that minimizes the overall reneging rate, for three network configurations that correspond to the balance between good and bad radio conditions in the cell. It can be observed that, when the weights of users with good and bad radio conditions are equivalent, the PF scheduler is the optimal one as it achieves the lowest reneging rate. In that figure, $\lambda_c$ (resp. $\lambda_e$) denotes the arrival rate of users at the center (resp. the edge) of the cell.

However, when radio conditions in the cell tend to be biased towards low SINRs ($\lambda_c/\lambda_e < 1$), a large $\alpha$ is optimal as it privileges cell edges and vice versa. The gain of this optimal scheme compared to a classical PF scheduler is illustrated in Figure 6 that shows a reduction in the reneging rate of up to 50%.

To conclude this section, let us mention that we have performed simulations for underload conditions. We have




(a) Cell edge.




(b) Cell center.

Fig. 4. QoE perturbation metric for users for different scheduler parameters (Proportional fair, $\alpha = 0.5$ and $\alpha = 2$).

observed that the modulation of the parameter $\alpha$ has much less impact on the reneging probabilities than in the case of overload. This is why the proposed scheme is well adapted to manage the cell under overload conditions.

## VI. Conclusion

By using a fluid limit approximation, we have studied in this paper the reneging probability of customers sharing the radio resources of a cell in a cellular network under different scheduling disciplines. We show that under heavy load conditions the reneging probability can easily be derived by using a simple fixed point equation.

In addition, we have introduced a QoE perturbation metric corresponding to the impact that a particular communication class has on the QoE of other users. We then used this analytical framework in order to devise a radio resource management scheme that minimizes reneging by controlling the scheduler parameter and show that important performance gains can be achieved. By using a factor impacting QoE, we can thus perform some kind of admission control without using explicit signalling between the user and the network.

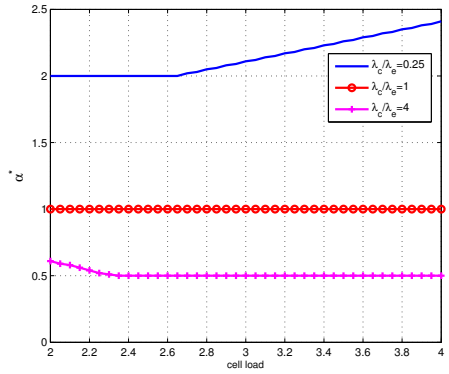



Fig. 5. Optimal scheduler parameter for different traffic loads.

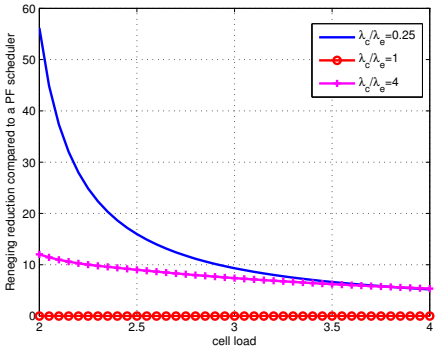



Fig. 6. Gain on the reneging rate when implementing the optimal scheduler parameter compared to a classical PF scheduler.

The choice of the optimal fairness parameter moreover does not depend on the impatience rate if this last factor is identical for all users, which seems to be a reasonable assumption. Concerning statistical assumptions, it should be noted that, due to the processor-sharing discipline the hypothesis of exponential service distribution has some impact on the results obtained in the paper. See Jean-Marie and Robert [15] for example. Further investigations are required in this domain.

## Appendix

### First Order

*Proof of Lemma 2:* Let

$$S^N(t) = \sup_{0 \le s \le t} \left( \sum_{k=1}^{K} \frac{\bar{n}_k^N(s)}{\bar{\mu}_k(\alpha)} \right)$$

then, with Relation (10), one obtains

$$S^N(t) \le S^N(0) + (\rho + 1)t + \gamma^* \int_0^t S^N(u)\, \mathrm{d}u + \sup_{0 \le s \le t} \left( \sum_{k=1}^{K} \frac{M_k^N(s)}{\bar{\mu}_k(\alpha)} \right),$$

where $\gamma^* = \max_{1\leq k\leq K}\gamma_k$. Doob's Inequality subsequently gives the relation

$$\mathbb{E}(S^N(t)) \leq \mathbb{E}(S^N(0)) + (\rho+1)t + \gamma^*\int_0^t \mathbb{E}(S^N(u))\,\mathrm{d}u + \frac{2}{N}\left(\left(\sum_{k=1}^K\left(\frac{\lambda_k}{\bar\mu_k(\alpha)^2}+\sup_k\frac{1}{\bar\mu_k(\alpha)}\right)\right)t + \gamma^*\int_0^t \mathbb{E}(S^N(u))\,\mathrm{d}u\right).$$

From Gronwall's Inequality, one derives the relation

$$\mathbb{E}(S^N(t)) \leq \left(\mathbb{E}(S^N(0)) + \frac{K_1^N}{K_2^N}\right)\exp\left(K_2 t\right), \qquad (19)$$

with

$$\begin{aligned} K_1^N &= \rho + 1 + \frac{2}{N}\sum_{k=1}^K\left(\frac{\lambda_k}{\bar\mu_k(\alpha)^2}+\sup_k\frac{1}{\bar\mu_k(\alpha)}\right), \\ K_2^N &= \frac{N+2}{N}\gamma^*. \end{aligned}$$

In particular, by using again Doob's Inequality and the expression for $\left\langle M_k^N\right\rangle(t)$, for $\eta>0$,

$$\mathbb{P}\left(\sup_{0\leq s\leq t}|M_k^N(s)|\geq\eta\right) \leq \frac{1}{N\eta^2}\left(\left(\lambda_k+\bar\mu_k(\alpha)\right)t + \gamma^*\bar\mu_k(\alpha)\int_0^t\mathbb{E}(S^N(u))\,\mathrm{d}u\right).$$

The uniform boundedness in $N$ implied by Relation (19) shows therefore that the sequence of martingales $(\mathbf{M}^N(t))$ converges in distribution to 0 for the uniform convergence on compact sets. ∎

*Proof of Lemma 3:* The modulus of continuity of process $(\bar n_k^N(t))$ is defined as

$$\omega_{\bar n_k^N}(\delta) = \sup_{\substack{0\leq s,s'\leq t\\ |s-s'|\leq\delta}}\left|\bar n_k^N(s) - \bar n_k^N(s')\right|.$$

For $\eta>0$, choose $\delta$ such that $\eta > 4\delta\max_k(\lambda_k,\bar\mu_k(\alpha))$

$$\mathbb{P}(\omega_{\bar n_k^N}(\delta)\geq\eta) \leq \mathbb{P}\left(\sup_{s,s'\leq t}|M_k^N(s)-M_k^N(s')|\geq\eta/4\right) + \mathbb{P}\left(\sup_{\substack{0\leq s\leq s'\leq t\\ |s-s'|\leq\delta}}\gamma^*\int_s^{s'}\bar n_k^N(u)\,\mathrm{d}u\geq\eta/4\right).$$

Because of the convergence of the martingales, the first term of the right-hand side of this inequality converges to 0 when $N$ gets large. The last term of the right-hand side can be bounded as follows

$$\mathbb{P}\left(\sup_{\substack{0\leq s\leq s'\leq t\\ |s-s'|\leq\delta}}\gamma^*\int_s^{s'}\bar n_k^N(u)\,\mathrm{d}u\geq\eta/4\right) \leq \mathbb{P}\left(S^N(t)\geq\frac{\eta}{4\gamma^*\bar\mu_1\delta}\right)\leq\frac{4\gamma^*\bar\mu_1\delta}{\eta}\mathbb{E}(S^N(t))$$

The sequence $\mathbb{E}(S^N(t))$ being bounded by Relation (19) this expression can be made arbitrarily small with a convenient $\delta$. One concludes that the sequence $(n_k^N(t))$ is tight. ∎

SECOND ORDER

Define

$$\widehat n_k^N(t) = \frac{n_k^N(Nt) - N\ell_k(t)}{\sqrt N},$$

The stochastic differential equations (10) and ordinary differential equation (12) give the relation

$$\widehat n_k^N(t) = \widehat n_k^N(0) - \gamma_k\int_0^t\widehat n_k^N(u)\,\mathrm{d}u + \widehat M_k^N(t) - \sqrt N\bar\mu_k(\alpha)\int_0^t\left(\frac{n_k^N(Nu)}{|\mathbf{n}^N(Nu)|} - \frac{\ell_k(u)}{\sum_{j=1}^K\ell_j(u)}\right)\mathrm{d}u \qquad (20)$$

where the process $(\widehat M_k^N(t))$ is the martingale $(M_k^N(t)/\sqrt N)$ whose predictable increasing process is given by

$$\left\langle\widehat M_k^N\right\rangle(t) = \lambda_k t + \int_0^t\frac{\bar\mu_k(\alpha)\bar n_k^N(u)}{\sum_{j=1}^K\bar n^N(u)}\,\mathrm{d}u + \gamma_k\int_0^t\bar n_k^N(u)\,\mathrm{d}u.$$

### A. Convergence of the martingale

The convergence in distribution of $(\bar n_k^N(t))$ to $(\ell_k(t))$ shows that the processes $\left(\left\langle\hat M_k^N\right\rangle(t)\right)$ converges in distribution to

$$\lambda_k t + \int_0^t\frac{\bar\mu_k(\alpha)\ell_k}{\sum_{j=1}^K\ell_j(u)}\,\mathrm{d}u + \gamma_k\int_0^t\ell_k(u)\,\mathrm{d}u = 2\lambda_k t + \ell_k(0) - \ell_k(t).$$

For $1\leq k\neq k'\leq K$, one has $\left\langle\widehat M_k^N,\widehat M_{k'}^N\right\rangle(t)=0$, consequently Theorem 1.4, page 339 of Ethier and Kurtz [12] shows that, for the convergence in distribution,

$$\lim_{N\to+\infty}\left(\widehat M_k^N(t), 1\leq k\leq K\right) = \left(\int_0^t\sqrt{2\lambda_k - \dot\ell_k(u)}\,\mathrm{d}B_k(u), 1\leq k\leq K\right)$$

where $(B_k(t))$, $1\leq k\leq K$ are independent standard Brownian motions.

### B. Tightness

Fix some $T>0$. The integrand of the last term of the right-hand side of Equation (20) is

$$\begin{aligned} &\frac{n_k^N(Nu)}{|\mathbf{n}^N(t)|} - \frac{\ell_k(u)}{\sum_{j=1}^K\ell_j(u)} = \\ &\frac{\widehat n_k^N(Nu)/\sqrt N + \ell_k(u)}{\sum_{k=1}^K\widehat n_k^N(Nu)/\sqrt N + \sum_{j=1}^K\ell_j(u)} - \frac{\ell_k(u)}{\sum_{j=1}^K\ell_j(u)} \\ &= \frac{1}{\sqrt N}\frac{\left(\widehat n_k^N(u)\sum_{j=1}^K\ell_k(u) - \ell_j(u)\sum_{j=1}^K\widehat n_j^N(u)\right)}{\sum_{j=1}^K\overline{n}_j^N(u)\sum_{j=1}^K\ell_j(u)}. \end{aligned}$$

Equation (20) becomes

$$\widehat{n}_k^N(t) = \widehat{n}_k^N(0) - \gamma_k \int_0^t \widehat{n}_k^N(u)\ \mathrm{d}u + \widehat{M}_k^N(t)$$
$$- \bar{\mu}_k(\alpha) \int_0^t \frac{\left(\sum_{j=1}^K \widehat{n}_k^N(u)\ell_j(u) - \ell_k(u)\widehat{n}_j^N(u)\right)}{\sum_{j=1}^K \overline{n}_j^N(u) \sum_{j=1}^K \ell_j(u)}\,\mathrm{d}u. \quad (21)$$

The convergence results obtained for the fluid limit show that the sequence of processes $\left(\sum_{k=1}^K \overline{n}_k^N(t)\right)$ converges in distribution to the process $\left(\sum_{k=1}^K \ell_k(t)\right)$. The assumptions on $\rho$ and on the initial state imply that $t \mapsto \sum_{k=1}^K \ell_k(t)$ is lower bounded by some $\alpha > 0$. In particular for any $\varepsilon > 0$, there exists some $N_0$ such that if $N \geq N_0$ and

$$\mathcal{A}_N \stackrel{\text{def.}}{=} \left\{ \inf_{0\leq t\leq T} \sum_{k=1}^K \overline{n}_k^N(t) \geq \frac{\alpha}{2} \right\},$$

then $\mathbb{P}(\mathcal{A}_N^{\mathrm{c}}) \geq 1 - \varepsilon$.

If $a = (a_k) \in \mathbb{R}^K$ and $(H(t)) = (H_k(t))$ is some locally bounded function with values in $\mathbb{R}^K$, one denotes

$$a^* = \max_{1\leq k\leq K} |a_k| \text{ and } H_t^* = \sup_{0\leq s\leq t} |H(s)^*|\,.$$

On the set $\mathcal{A}_N$, one has, for $0 \leq t \leq T$,

$$\widehat{n}_t^{N*} \leq \widehat{n}_0^{N*} + \widehat{M}_T^{N*} + \left(\gamma^* + \frac{8K\bar{\mu}(\alpha)^*\ell_T^*}{\alpha^2}\right) \int_0^t \widehat{n}_s^{N*}\ \mathrm{d}s,$$

Gronwall's Inequality gives

$$\widehat{n}_t^{N*} \leq \left(\widehat{n}_0^{N*} + \widehat{M}_T^{N*}\right) \exp\left(\left(\gamma^* + \frac{8K\bar{\mu}(\alpha)^*\ell_T^*}{\alpha^2}\right) t\right),$$

since $\mathbb{P}(\mathcal{A}_N) \leq \varepsilon$ and the sequence $(\widehat{M}_T^{N*})$ is converging in distribution, one gets that there exists some $C_1 > 0$ such that

$$\mathbb{P}\left(\widehat{n}_T^{N*} \geq C_1\right) \leq 2\varepsilon. \quad (22)$$

Let

$$W_H(\delta) = \max_{1\leq k\leq K} \sup_{\substack{0\leq s,t\leq T \\ |s-t|\leq\delta}} |H_k(s) - H_k(t)|.$$

From Relation (21), one gets that, for $\eta > 0$ and $\delta > 0$,

$$\mathbb{P}\left(W_{\widehat{n}^N}(\delta) > \eta\right) \leq$$
$$\mathbb{P}\left(\gamma^*\delta\widehat{n}_T^{N*} \geq \frac{\eta}{3}\right) + \mathbb{P}\left(W_{\widehat{M}^N}(\delta) > \frac{\eta}{3}\right)$$
$$+ \mathbb{P}\left(\delta\left(\gamma^* + \frac{8K\bar{\mu}(\alpha)^*\ell_T^*}{\alpha^2}\right)\widehat{n}_T^{N*} \geq \frac{\eta}{3}\right).$$

By tightness of the sequence of processes $(\widehat{M}_k^N(t))$, one can find a $\delta_0 > 0$ and $N_1 \geq N_0$ such that if $N \geq N_1$ and $\delta < \delta_0$, then

$$\mathbb{P}\left(W_{\widehat{M}^N}(\delta) > \frac{\eta}{3}\right) \leq \varepsilon.$$

If one takes

$$\delta_1 < \min\left(\delta_0, \frac{\eta}{3C_1\gamma^*}, \frac{\eta}{3C_1}\left(\gamma^* + \frac{8K\bar{\mu}(\alpha)^*\ell_T^*}{\alpha^2}\right)^{-1}\right),$$

from Relation (22), one gets that, for $\delta < \delta_1$ and $N \geq N_1$,

$$\mathbb{P}\left(W_{\widehat{n}^N}(\delta) > \eta\right) \leq 5\varepsilon.$$

From Theorem 15.5, page 127 of Billingsley [16], one gets that the sequence of processes $\left(\widehat{n}_k^N(t), 1 \leq k \leq K\right)$ is tight and that any of its limiting points is a continuous process.

### C. Proof of the theorem

If $(Z(t)) = (Z_k(t), 1 \leq k \leq K)$ is a limiting point of the sequence

$$\left(\widehat{n}_k^N(t), 1 \leq k \leq K\right),$$

Relation (21) gives that it must satisfy the relation

$$Z_k(t) = \delta_k$$
$$- \gamma_k \int_0^t Z_k(u)\ \mathrm{d}u + \int_0^t \sqrt{2\lambda_k - \dot{\ell}_k(u)}\ \mathrm{d}B_k(u)$$
$$- \int_0^t \frac{\bar{\mu}_k(\alpha)\left(\sum_{j=1}^K Z_k(u)\ell_j(u) - \ell_k(u)Z_j(u)\right)}{\left(\sum_{j=1}^K \ell_j(u)\right)^2}\,\mathrm{d}u.$$

The uniqueness of such a process $(Z(t))$ (as the solution of a classical SDE) concludes the proof of the result.

## REFERENCES


[1] R. Stanford, "Reneging phenomena in single channel queues," *Mathematics of Operations Research*, vol. 4, no. 2, pp. 162–178, 1979.

[2] ——, "On queues with impatience," *Advances in Applied Probability*, vol. 22, no. 3, pp. 768–769, 1990.

[3] F. Baccelli and G. Hebuterne, "On queues with impatient customers," *Performance*, vol. 81, pp. 159–179, 1981.

[4] O. B. Jennings and A. L. Puha, "Fluid limits for overloaded multiclass FIFO single-server queues with general abandonment," *Stochastic Systems*, vol. 3, no. 1, pp. 262–321, 2013.

[5] J. Boyer, F. Guillemin, P. Robert, and B. Zwart, "Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks," in *Proc. INFOCOM*, 2003.

[6] F. Guillemin, P. Robert, and B. Zwart, "Tail asymptotics for processor-sharing queues," *Advances in Applied Probability*, pp. 525–543, 2004.

[7] S.-C. Yang and G. de Veciana, "Bandwidth sharing: The role of user impatience," in *IEEE Globecom*, 2001.

[8] T. Bonald and J. Roberts, "Congestion at flow level and the impact of user behaviour," *Computer networks*, vol. 42, pp. 521–536, 2003.

[9] R. Combes, Z. Altman, and E. Altman, "Scheduling gain for frequency-selective rayleigh-fading channels with application to self-organizing packet scheduling," *Perform. Evaluation*, vol. 68, no. 8, pp. 690–709, 2011.

[10] R. Combes, S. Elayoubi, and Z. Altman, "Cross-layer analysis of scheduling gains: Application to lmmse receivers in frequency-selective rayleigh-fading channels," in *Proceedings of IEEE WiOpt 2011*, 2011.

[11] C. Gromoll, P. Robert, and B. Zwart, "Fluid limits for processor sharing queues with impatience," *Mathematics of Operations Research*, vol. 33, no. 2, pp. 375–402, May 2008.

[12] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. WILEY, 1986.

[13] S. E. Elayoubi, C. Fricker, F. Guillemin, P. Robert, and B. Sériola, "Impatience in mobile networks and its application to data pricing," in *IEEE ICC 2015 Communications QoS, Reliability and Modeling Symposium*. IEEE Communications Society, Jun. 2015.

[14] F. Bergren and R. Jantti, "Asymptotically fair transmission scheduling over fading channels," *IEEE transactions on wireless communications*, vol. 3, no. 8, pp. 326–336, 2004.

[15] A. Jean-Marie and P. Robert, "On the transient behavior of some single server queues," *Queueing Systems, Theory and Applications*, vol. 17, pp. 129–136, 1994.

[16] P. Billingsley, *Convergence of Probability Measures*. Wiley, New York, 1968.