Quality of Real-Time Streaming in Wireless Cellular Networks: Stochastic Analytic Evaluation

Bartlomiej Blaszczyszyn, Miodrag Jovanovic, Mohamed Kadhem Karray

Quality of Real-Time Streaming in Wireless Cellular Networks

Stochastic Analytic Evaluation

Bartłomiej Błaszczyszyn, Miodrag Jovanovic
and Mohamed Kadhem Karray, *Member, IEEE,*

Abstract

We present a new stochastic service model with service capacity sharing and interruptions, meant to be useful for the performance evaluation and dimensioning of wireless cellular networks offering real-time streaming, like e.g. mobile TV. Our general model takes into account Markovian, multi-class process of call arrivals, arbitrary streaming time distribution, and allows for a general service (outage) policy saying which users are temporarily denied the service due to insufficient service capacity. Using Palm theory formalism, we develop expressions for several important characteristics of this model, including mean time spent in outage and mean number of outage incidents for a typical user of a given class. We also propose some natural class of *least-effort-served-first* service policies, for which the aforementioned expressions can be efficiently evaluated on the basis of the Fourier analysis of Poisson process. Last but not least, we show how our model can be used to analyse the quality of real-time streaming in 3GPP Long Term Evolution (LTE) cellular networks. We identify and evaluate an optimal and a fair service policy, the latter being suggested by LTE implementations, as well as propose some intermediate policies which allow to solve the optimality/fairness trade-off caused by unequal user radio-channel conditions.

Index Terms

real-time streaming, mobile TV, LTE, outage times, interruptions, processor-sharing model, Poisson process.

1 INTRODUCTION

Wireless cellular networks offer nowadays possibility to watch TV on mobile devices, which is an example of a real-time content streaming. This type of traffic demand is expected to increase significantly in future. In order to cope with this process, network operators need to implement in their dimensioning tools efficient methods allowing to predict the quality of this type of service, related in particular to the number and duration of *outage incidents* — (hopefully short) periods when the network cannot deliver to a given user in real-time the requested content (of the required quality). In this paper we propose a stochastic model allowing for efficient,

B. Błaszczyszyn is with Inria-ENS, 23 Avenue d'Italie, 75214 Paris, France; email: Bartek.Blaszczyszyn@ens.fr

M. Jovanovic and M. K. Karray are with Orange Labs, 38/40 rue Général Leclerc, 92794 Issy-les-Moulineaux, France; email: {miodrag.jovanovic, mohamed.karray}@orange.com

This paper reports the results of the research undertaken under CRE-CIFRE thesis co-advising agreement between Inria and Orange Labs.

analytic evaluation of these metrics of the real-time streaming. Besides the traffic demand, our model can be specified to take into account all main characteristics of a given wireless cellular technology, as well as the spatial distribution of the signal-to-interference-and-noise ratio (SINR) resulting from the network geometry and the topography of its deployment area.

Our general model is a new stochastic service model with service capacity sharing and interruptions. It comprises Markovian, multi-class process of call arrivals and their independent, arbitrarily distributed, sojourn (streaming) times. These calls are served by a server whose service capacity is limited. Depending on numbers of calls of different classes present in the system, the server may not be able to serve some classes of users. If such a congestion occurs, these classes are temporarily denied the service, until the next call arrival or departure, when the situation is reevaluated. These service denial periods, called outage periods, do not alter the call sojourn times in the system. Our model allows for a very general service (outage) policy saying which classes of users are temporarily denied the service due to insufficient service capacity.

Using the formalism of point processes and their Palm theory, often used in the modern approach to stochastic networking [1], we evaluate the following key performance characteristics of the new service model: the intensity of outage incidents, the mean inter-outage times and the outage durations of a given class, seen from the server perspective, as well as the probability of outage at the arrival epoch, mean total time in outage and mean number of outage incidents experienced by a typical user of a given class. The expressions developed for these characteristics involve only stationary probabilities of the (free) traffic demand process, which in our case is a vector of independent Poisson random variables. Recall that such a representation is possible e.g. for the well known Erlang-B formula, giving the blocking probability in the classical (possibly multi-class) Erlang's loss model. Indeed, our model can be seen as an extension of the classical loss model, where the losses (i.e., service denials) are not definitive for a given call, but only temporal — having the form of outage periods.

Of particular interest in the modeling of heterogeneous traffic service are the so-called *multi-rate Erlang resource constraints*, where each user of a given class requires some given fixed fraction of the normalized capacity of the server, making the simultaneous service of the whole configuration of users feasible if and only if the total (sum) required capacity fraction is not larger than one. In the case of loss and processor sharing models, this assumption has already proved to lead to efficient model evaluation methods, cf. e.g. Kaufman-Roberts algorithm [2, 3]. Assuming these specific multi-rate Erlang resource constraints in our general model with service interruptions has the following important consequences:

- It allows us to propose some natural parametric class of *least-effort-served-first* service policies, which assign service to users in order of their increasing capacity demand, until the full capacity (possibly with some margin) is reached. The capacity margin may be used to offer some "lower quality" service to users temporarily in outage, thus realizing a trad-off between the optimality and fairness.
- For these policies, we use a method described in [4], based on the Bromwich contour inversion of the Fourier transform, to evaluate very explicitly the general expressions for key performance characteristics of the model, involving the stationary probabilities of the Poisson traffic demand process.
- Fundamental and practical limitations observed in the radio part of wireless cellular networks at the link layer and the multiple access layer are related to the scarceness of resources, such as signal power, frequency bandwidth, time and space. We show how these limitations can naturally be cast into the multi-rate Erlang

resource constraint model. This let us propose a complete model, compliant with the 3GPP Long Term Evolution (LTE) cellular network specification, allowing for the analytic evaluation of the performance of the real-time streaming services.

- We use a multi-class model (with a multi-rate Erlang condition) not only for the representation of users with different required streaming bit-rates or times, but primarily in order to represent users with different radio channel conditions. This allows us to evaluate the *quality of streaming in function of the user location in the network characterized by the values of the SINR it experiences*. This is very different from many prior works, in which one studies population-averaged quality of service characteristics.

Let us now recollect a few *related works* on the performance evaluation of cellular networks. In early 80's, wireless cellular networks were carrying essentially voice calls, which require constant bit-rates (CBR) and are subject to admission control policies with blocking (at the arrival epoch) to guarantee these rates for calls already in service. An important amount of work has been done to propose efficient call admission conditions [5–7]. Policies with admission conditions in the multi-Erlang form have been considered e.g. in [8–10].

Progressively, cellular networks started carrying also calls with variable bit-rates (VBR), used to transmit data files. The available resources are (fairly) shared between such calls and when the traffic demand increases, the file transfer delays increase as well, but (in principle) no call is ever blocked. These delays may be evaluated analytically using multi-rate Erlang resource constraint in conjunction with multi-class processor sharing models; cf e.g. [10, 11].

Recently, users may access multimedia streaming services through their mobile devices [12]. They are provided via CBR connections, essentially without admission control, but they tolerate temporary interruptions, when network congestions occur. One may distinguish two types of streaming traffic. In *real-time streaming* (as e.g. in mobile TV), considered in this paper, the portions of the streaming content emitted during the time when the transmission to a given user is interrupted (is in outage) are definitely lost for him (unless a "secondary", lower-rate streaming is provided during these periods). In *non-real-time streaming* (like e.g., video-on-demand, YouTube, Dailymotion, etc), a user starts playing back the requested multimedia content after some initial delay, required to deliver and buffer on the user device some initial portion of it. If further transmission is interrupted for some time making the user buffer content drop to zero (buffer starvation) then the play-back is stopped until some new required portion of the content is delivered. Several papers study the effect of the variability of the wireless channel on the performance of a single streaming call; see for e.g. [13], [14]. In [15] VBR transmissions and real-time streaming are considered jointly in some analytical model, however the number and duration of outage periods are not evaluated. In [16] the tradeoff between the start-up delay and the probability of buffer starvation is analyzed in a Markovian queuing framework for non-real-time streaming.

The remaining part of this paper is organized as follows. In Section 2 we present our general model. General analytical results are formulated and proved in Section 3. In Section 4 we show how the fundamental and practical limitations observed in the radio part of wireless cellular networks can be cast into the multi-rate Erlang resource constraint condition to be used with our general model. Finally, in Section 5 we present some numerical results regarding the quality of real-time streaming, obtained using our model specified to be compliant with the LTE cellular network specification.

2 A STOCHASTIC SERVICE MODEL WITH CAPACITY SHARING AND INTERRUPTIONS

2.1 Traffic demand

Consider $J \geq 1$ classes of users identified with calls. Classes of users/calls are characterized e.g. by different requested streaming bit-rates, wireless channel conditions, mean streaming times. We assume that users of class $k \in \{1, \ldots, J\}$ arrive in time according to a Poisson process $N_k = \{T_n^k : n\}$ with intensity $\lambda_k > 0$ and stay in the system for independent requested streaming times W_n^k having some general distribution with mean $1/\mu_k < \infty$. All the results presented in this paper do not depend on the particular choice of the streaming time distributions. [1] Denote by $\tilde{N}_k = \{(T_n^k, W_n^k) : n\}$ the process of arrival epochs and streaming times (call durations) of users of class k. We assume that $\tilde{N}_k$ are independent across $k = 1, \ldots, J$. Denote by $X_k(t) = \sum_n \mathbb{1}_{[T_n^k, T_n^k + W_n^k)}(t)$, with $\mathbb{1}_A(x) = 1$ being the indicator function of set A, the number of users of class k present in the system at time t and let $\boldsymbol{X}(t) = (X_1(t), \ldots, X_J(t))$; we call it the (vector of) user configuration at time t. The stationary distribution π of $\boldsymbol{X}(t)$ coincides with the distribution of the vector $(X_1, \ldots, X_J)$ of independent Poisson random variables with means $\mathbf{E}[X_k] := \rho_k = \lambda_k/\mu_k$, $k = 1, 2, \ldots, J$. We call ρ_k the *traffic demand* of class k.

Throughout the whole paper we adopt the usual convention for the numbering of the arrival epochs $T_0^k \leq 0 < T_1^k$. The same convention is used with respect to all point processes denoting some time epochs.

2.2 Resource constraints and outage policy

Users are supposed to be served by some (streaming) server for their whole sojourn times. The service corresponds to downloading some given content from the server with the requested bit-rate. However, due to limited service resources, for some configuration of users, service can be temporarily unavailable to some classes of users. We will call such classes being *in outage*. This means they cannot be offered the requested streaming bit-rate. [2] We assume that the users' requested streaming times are *not* altered by the eventual outages of their services. To be more specific, for class $k = 1, \ldots, J$, let a subset of user configurations $\mathcal{F}_k \subset \bar{\mathbb{N}}^J$ be given, where $\bar{\mathbb{N}} = \{0, 1, \ldots\}$, such that all X_k users of class k present in the configuration $\boldsymbol{X} = (X_1 \ldots, X_k, \ldots, X_J)$ are served if and only if $\boldsymbol{X} \in \mathcal{F}_k$ and no user of class k is served if $\boldsymbol{X} \notin \mathcal{F}_k$. We call $\mathcal{F}_k$ the *k th class (service) feasibility set*. (Particular choice of the feasibility sets in our model depends on the technological aspect of the streaming and on the outage policy implemented at the streaming server. Specific assumptions regarding streaming in wireless networks in will be considered in Section 4.)

Denote by $\pi_k = \pi(\mathcal{F}_k)$ the probability that the stationary configuration of uses is in k th class feasibility set.

In what follows we will assume that no user departure can cause outage of any class of users i.e., switch a given configuration from $\mathcal{F}_k$ to $\mathcal{F}'_k = \bar{\mathbb{N}}^J \setminus \mathcal{F}_k$. (However a user departure may make some class j switch from $\mathcal{F}'_j$ to $\mathcal{F}_j$.)

Regarding outage policy, we assume that, upon each arrival or departure of a user, the system updates its decision and, for any class k, it assigns the service to all users of class k if the updated configuration of users is in $\mathcal{F}_k$. All users of any class j for which the updated configuration is in $\mathcal{F}'_j$ will be placed in outage (at

1. This property is often referred to in the queuing context as the insensitivity property.

2. In practice, during the outage the streaming bit-rate can drop to a lower value or to zero; cf Remark 3.6.

least) until the next user arrival or departure. Consequently, $\tilde{X}_k(t) := X_k(t)\mathbb{1}_{\mathcal{F}_k}(\boldsymbol{X}(t))$ is the number of users of class k *not in outage* at time t. Denote by $\tilde{\boldsymbol{X}}(t) = (\tilde{X}_i(t), \ldots, \tilde{X}_J(t))$ the configuration of users not in outage at time t.

2.3 Performance metrics

In what follows we will be interested in the following characteristics of the model.

2.3.1 Virtual system metrics

During its time evolution, the user configuration $\boldsymbol{X}(t)$ alternates visits in the feasibility set $\mathcal{F}_k$ and its complement $\mathcal{F}'_k$, for each class $k = 1, \ldots, J$. We are interested in the expected visit durations in theses sets as well as the intensities (frequencies) of the alternations. More formally, for each given $k = 1, \ldots, J$, we define the point process $B_k := \{\tau_n^k : n\}$ of exit epochs of $\boldsymbol{X}(t)$ from $\mathcal{F}_k$; i.e., all epochs t such that $\Big(\boldsymbol{X}(t-), \boldsymbol{X}(t)\Big) \in \mathcal{F}_k \times \mathcal{F}'_k$ (with the convention $\tau_0^k \leq 0 < \tau_1^k$). These are epochs when all users of class k present in the system (if any) have their service interrupted.

Denote by $\sigma_n'^k := \sup\{t - \tau_n^k : \boldsymbol{X}(s) \in \mathcal{F}'_k \ \forall s \in [\tau_n^k, t)\}$ the duration of the n th visit of the process $\boldsymbol{X}(t)$ in $\mathcal{F}'_k$ and by $\sigma_n^k := \tau_{n+1}^k - \tau_n^k - \sigma_n'^k$ the duration of the n th visit of the process $\boldsymbol{X}(t)$ in $\mathcal{F}_k$. We define for each class $k = 1, \ldots, J$:

- *The intensity of outage incidents of class k*, i.e., the mean number of outage incidents of this class per unit of time
$$\Lambda_k := \lim_{T\to\infty} \frac{1}{T} \sum_n \mathbb{1}_{[0,T)}(\tau_n^k)\,.$$
Obviously Λ_k is also the intensity of entrance to the k th class feasibility set $\mathcal{F}_k$.
- *The mean service time between two outage incidents of class k*
$$\bar{\sigma}_k := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sigma_n^k\,.$$
- *The mean outage duration of class k*
$$\bar{\sigma}'_k := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sigma_n'^k\,.$$

Note that the above metrics characterize a "virtual" quality of the service, since some visits in $\mathcal{F}_k$ and $\mathcal{F}'_k$ may occur when there is no k th class user in the system (in the latter case the outage of this class is not experienced by any user).

2.3.2 User metrics

We adopt now a user point of view on the system. We define for each class $k = 1, \ldots, J$:

- *The probability of outage at the arrival epoch for user of class k*
$$P_k = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\mathcal{F}'_k}(\boldsymbol{X}(T_n^k))\,.$$
- *The mean total time in outage of user of class k*
$$D_k = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int_{[T_n^k, T_n^k + W_n^k)} \mathbb{1}_{\mathcal{F}'_k}(\boldsymbol{X}(t))\,\mathrm{d}t\,.$$

- *The mean number of outage incidents experienced by user of class k after its arrival*

$$M_k = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{m} \mathbb{1}_{(T_n^k, T_n^k + W_n^k)}(\tau_m^k)\,.$$

Note that eventual outage experienced at the arrival of a given user is not counted in M_k. The mean total number of outage incidents (including possibly at the arrival epoch) is hence $P_k + M_k$.

3 MATHEMATICAL RESULTS

For a given class $k = 1, \ldots, J$, denote by $\varepsilon_k = (0, \ldots, 1, \ldots, 0) \in \bar{\mathbb{N}}^J$ the unit vector having its kth component equal to 1. Hence $\boldsymbol{x} + \varepsilon_k$ represents adding one user of class k to the configuration of users $\boldsymbol{x} \in \bar{\mathbb{N}}^J$. Denote by $\mathbf{P}$ the probability under which $\{\boldsymbol{X}(t) : t\}$ is stationary and by $\mathbf{E}$ the corresponding expectation. Recall that $\pi\{\boldsymbol{x} \in \cdot\} = \mathbf{P}\{\boldsymbol{X}(t) \in \cdot\}$ is the distribution of the stationary configuration of users $\boldsymbol{X}(t)$ (it corresponds to independent Poisson variables of mean ρ_k).

3.1 General results

We present first results regarding the virtual system metrics. These results will be next used to evaluated the user metrics.

Lemma 3.1: *The intensity of outage incidents of class k is $\mathbf{P}$-almost surely equal to*

$$\Lambda_k = \sum_{j=1}^{J} \lambda_j \pi\left\{\boldsymbol{x} \in \mathcal{F}_k, \boldsymbol{x} + \varepsilon_j \in \mathcal{F}_k'\right\} \qquad k = 1, \ldots, J.$$

Proof: Let $N = \sum_{j=1}^{J} N_j$ be the point process counting the arrival times of users of all classes. By independence, N is the Poisson point process of intensity $\lambda = \sum_{j=1}^{J} \lambda_j$. Then, by the ergodicity of the process $\{\boldsymbol{X}(t) : t\}$ and the fact that the exists from $\mathcal{F}_k$ can take place only at some user arrival epoch we have by the Campbell formula [cf. e.g. 1, Equation (1.2.19)[3]],

$$\Lambda_k = \mathbf{E}\left[\int_{[0,1)} \mathbb{1}_{\mathcal{F}_k \times \mathcal{F}_k'}\left(\boldsymbol{X}\left(t-\right), \boldsymbol{X}\left(t\right)\right) N\left(\mathrm{d}t\right)\right] = \lambda \mathbf{P}_N^0\left\{\boldsymbol{X}(0-) \in \mathcal{F}_k, \boldsymbol{X}(0) \in \mathcal{F}_k'\right\}\,,$$

where $\mathbf{P}_N^0$ designates the Palm probability associated to N (which is, roughly speaking, the conditional probability given an arrival at time 0). By PASTA (Poisson Arrivals See Time Averages) property [cf. 1, Equation (3.3.4)] the configuration of users $\boldsymbol{X}(0-)$ under $\mathbf{P}_N^0$ has distribution π. Moreover, $\boldsymbol{X}(0) = \boldsymbol{X}(0-) + \varepsilon_\xi$ where $\xi \in \{1, \ldots, J\}$ is under $\mathbf{P}_N^0$ independent of $X(0-)$ and takes value j with probability λ_j/λ. This completes the proof. □

Lemma 3.2: *The mean service time between two outage incidents and the mean outage duration of class k are $\mathbf{P}$-almost surely equal to, respectively,*

$$\bar{\sigma}_k = \frac{\pi\left(\mathcal{F}_k\right)}{\Lambda_k}, \quad \bar{\sigma}_k' := \frac{\pi(\mathcal{F}_k')}{\Lambda_k} \qquad k = 1, \ldots, J,$$

where Λ_k is given in Lemma 3.1.

Proof: First we prove the expression for $\bar{\sigma}_k$. By ergodicity $\bar{\sigma}_k = \mathbf{E}_{B_k}^0\left[\sigma_0^k\right]$ $\mathbf{P}$-almost surely, where $\mathbf{E}_{B_k}^0$ designates the expectation with respect to the Palm probability associated to B_k, and $\mathbf{E}_{B_k}^0\left[\tau_0^k\right] = 1/\Lambda_k$; [see

3. with $Z_n := (\boldsymbol{X}(T_n-), \boldsymbol{X}(T_n))$ and $f(t,z) = \mathbb{1}_{[0,1)}(t)\mathbb{1}_{\mathcal{F}_k \times \mathcal{F}_k'}(z)$

e.g. 1, Equation (1.6.8) and Equation (1.2.27)]. Applying the mean value formula [see 1, Equation (1.3.2)[4]] we get $\pi(\mathcal{F}_k) = \Lambda_k \mathbf{E}^0_{B_k}\left[\sigma^k_0\right]$, which completes the proof of the expression for $\bar{\sigma}_k$. For the other expression, note by the definition of the sequence σ^k_n, σ'^k_n and τ^k_n that $\mathbf{P}$-almost surely,

$$\bar{\sigma}'_k = \mathbf{E}^0_{B_k}\left[\sigma'^k_0\right] = \mathbf{E}^0_{B_k}\left[\tau^k_1 - \sigma^k_0\right] = \frac{1}{\Lambda_k} - \frac{\pi(\mathcal{F}_k)}{\Lambda_k} = \frac{\pi(\mathcal{F}'_k)}{\Lambda_k},$$

which completes the proof. □

Proposition 3.3: *The probability of outage at the arrival epoch for user of class k is equal to*

$$P_k = \pi\left\{\boldsymbol{x} + \varepsilon_k \in \mathcal{F}'_k\right\} \qquad k = 1, \ldots, J \tag{1}$$

$\mathbf{P}$*-almost surely.*

Proof: By ergodicity we have $P_k = \mathbf{P}^0_{N_k}\left\{\boldsymbol{X}(0) \in \mathcal{F}'_k\right\}$, where $\mathbf{P}^0_{N_k}$ designates the Palm probability associated to N_k (arrival process of the users of class k). By PASTA property the configuration of users $\boldsymbol{X}(0-)$, just before arrival of the user of class k at time 0, has distribution π. Once the user enters the system, the users configuration becomes $\boldsymbol{X}(0-) + \varepsilon_k$, whence the result. □

Proposition 3.4: *The mean total time in outage of user of class k is $\mathbf{P}$-almost surely equal to*

$$D_k = \frac{1}{\mu_k}\pi\left\{\boldsymbol{x} + \varepsilon_k \in \mathcal{F}'_k\right\} \qquad k = 1, \ldots, J\,.$$

Proof: Again using the ergodicity of $\{\boldsymbol{X}(t)\}$ we can write

$$D_k = \mathbf{E}^0_{N_k}\left[\int_{[0,W^k_0)} \mathbb{1}_{\mathcal{F}'_k}(\boldsymbol{X}(t))\,\mathrm{d}t\right].$$

Denote by $\boldsymbol{Y}(t) := \boldsymbol{X}(t) - \varepsilon_k \mathbb{1}_{[T^k_0, T^k_0 + W^k_0)}(t)$ the process of configurations of users other than the user number 0 of class k (which arrives at time 0 under $\mathbf{E}^0_{N_k}$). By Slivnyak theorem [see e.g. 17, Theorem 1.13] the distribution of the process $\{\boldsymbol{Y}(t) : t\}$ under $\mathbf{E}^0_{N_k}$ is the same as this of $\{\boldsymbol{X}(t) : t\}$ under $\mathbf{P}$. Using the fact that W^k_0 and $\boldsymbol{Y}(t)$ are independent under $\mathbf{P}^0_{N_k}$ with $\mathbf{E}^0_{N_k}[W^k_0] = 1/\mu_k$ we obtain

$$D_k = \int_0^\infty \mathbf{E}^0_{N_k}\left[\mathbb{1}_{[0,W^k_0)}(t)\mathbb{1}_{\mathcal{F}'_k}(\boldsymbol{Y}(t) + \varepsilon_k)\right]\mathrm{d}t = \frac{1}{\mu_k}\,\pi\left\{\boldsymbol{x} + \varepsilon_k \in \mathcal{F}'_k)\right\},$$

which completes the proof. □

Proposition 3.5: *The mean number of outage incidents experienced by user of class k after its arrival is $\mathbf{P}$-almost surely equal to*

$$M_k = \frac{1}{\mu_k}\sum_{j=1}^{J} \lambda_j \pi\left\{\boldsymbol{x} + \varepsilon_k \in \mathcal{F}_k, \boldsymbol{x} + \varepsilon_k + \varepsilon_j \in \mathcal{F}'_k\right\}, \qquad k = 1, \ldots, J\,. \tag{2}$$

Proof: Again using the ergodicity of $\{X(t)\}$ we know that, $\mathbf{P}$-almost surely,

$$M_k = \mathbf{E}^0_{N_k}\left[\int_{(0,W^k_0)} B_k(\mathrm{d}t)\right].$$

Using the same arguments as these used in the proof of Proposition 3.4 we obtain

$$M_k = \mathbf{E}^0_{N_k}\left[B^*_k(0, W^k_0)\right] = \frac{\Lambda^*_k}{\mu_k}$$

where $B^*_k =: \{\tau^{*k}_n : n\}$ is the point process of exit epochs of $\boldsymbol{X}(t)$ from $\mathcal{F}^*_k = \{\boldsymbol{x} : \boldsymbol{x} + \varepsilon_k \in \mathcal{F}_k$ and Λ^*_k its intensity. Using Lemma 3.1 with $\mathcal{F}_k$ replaced by $\mathcal{F}^*_k$ concludes the proof. □

4. with $Z_k(t) = \mathbb{1}_{\mathcal{F}_k}(\boldsymbol{X}(t))$

Note that expressions given in the general results presented in this section, albeit complicated at first glance, involve only probabilities $\pi\{\boldsymbol{x} \in \cdot\}$ of the vector of independent Poisson random variables. Thus, they can be numerically calculated, in principle, for any choice of feasibility sets $\mathcal{F}_k$. Further analytical evaluation is also possible for some particular but natural definition of these sets, as will be shown in the next section.

3.2 Multi-Erlang resource constraint and least-effort-served-first service policies

In this section we will assume a multi-Erlang form of the resource constraint and define in this context some natural class of outage policies, for which our previous general expressions regarding the mean total time in outage and number of outage incidents can be evaluated quite explicitly on the ground of the Fourier analysis.

In what follows we assume that each user of class $k = 1, \ldots, J$ requires some amount φ_k, $0 < \varphi_k \leq 1$, of the total service capacity of the server normalized to 1; i.e., the resource constraint has the following *multi-Erlang form*

$$\sum_{\text{served classes } k} x_k \varphi_k \leq 1. \tag{3}$$

(In Section 4 we will explain under what assumption the above representation of the resource constraint appears in the context of wireless streaming and how φ_k depend on the requested streaming bit-rates and wireless channel conditions.)

We present now a parametric family of service polices for which *classes with smaller resource demands have higher service priority*. (See Remarks 3.6 and 4.1 fore the details on the pertinence of this assumption.) In this regard, we assume (without loss of generality) that the resource demands are ordered $\varphi_1 < \varphi_2 < \ldots < \varphi_J$. Recall also from Section 2.2 that assuming particular sets $\mathcal{F}_k$ entirely characterizes the outage policy.

For a given constant δ, $(0 \leq \delta \leq \infty)$ assume

$$\mathcal{F}_k = \mathcal{F}_k^\delta = \left\{ \boldsymbol{x} = (x_1, \ldots, x_J) \in \bar{\mathbb{N}}^J : \sum_{j=1}^{k-1} \varphi_j x_j + \varphi_k \sum_{j=k}^{J} x_j \mathbb{1}(\varphi_j \leq \varphi_k(1+\delta)) \leq 1 \right\}. \tag{4}$$

In what follows we will call the policy corresponding to (4) the *least-effort-served-first policy with δ-margin* (*LESF(δ)* for short).

Remark 3.6: The LESF(0) policy is *optimal* in the following sense: given constraint (3) and the assumption that the classes with smaller resource demands have higher priority, this policy allows to serve the maximal subset of users present in the system. For the same reason any LESF(δ) policy with $\delta > 0$ is clearly sub-optimal. In order to explain the motivation for considering such policies, one needs to extend the model and explain what actually happens with classes of users which experience service interruption. In this regard, for a given configuration of users $\boldsymbol{x} = (x_1, \ldots, x_J)$ denote by $K = K^\delta(\boldsymbol{x}) = \max\{k \in \{1, \ldots, J\} : \boldsymbol{X} \in \mathcal{F}_k^\delta\}$ the least-priority class which is still served by the policy LESF(δ) and by $C = C(\boldsymbol{x}) = \sum_{j=1}^{K} \varphi_j x_j \leq 1$ the actual fraction of the server capacity consumed by this task. The remaining server capacity $1 - C$ (which is not needed to serve users in classes $1, \ldots, K$) can be used to offer some "lower quality" service (e.g. streaming with lower video resolution, etc) to the users in classes $K+1, \ldots, J$ which are in outage. Note by (4) that the remaining server capacity under the policy LESF(δ) is at least

$$1 - C \geq \varphi_K \sum_{j=K+1}^{J} x_j \mathbb{1}(\varphi_j \leq \varphi_K(1+\delta)).$$

Hence, the server accepting the class K as the least-priority class being "fully" served, leaves enough remaining capacity to be able to make the same effort (allocate service capacity φ_K) for all users in outage in classes whose service demand exceeds φ_K by no more than $\delta \times 100\%$. These latter users will not have "full" required service (since they require $\varphi_j > \varphi_K$ for the full service) but only some "lower quality" service (see Remark 4.1 for more details). Consequently, one can conclude that policies LESF(δ) with $\delta > 0$, being sub-optimal, ensure some *fairness*, in the sense explained above. Clearly the policy LESF(∞) (i.e., with $\delta = \infty$) is the most fair in this class, in the sense that it reserves enough remaining capacity to offers the "lower quality" service for *all* users in outage. Thus, we will call LESF(∞) the *LESF fair* policy.

In order to evaluate the general expressions developed in Section 3.1 we need to calculate probabilities $\pi\{\boldsymbol{x} \in \mathcal{F}_k\}$, $\pi\{\boldsymbol{x} \in \mathcal{F}_k, \boldsymbol{x}+\varepsilon_j \in \mathcal{F}'_k\}$ and $\pi\{\boldsymbol{x}+\varepsilon_k \in \mathcal{F}_k, \boldsymbol{x}+\varepsilon_k+\varepsilon_j \in \mathcal{F}'_k\}$. In the case of policies LESF(δ) the evaluation of the above probabilities boils down to the calculation of the probability distribution functions $F_k^\delta(t) := \mathbf{P}\left\{\sum_{j=1}^k X_j^\delta \varphi_j \le t\right\}$ where $X_j^\delta = X_j$ for $j = 1, \ldots, k-1$ and $X_k^\delta = \sum_{j=k}^J X_j \mathbb{1}(\varphi_j \le \varphi_k(1+\delta))$ with $X_1, \ldots, X_J$ being independent, Poisson random variables with parameters, ρ_j, respectively. Indeed, note for example that $\pi\{\boldsymbol{x} \in \mathcal{F}_k\} = F_k^\delta(1)$, $\pi\{\boldsymbol{x}+\varepsilon_k \in \mathcal{F}_k, \boldsymbol{x}+\varepsilon_k+\varepsilon_j \in \mathcal{F}'_k\} = F_k^\delta(1-\varphi_k) - F_k^\delta(1-\varphi_k+\varphi_j)$ and similarly for other probabilities.

Note that for a given k, the random variables $X_1^\delta, \ldots, X_k^\delta$ are also independent, of Poisson distribution, with parameters $\rho_1^\delta, \ldots, \rho_k^\delta$, respectively, where $\rho_j^\delta = \rho_j$ for $j = 1, \ldots, k-1$ and $\rho_k^\delta = \sum_{j=k}^J \rho_j \mathbb{1}(\varphi_j \le \varphi_k(1+\delta))$. Consequently, the evaluation of $F_k^\delta(t)$ can be done using is Laplace transform $\mathcal{L}_k^\delta(\theta) := \int_0^\infty e^{-\theta s} F_k^\delta(s)\mathrm{d}s$, which are explicitly known.

Fact 3.7: We have

$$\mathcal{L}_k^\delta(\theta) = \frac{1}{\theta} \exp\left[\sum_{j=1}^k \rho_j^\delta \left(e^{-\theta\varphi_j} - 1\right)\right]$$

Proof: Use [17, Proposition 1.2.2] and a general relation $\int_0^\infty e^{-\theta s} F(s)\,\mathrm{d}s = \frac{1}{\theta}\int_0^\infty e^{-\theta s} F(\mathrm{d}s)$. □

The probabilities $F_k^\delta(\cdot)$ may be retrieved from $\mathcal{L}_k^\delta(\cdot)$ using standard techniques. For example [18, with the algorithm implemented by Hollenbeck [19] in Matlab]. In what folloos we present a more explicit result based on the Bromwich contour inversion integral. In this regard, denote $\overline{\mathcal{L}}_k^\delta(\theta) = 1/\theta - \mathcal{L}_k^\delta(\theta)$ (which is the Laplace transform of complementary distribution function $1 - F_k^\delta(t)$). Also, denote by $\mathcal{R}(z)$ the real part of the complex number z.

Fact 3.8: We have

$$F_k^\delta(t) = 1 - \frac{2e^{at}}{\pi}\int_0^\infty \mathcal{R}\left(\overline{\mathcal{L}}_k^\delta(a+iu)\right)\cos ut\,\mathrm{d}u\,, \tag{5}$$

where $a > 0$ is an arbitrary constant.

Proof: See [4]. □

Remark 3.9: As shown in [4], the integral in (5) can be numerically evaluated using the trapezoidal rule, with the parameter a allowing control the approximation error. Specifically, for $n = 0, 1, \ldots$ define

$$h_n(t) = h_n(t; a, k, \delta) := \frac{(-1)^n e^{a/2}}{t}\mathcal{R}\left(\overline{\mathcal{L}}_k^\delta\left(\frac{a+2n\pi i}{2t}\right)\right)\,,$$

$S_n(t) := \frac{h_0(t)}{2} + \sum_{i=1}^n h_i(t)$, and $S(t) = \lim_{n\to\infty} S_n(t)$. Then $\left|F_k^\delta(t) - (1 - S(t))\right| \le e^{-a}$. Finally, the

(alternating) infinite series $S(t)$ can be efficiently approximated using for example the Euler summation rule

$$S(t) \approx \sum_{i=0}^{M} \binom{M}{i} 2^{-M} S_{N+i}(t)$$

with a typical choice $N = 15$, $M = 11$.

4 STREAMING IN WIRELESS CELLULAR NETWORKS

In this section we will further specify our model to fit the context of wireless cellular networks.

Our present context is a configuration of users $\boldsymbol{x} = (x_1, \ldots, x_J)$ wishing to obtain wireless transmission from a given base station at some predefined bit-rates $\boldsymbol{r} = (r_1, \ldots, r_J)$, where r_k denotes the bit-rate required by the users of class k.

4.1 Multiple access with orthogonal AWGN channels

Denote by $r_k^{\max}$ the *maximal bit-rate* of a user of class k, achievable when it is served alone by the base station. Recall that in multiple-access channels the achievable data rates of a given user depend on data rates selected by other users. Assume the following *rate-region*, i.e., of the set of *mutually* achievable rates.

Orthogonal channels: Assume that rates $\boldsymbol{r}$ are achievable for the configuration $\boldsymbol{x}$, if $x_k r_k = \lambda_k r_k^{\max}$ for some non-negative vector $(\lambda_1, \ldots, \lambda_J)$, such that $\sum_{k=1}^{J} \lambda_k \leq 1$. It is straightforward to see that the above assumption is equivalent to the resource constraint in the previously considered multi-Erlang form (3) with resource demands $\varphi_k = \lambda_k / x_k = r_k / r_k^{\max}$.

The above assumption corresponds to the situation, when users neither hamper nor assist each other's transmission. They use channels which are perfectly separated in time, frequency or by orthogonal codes, nevertheless sharing these resources. [5]

We make now an assumption on the maximal achievable rates. In order to keep this exposition simple, but also to be able to consider "real" coding schemes, we will express $r_k^{\max}$ as some fraction of the maximal theoretical bit-rate achievable in the *additive white Gaussion noise* (AWGN) *single input single output* channel; cf [20, Th .9.1.1].

AWGN SISO-like channels: We assume that the maximal bit-rate of a user of class k, achievable when it is served alone by the base station is equal to $r_k^{\max} = \gamma W \log(1 + \mathrm{SNR}_k)$, where W is the frequency bandwidth, SNR_k is the *signal to noise ratio* (to be explained in the next section) of k-class user. Moreover, γ (with $0 < \gamma \leq 1$) is the coefficient telling how close a given coding scheme approaches the theoretical Shanon's bound for AWGN SISO channel (corresponding to $\gamma = 1$). [6]

Concluding this part of the model, we assume the vector of resource demands to be the vector of *order statistics* of

$$\varphi(r_k, \mathrm{SNR}_k) := \frac{r_k}{\gamma W \log(1 + \mathrm{SNR}_k)}\,, \tag{6}$$

5. It is the case for current LTE (Long Term Evolution) norm for cellular networks based on OFDMA, as well as for other multiple access techniques as FDA, TDMA, CDMA assuming perfect in-cell orthogonality, and even HDR neglecting the scheduler gain.

6. It was also shown in [21] how that the performance of AWGN *multiple input multiple output* (MIMO) channel can be approximated by taking values of $\gamma \geq 1$. Another possibility to consider MIMO channel is to use the exact capacity formula given in [22].

where r_k and SNR_k are, respectively, required bit-rates and SNR's of different classes of users.[7] Remark that different classes of users correspond to different streaming rate requests r_k as well as to different SNR conditions of users. In other words, even if we assume that all users require the same streaming rate $r_k = r$, we still need a multi-class model due to (typically) different SNR's of users in wireless cellular networks, and that in this case the service priority corresponds to the ordering of SNR's (inversely with respect to the values of φ). We will complete this line of thought in Section 4.3.

Remark 4.1: Following Remark 3.6 we will specify now a natural model for the best-effort "lower quality" streaming that can be offered for users in outage in association with a given LESF(δ) policy. Assuming that $\varphi_k = r_k/r_k^{\max}$ are increasing in k, recall that $K = K^\delta(\boldsymbol{x})$ denotes the largest class-index k such that $\boldsymbol{x} \in \mathcal{F}_k^\delta$. For $k > K$ denote

$$r'_k = r_k^{'\delta}(\boldsymbol{x}) = r_k^{\max} \frac{1 - \sum_{j=1}^{K} x_j \varphi_j}{\sum_{j=K+1}^{J} x_j \mathbb{1}(\varphi_j \le (1+\delta)\varphi_K)} \quad \text{if } \varphi_k \le (1+\delta)\varphi_K \text{ and 0 otherwise.} \tag{7}$$

Note that the rates $(r_1, \ldots, r_K, r'_{K+1}, \ldots, r'_J)$ are achievable for the configuration $\boldsymbol{x}$ of users given orthogonal channels described above. Assuming that users of class k, when in outage, are given the rate r'_k, we denote by

$$T_k = T_k^\delta = \mu_k \mathbf{E}_{N_k}^0 \left[\int_{[0, W_0^k)} r_k \mathbb{1}(\boldsymbol{X}(t) \in \mathcal{F}_k^\delta) + r_k^{'\delta}(\boldsymbol{X}(t)) \mathbb{1}(\boldsymbol{X}(t) \notin \mathcal{F}_k^\delta) \, \mathrm{d}t \right]$$

the *mean throughput of user of class k during his service time*. It is easy to see, as in the proof of Proposition 3.4, that $T_k = r_k \pi \left\{ \boldsymbol{x} + \varepsilon_k \in \mathcal{F}_k^\delta \right\} + T'_k$, where

$$T'_k = \mathbf{E} \left[r_k^{'\delta}(\boldsymbol{X}(t) + \varepsilon_k) \mathbb{1}((\boldsymbol{X}(t) + \varepsilon_k) \notin \mathcal{F}_k^\delta) \right] \tag{8}$$

can be interpreted as the *part of the throughput obtained by user of class k during its outage time*. Recall from (7) that this throughput is realized by user k when it is in outage, i.e., $k > K$, but $\varphi_k \le (1+\delta)\varphi_K$. In the case of equal requested rates r_k, the latter two conditions are equivalent to

$$(1 + \mathrm{SINR}_K)^{1/(1+\delta)} - 1 \le \mathrm{SINR}_k \le \mathrm{SINR}_K \, . \tag{9}$$

Finally, note that T'_k is expressed in (8) as an integral with respect to the stationary distribution of the free arrival process $\boldsymbol{X}(t)$, which is the distribution π of the vector of independent Poisson random variables with parameters ρ_k. It can be evaluated easily by Monte Carlo simulation. In Section 5.3 we will study T'_k for LESF(δ) policies with different values of δ.

4.2 AWGN broadcast channel

Before taking about SNR in wireless cellular networks (in the next section), let us make in this section a brief digression on how the assumption of orthogonal channels can be relaxed. From information theory point of view, the assumption that users neither hamper nor assist each other's transmission is a suboptimal assumption. In fact, the theoretically optimal performance is offered by the *broadcast channel* model. Let us briefly show how a least-effort-served-first policy could be defined in the case of AWGN broadcast channel. This model does not lead to the multi-Erlang form of the resource constraint considered in Section 3.1 but can be studied

7. Formally, to have φ_i ordered, we define $(\varphi_1, \ldots, \varphi_J) = \mathcal{O}(\varphi(r_1, \mathrm{SNR}_1), \ldots, \varphi(r_J, \mathrm{SNR}_J))$ where $\mathcal{O}(\xi_1, \ldots, \xi_J) = (\xi_{(1)}, \ldots, \xi_{(J)})$ denotes a permutation of $(\xi_1, \ldots, \xi_J)$ such that $\xi_1 \le \ldots \le \xi_J$.

using our results of Section 3.1 for general feasibility sets $\mathcal{F}_k$. Assume that higher priority for service is given to the classes with higher values of SNR and that they are ordered such that $\mathrm{SNR}_1 \geq \mathrm{SNR}_2 \geq \ldots \geq \mathrm{SNR}_J$. It is known that in the case of AWGN broadcast channel the rates $\boldsymbol{r}$ are (theoretically) achievable for the configuration $\boldsymbol{x}$ if (and only if) there exists a vector $(\lambda_1, \ldots, \lambda_J)$, such that $\sum_{k=1}^{J} \lambda_k \leq 1$ and

$$x_k r_k = W \log\left(1 + \frac{\lambda_k}{1/\mathrm{SNR}_k + \sum_{i=1}^{k-1} \lambda_i}\right) \qquad k = 1, \ldots, J;$$

cf [23, Eq. 6.29]. Note that the above equation can be iteratively solved for the vector $(\lambda_1, \ldots, \lambda_J)$ as follows $\lambda_1 := \lambda_1(x_1 r_1, \mathrm{SNR}_1)$ and $\lambda_k := \lambda_k(x_k r_k, \mathrm{SNR}_k, \lambda_1, \ldots, \lambda_{k-1})$ for $k = 2, \ldots, J$. Consequently, the kth class feasibility set $\mathcal{F}_k$ of the policy that gives higher priority to classes with higher values of SNR is $\mathcal{F}_k := \{\boldsymbol{x} : \sum_{i=1}^{k} \lambda_i \leq 1\}$. Finding efficient method for the calculation of the corresponding probabilities $\pi(\mathcal{F}_k)$ is left for further research and the broadcast channel will not be considered in the remaining part of the paper.

4.3 SINR in multicellular network context

Coming back to the line of thought from the end of Section 4.1, we need to characterize the classes of users by the values $(\mathrm{SNR}_1, \ldots, \mathrm{SNR}_J)$ of SNR they experience in a given wireless network. This SNR depends on the path-loss of the user from the serving (streaming) base station, typically including shadowing, and the thermal noise power. Moreover, in multicellular network scenario the noise comprises *interference* from other base stations, which depends on the path-loss from these other base stations. Thus, in the remaining part of this paper we will be taking SNR:=SINR (signal to interference and noise ratio) in (6).

In order to choose some representative values of SINR in a given network and to know what fraction of users experience a given value, we need to know the distribution of the SINR (with respect to the serving base station) experienced in this network, biased by the spatial repartition of arrivals of streaming calls. It can be obtained from measurements, simulations or analytics evaluation of an appropriate stochastic model. It is not our goal in this paper to describe these methods in details or to present stochastic-geometric modeling of wireless cellular networks. (We will be more specific, however, regarding our choice when presenting numerical results in Section 5.) Assume simply, that we are given a cumulative distribution function (CDF) of the SINR expressed in dB, $F(x) := \mathbf{P}\{10 \log_{10}(\mathrm{SINR}) \leq x\}$, obtained from either of these methods.

Consider a discrete probability mass function

$$p_k := F\Big(\frac{x_{k+1} + x_k}{2}\Big) - F\Big(\frac{x_k + x_{k-1}}{2}\Big) \qquad k = 1, 2, \ldots, J\,, \tag{10}$$

with $x_0 = -\infty$, $x_{J+1} = \infty$, approximating the given CDF F. We define the class $k = 1, \ldots, J$ of users as all users having the SINR expressed in dB in the interval $\big((x_k + x_{k-1})/2, (x_{k+1} + x_k)/2\big]$, and approximate their SINR by the common value $\mathrm{SINR}_k = 10^{x_k/10}$. Note that the intensity of arrivals λ_k of users of class k is equal to $\lambda_k = p_k \lambda$ where $\lambda = \sum_{i=k}^{J} \lambda_k$ is the total arrival intensity, to be specified together with the CDF F of the SINR.

5 Numerical results — quality of streaming in LTE

In this section we will use the analytical approach developed in previous sections to evaluate the quality of streaming in LTE networks.

5.1 LTE model and traffic specification

5.1.1 CDF of the SINR

We obtain the CDF F of SINR presented on Figure 1 from the simulation compliant with the 3GPP recommendation in the so-called calibration case, (compare to [24, Figure A.2.2-1(right)]).

More precisely, we consider the geometric pattern of BS placed on the 6×6 hexagonal lattice. In the middle of each hexagon there are three symmetrically oriented BS antennas , which gives a total of 108 BS antennas. The distance between the centers of two neighboring hexagons is 500m. Each BS antenna is characterised by the following horizontal pattern $A(\phi) = -\min(12(\phi/\theta)^2, A_m)$, where ϕ is the angle in degrees, with $\theta = 70°$, $A_m = 20dB$, and uses transmission power $P = 60$dBm (including omnidirectional gain of 14dBi).

The distance-loss model (corresponding to the frequency carrier 2GHz) is $L(r) = 128.1 + 37.6 \log_{10}(r)$[dB] where r is the distance in km. A supplementary penetration loss of 20dB is added.

The shadowing is modeled as a centered log-normal random variable of standard deviation 8dB. The noise power equals -95dBm.

In order to obtain the empirical CDF of the SINR we generate 3600 random user locations uniformly in the network (100 user locations per hexagon on average). Each user is connected to the antenna with the strongest received signal (smallest propagation-loss including distance, shadowing and antenna pattern) and the SINR is calculated. The obtained empirical CDF F of the SINR is shown on Figure 1.

5.1.2 Link characteristics

3GPP shows in [25, §A.2] that there is a 25% gap between the practical coding schemes and the Shannon's limit for the AWGN channel. Moreover, some of the transmitted bits are used for signaling, which induces a supplementary capacity loss of about 30% (see [26, §6.8]). This made us assume $\gamma = 0.5 (\approx 0.75(1 - 0.3))$ in (6). The system bandwidth is $W = 10$MHz.

5.1.3 Streaming traffic

We assume that all calls require the same streaming rate $r_k = 256$ kbit/s and have the same streaming (sojourn) time distribution. We split them into $J = 100$ user classes characterized by values of the SINR falling into different intervals regularly approximating the SINR domain from $x_1 = -10$dB to $x_J = 17$dB as explained in Section 4.3. In our performance evaluation we will consider two values of the spatially uniform traffic demand: 900 and 600 Erlang/km^2. (Results presented in what follows do not depend on the mean streaming time but only on the traffic demand). Consequently, k th class traffic demand per unit of surface (λ_k/μ_k)/km^2 is equal to, respectively, $p_k \times 900$ and $p_k \times 600$Erlang/km^2, where p_k are given by (10).

5.2 Performance evaluation

Assuming the LTE and traffic model described above, we consider now streaming policies LESF(δ) defined in Section 3.2 and further specified in Section 4.1 for wireless scenario. Recall that in doing so, we assume that users are served by the antenna offering the smallest path-loss, and dispose orthogonal down-link channels, with the maximal rates $r_k^{\max}$ depending on the value of the SINR (interference comes from non-serving BS) characterizing class k. Roughly speaking, LESF(δ) policy assigns the total requested streaming rate $r_k =$

256kbit/s for the maximal possible subset of classes in the order of decreasing SINR, leaving some capacity margin to offer some "best-effort" streaming rates for (some) users remaining in outage. These streaming rates r'_k given by (7) depend on the current configuration of users and are non-zero for users with SINR within the interval $(1 + \text{SINR}_K)^{1/(1+\delta)} - 1 \leq \text{SINR} \leq \text{SINR}_K$, where SINR_K is the minimal value of SINR for which users are assigned the total requested streaming rate; cf (9). In particular, LESF(0), called the *optimal* policy, leaves no capacity margin for users in outage, while LESF(∞), called the *fair* one, offers a "best-effort" streaming rate for all users in outage at the price of assigning the full requested rate 256kbit/s to a smaller number of classes (higher value of SINR_K) [8]. In what follows, we use our results of Section 3 to evaluate performance of these streaming policies in the LTE network model.

5.2.1 Outage time

Figure 2 shows the mean fraction of the requested streaming time spent in outage, $\mu_k D_k$ evaluated in Proposition 3.4, in function of the SINR value characterizing class k, for the traffic 900 Erlang/km^2 and different policies LESF(δ). Figure 3 shows the analogous results assuming traffic load of 600 Erlang/km^2. The main observations are as follows:

- All LESF policies exhibit a cut-off behaviour: the fraction of time in outage drops rapidly from 100% to 0% when SINR transgresses some critical values. This cut-off is more strict for the optimal policy.
- For the traffic of 900 Erlang/km^2, users with SINR$\geq$ 3dB are practically never in outage, when the optimal policy is used. The same holds true for users with SINR$\geq$ 13dB, when the fair policy is used.
- When the traffic drops to 600 Erlang/km^2, these critical values of SINR decrease by 2dB and 5dB, respectively, for the optimal and the fair policy. Note that the fair policy is more sensitive to higher traffic load.

5.2.2 Number of outage incidents

Figure 4 shows the mean number of outage incidents per service time, M_k evaluated in Proposition 3.5, in function of the SINR value characterizing class k, for the traffic 900 Erlang/km^2 and different policies LESF(δ). Figure 5 shows the analogous results assuming traffic load of 600 Erlang/km^2. The main observations are as follows:

- For all policies, the number of outage incidents (during the service) is non-zero only for users with the SINR close to the critical values revealed by the analysis of the outage times. Users with SINR below these values are constantly in outage while users with SINR above them never in outage.
- More fair policies generate slightly more outage incidents. The worst values are 2 to 2.2 interruptions per service for the optimal policy, depending on the traffic value, and 2.4 to 3 interruptions per service for the fair policy.

Note that studying outage times and outage incidents there is no apparent reason for considering fair policies. This motivates our study of the throughput in the following section.

8. The LESF fair policy seems to be adopted in some implementations of the LTE.

5.3 "Best effort" service, outage and "deep outage"

Figure 6 shows the fraction of time spent by a given user in outage and in "deep outage" in function of the SINR, assuming traffic 900 Erlang/km^2. Recall, outage denotes the situation when the full requested streaming rate (assumed 256kbit/s in our example) is not offered. By "deep outage" we call the situation when the user does not receive even the "best effort" service offered to some users in outage by LESF(δ) policies with $\delta > 0$. More precisely, when his SINR is smaller than $(1 + \text{SINR}_K)^{1/(1+\delta)} - 1$, where SINR_K is the minimal value of the SINR for which users are assigned the total requested streaming rate by a given policy; cf (9). For comparison, the fraction of time spend in outage (as on Figure 2, when the requested full service rate is not offered) is also plotted.

Figure 7 shows also two curves for all policies LESF(δ) assuming traffic 900 Erlang/km^2. The upper ones represent the mean total throughput realized during the service, normalized to its maximal value; i.e., $T_k/(256\text{kbit/s})$, in function of the SINR value characterizing class k (cf. Remark 4.1). The fractions of this throughput realized during outage periods, $T'_k/(256\text{kbit/s})$, are represented by the lower curves.

Figures 7 and 6 teach us that the role of the LESF(δ) policies with $\delta > 0$ may be two-fold.

- LESF(δ) policies with small values of δ, e.g. $\delta = 0.5$, *improve "temporal homogeneity" of service with respect to the optimal policy, for users having SINR near the critical value.* For example, a user having SINR equal to 1dB is served by the optimal policy during 80% of time with the full requested streaming rate (cf. Figure 6). However, for the remaining 20% of time it does not receive any service ("deep outage", rate 0bits/s). The policy LESF(0.5) offers to such a user 80% of the requested streaming rate during the whole streaming time (cf. Figure 7), with no "deep outage" periods (cf. Figure 6). The price for this is that a slightly higher SINR is required to receive the full requested streaming rate (at least 5dB, instead of 3dB for the optimal policy).
- The fair policy LESF(∞) *improves the spatial homogeneity of service*. It leaves no user in "deep outage", however a much larger SINR$=$ 13dB is required for not to be in outage (cf. Figure 6). Moreover, the throughput of all users which are in outage but not in "deep outage" with intermediate LESF policies is reduced, e.g. from 80% to 40% for SINR$=$ 1dB (cf. Figure 7).

6 Conclusions

We proposed a new, stochastic call-service model with service capacity sharing and interruptions. It can be seen as an extension of the classical Erlang's loss model, where service denials (losses) are not definitive for a given call, but only temporal, having a form of outage periods. As in the classical loss model, one can express its key performance characteristics (as e.g. the number and total duration of the outage periods for a typical call), in terms of the stationary probabilities of the (free) traffic demand process. We specify and use a multi-class version of this new model, with a multi-rate Erlang resource constraint, to study the quality of real-time streaming experienced by users of 3GPP LTE cellular network, in function of their location in the network, characterized by the values of the SINR they experience. We identify and evaluate some natural parametric class of outage policies, in which users with larger SINR have higher priority for service, and which allow to chose arbitrarily between the optimality and fairness in real-time streaming.

REFERENCES

[1] F. Baccelli and P. Brémaud, *Elements of queueing theory; Palm martingale calculus and stochastic recurrences*. Springer, 2003.

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

[3] J. Roberts, "A service system with heterogeneous user requirements," in *Performance of Data Communications Systems and their Applications*, G. Pujolle, Ed., 1981.

[4] J. Abate and W. Whitt, "Numerical inversion of Laplace transforms of probability distributions," *ORSA Journal on Computing*, vol. 7, no. 1, pp. 38–43, 1995.

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

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

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

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

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

[10] M. K. Karray, "Analytical evaluation of qos in the downlink of OFDMA wireless cellular networks serving streaming and elastic traffic," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, May 2010.

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

[12] F. H. Fitzek, S. Hendrata, P. Seeling, and M. Reisslein, "Video streaming in wireless Internet," in *Mobile Internet: Enabling Technologies and Services*, ser. Electrical Engineering & Applied Signal Processing, S. Apostolis, Ed. CRC Press, 2004, ch. 11.

[13] G. Liang and B. Liang, "Effect of delay and buffering on jitter-free streaming over random VBR channels," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1128–1141, 2008.

[14] A. ParandehGheibi, M. Médard, S. Shakkottai, and A. Ozdaglar, "Avoiding interruptions — QoE tradeoffs in block-coded streaming media applications," in *Proc. of International Symposium on Information Theory*, june 2010, pp. 1778–1782.

[15] L. Rong, S. Elayoubi, and O. Haddada, "Performance evaluation of cellular networks offering TV services," *IEEE Trans. on Vehicular Technology*, vol. 60, no. 2, pp. 644–655, 2011.

[16] Y. Xu, E. Altman, R. E. Azouzi, M. Haddad, S.-E. Elayoubi, and T. Jiménez, "Probabilistic analysis of buffer starvation in markovian queues," in *in Proc. of Infocom'12*, Orlando, FL USA, 2012.

[17] F. Baccelli and B. Błaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume I — Theory*, ser. Foundations and Trends in Networking. NoW Publishers, 2009, vol. 3, No 3–4.

[18] F. R. de Hoog, J. H. Knight, and A. N. Stokes, "An improved method for numerical inversion of laplace transforms," *SIAM Journal of Scientific and Statistical Computation*, vol. 3, no. 3, pp. 357–366, 1982.

[19] K. J. Hollenbeck, "Invlap.m: A Matlab function for numerical inversion of Laplace transforms by the de Hoog algorithm," 1998. [Online]. Available: www.isva.dtu.dk/staff/karl/invlap.htm

[20] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, Inc., 2006.

[21] M. Karray and M. Jovanovic, "Theoretically feasible QoS in a MIMO cellular network compared to the practical LTE performance," in *Proc. of ICWMC*, Venice, Italy, 2012.

[22] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–596, November 1999.

[23] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[24] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA — physical layer aspects," Tech. Rep. 36.814-V900, 2010. [Online]. Available: www.3gpp.org/ftp/Specs/archive/36_series/36.814/

[25] ——, "Evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) system scenarios," Tech. Rep. 36.942-V830, 2010. [Online]. Available: www.3gpp.org/ftp/Specs/archive/36_series/36.942/

[26] ——, "Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation," Tech. Rep. 36.211-V910, 2010. [Online]. Available: www.3gpp.org/ftp/Specs/archive/36_series/36.211/

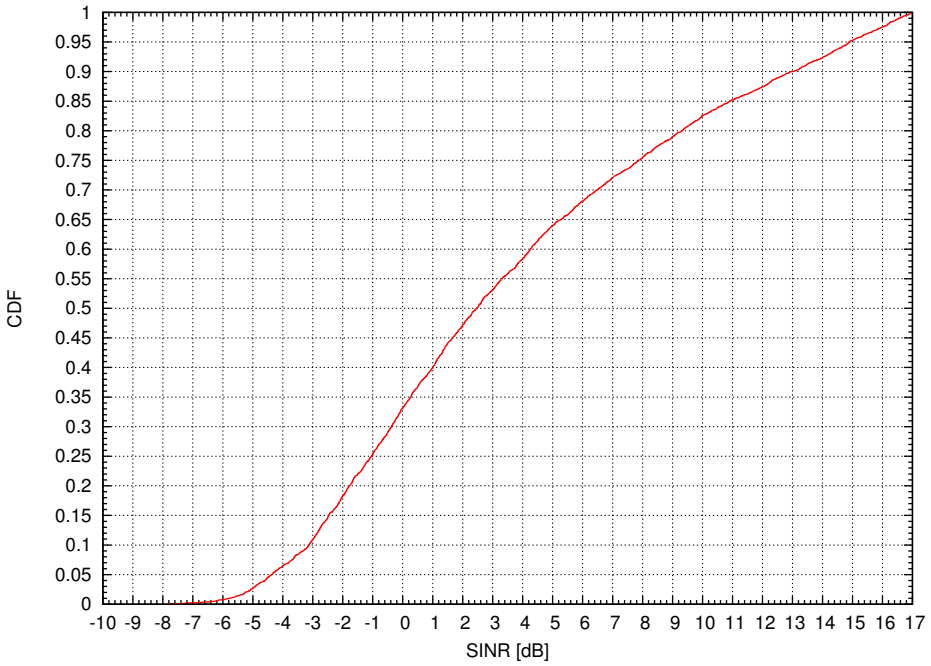

Fig. 1. Cumulative distribution function of the SINR obtained according to 3GPP specification; see Section 5.1.1.

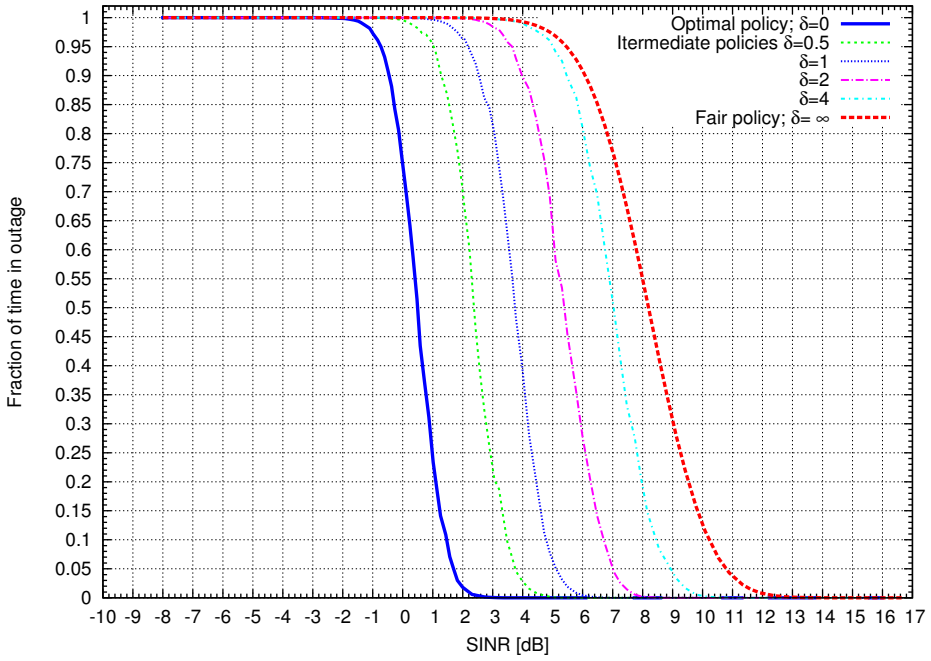

Fig. 2. Mean fraction of the requested streaming time in outage, in function of the user SINR for different policies LESF(δ); traffic 900 Erlang/km^2.

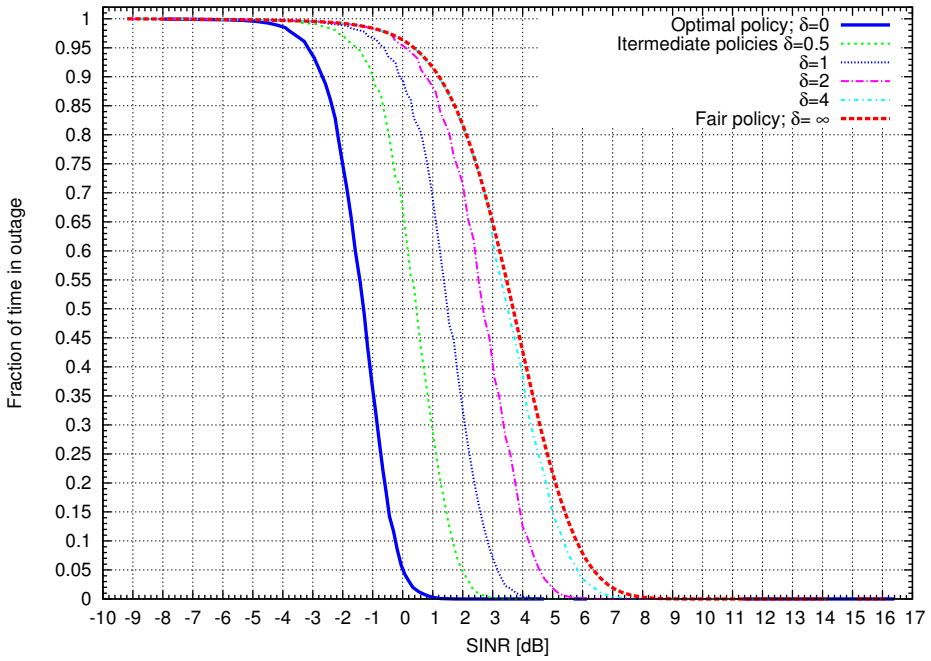

Fig. 3. Fraction of time in outage as on Figure 2 for traffic 600 Erlang/km^2.

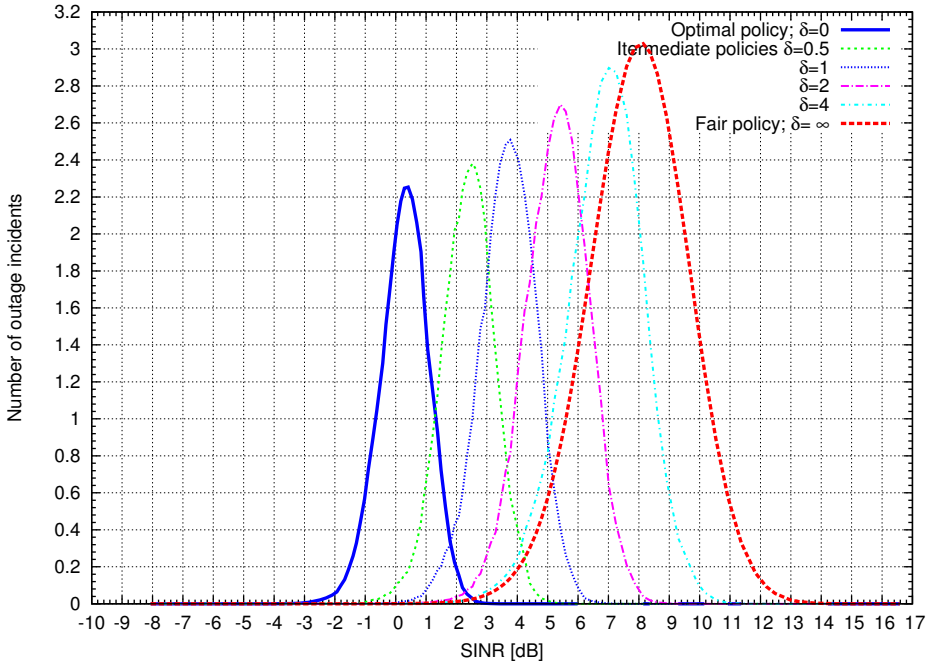

Fig. 4. Number of outage incidents during the requested streaming time, in function of the user SINR for different policies LESF(δ); traffic 900 Erlang/km^2.

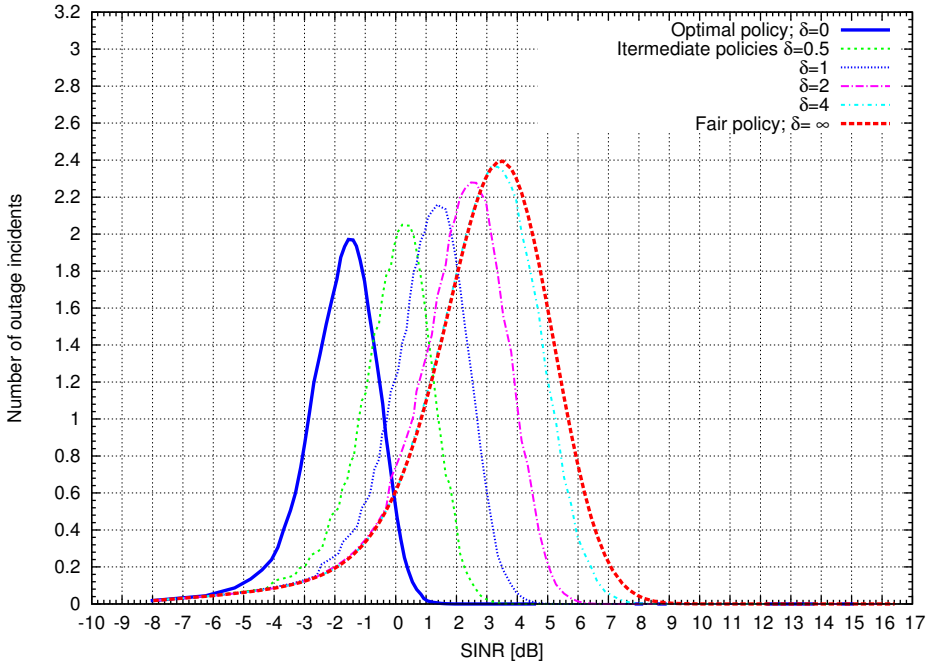

Fig. 5. Number of outage incidents as on Figure 4 for traffic 600 Erlang/km^2.

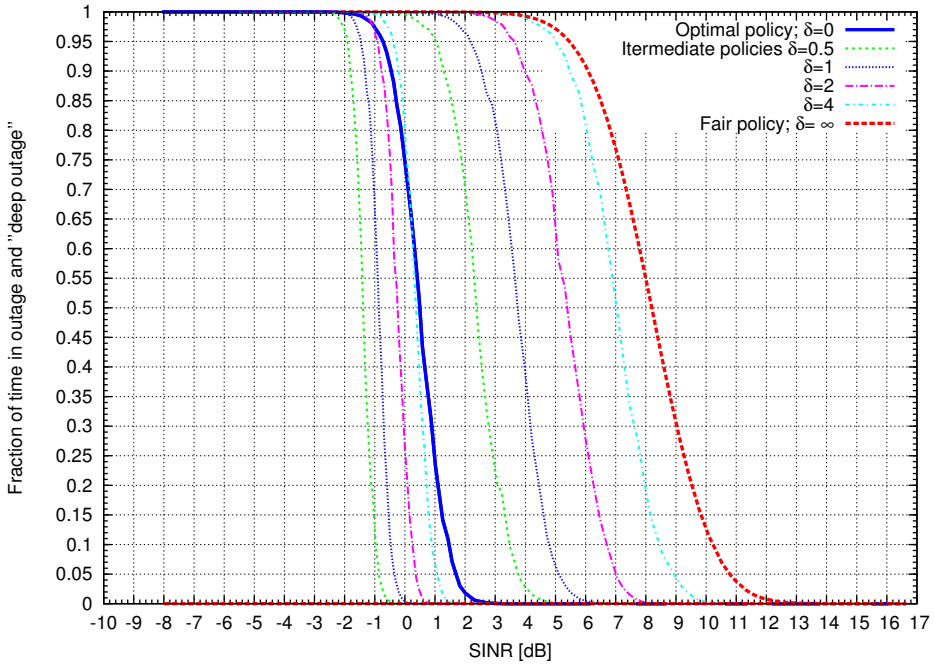

Fig. 6. "Best effort" reduction of outage time. For any policy LESF(δ), with $0 < \delta < \infty$, the left curve represents the fraction of time spent in "deep outage", when the "best effort" service is not offered. For comparison, the time spend in outage (as on Figure 2, when the requested full service rate is not offered) is also plotted (the right curve of a given style). The optimal policy ($\delta = 0$) does not offer any "best effort" service. The fair policy ($\delta = 0$) offers this service for all users in outage.

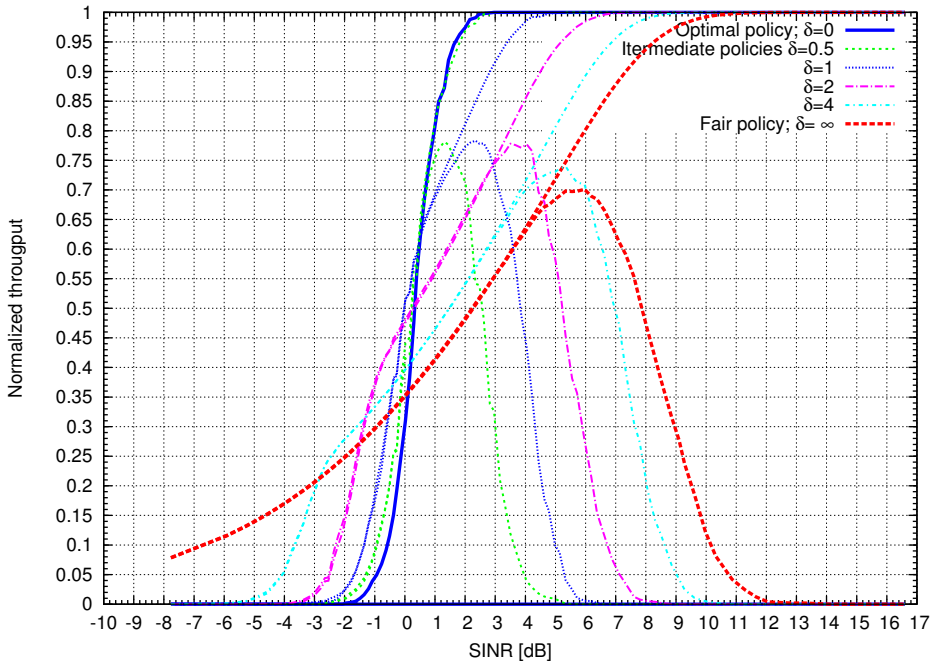

Fig. 7. Mean total throughput normalized to its maximal value 256kbit/s obtained during the service time (upper curves) and its fraction obtained when user in outage (lower curves) for different policies LESF(δ) traffic 900 Erlang/km^2.