Flow level performance evaluation in mobile networks: Analytical modeling and empirical validation

Salah Eddine Elayoubi, Younes Khadraoui, Bruno Baynat, Taoufik En-Najjary

Flow level performance evaluation in mobile networks: Analytical modeling and empirical validation

Salah Eddine Elayoubi[1], Younes Khadraoui[1], Bruno Baynat[2] and Taoufik En-Najjary[1]
[1] Orange Labs, Issy-les-Moulineaux, France
[2] Sorbonne Université, UPMC Univ Paris 06, CNRS, LIP6 Laboratory, Paris, France

***Abstract*—The purpose of this paper is to identify, through empirical validation, the best analytical model for QoS and capacity estimation in mobile data networks. We first present two different sets of models that have been proposed in the literature: the infinite source, Erlang-like models and the finite source, Engset-like models. While the former models are widely developed and adapted to mobile networks, the latter are less used in the mobile context. We thus derive novel analytical models to complete the finite source theory before moving to the comparison of the different models. We make use of network measurements originating from an HSPA network and compare the modeled performance with the observed one, for each of the available analytical models. Our results show that a Processor Sharing analysis that takes as cell capacity the harmonic average of the achievable throughputs and as traffic inputs the volumes generated in each CQI, gives results that are close to the field measurements. As of the finite source models, we observed a good match on some cells and a mismatch on other cells. This is due to the difficulty of extracting traffic parameters from the field. However, finite source models have a better predictive power than infinite source models, as their traffic characteristics can more easily be linked to changes in the network such as the introduction of new services or new devices.**

I. Introduction

With the explosion of data traffic over mobile networks, Quality of Service (QoS) and capacity prediction is becoming a hot topic for operators. Indeed, robust performance evaluation models are needed in order to be able to predict the evolution of QoS and plan capacity upgrade plans. These latter may consist of adding new radio carriers, deploying new sites or even supplementing a site that has 3G with the new 4G technology.

As the objective of mobile operators is to ensure an acceptable *user* QoS, performance models at flow level are needed. By flow level we mean that the dynamic users' behavior is taken into account. Indeed, in modern data networks, a user initiates a call, transfers a certain amount of data, and leaves the system. The pioneer work of Bonald et al. [8] proposed an analytical model, based on the Processor Sharing (PS) theory, that takes into account the heterogeneity of radio conditions over the cell. In this model, users will accumulate at cell edges as a transfer will take more time if the offered throughput is lower. The authors in [22] extended the model of [8] to the case where multiple classes of service (voice, streaming and data) share the radio resources. Note that these models give a cell capacity that is proportional to the *harmonic* mean of the throughputs observed over the cell surface, as this harmonic mean gives more weight to positions with lower rates [9]. Other models like that proposed in [16] use an *arithmetic* mean of throughputs as a measure for the capacity, the main arguments being that user mobility allows observing several radio conditions during a communication, thus preventing user accumulation at cell edge. Recent works have extended Processor sharing models to adaptive streaming traffic [19] and to different scheduling schemes [1], or to take into account mobility of users [4]. These models have also been extended to the cell coordination schemes in HSPA networks [18].

The above mentioned models suppose an *infinite source of users*: Users arrive from the outside to the network following a Poisson process that is independent from the network state; they are called Erlang-like models, in reference to the original Erlang model proposed by A. K. Erlang [12]. Other performance evaluation models for mobile networks are inspired from the Engset model [13] and take the hypothesis of a *finite source of users*: Users that are physically present in the cell are limited and their activity is described by an ON/OFF process. Examples of such works are [5], [6], that considered that each cell of the network has a finite number of users, each of them generating ON sessions carrying a pre-determined volume of data distributed following an exponential law that is independent from others' activities. Each ON session is then followed by an OFF duration, called reading duration. Note that the available models in the literature suppose that each user may visit all the possible radio conditions during its communication; which is somehow related to a high mobility assumption.

In this paper, we extend these models to the case with low mobility, where users stay with constant radio conditions during a typical communication, thus accumulating at cell edges. Then, in order to identify the best suitable analytical model among the ones of the literature and the new ones developed in the current paper, we perform an empirical validation based on measurements collected from a live HSPA network. The measurement data include radio parameters, indicating the distribution of radio conditions, for each of the considered cells, as well as traffic data. The latter include

* This work has been carried out in the framework of IDEFIX project, funded by the ANR under the contract number ANR-13-INFR-0006.

offered traffic volumes per radio condition and typical behavior per user (ON/OFF characteristics). We apply these radio and traffic measurements as inputs to the different analytical methods, and compare the performance results with the Key Performance Indicators (KPI) observed on the field.

The original contributions of this paper are as follows:

- We present and classify the main flow level performance evaluation methods that have been proposed in the literature.
- We extend the Engset-like methods to the case of elastic traffic where users do not change their average radio conditions during a communication.
- We propose a methodology to feed performance evaluation tools with inputs from live networks, in terms of radio conditions and traffic parameters.
- We compare empirically the different analytical models and show the advantages and drawbacks of each of them. To the best of our knowledge, this is the first work that assesses empirically the performance of flow level models based on field measurements.

The remainder of this paper is organized as follows. In Section II, we present the infinite source queuing models. Section III deals with the finite source queuing models and extends the theory to the low mobility case. Section IV shows how the analytical models can be adapted to take as inputs network measurements. Section V compares the outputs of the analytical models to the performance observed in the field and discusses the pros and cons of each method. Finally, Section VI concludes the paper.

II. Infinite source queuing models

This set of models assumes that there is an infinite source that generates connections in the cell. The theory is well elaborated in the literature, so that we give only an overview of the results that have been obtained in, e.g. [8] and show how that can be applied to an arbitrary number of classes of radio conditions.

A. Radio model

In order to model the capacity of mobile networks, we have to understand how the cell resources are shared among users and how users benefit from obtained resources. We begin by considering round robin scheduling, opportunistic scheduling is considered next. In this case, for a given number of active users, resources (for instance Time Slots in HSPA and Resource Blocks in LTE) are equally divided among users. A user that is alone in the cell will have different bit rates if he is close to the base station, compared to the case where he is far from it, as illustrated in Figure 1. A typical cell of the network can thus be divided into zones of equal radio conditions, or classes, each characterized by an achievable throughput, i.e., a throughput that can be obtained by a user when scheduled by the base station.

Suppose now that the radio conditions are known, obtained analytically, by simulations or from network measurements as will be shown later. We thus assume that the cell can be divided into K zones, each one being associated to a given class i of users, $i = 1, ..., K$. A user of class i will obtain a throughput C_i, if it is alone in the whole cell. These values are not sufficient for estimating the average capacity of the cell. Indeed, the radio interface is a shared broadcast medium, and the traffic dynamics are to be taken into account when considering the capacity, as explained in the following subsection.

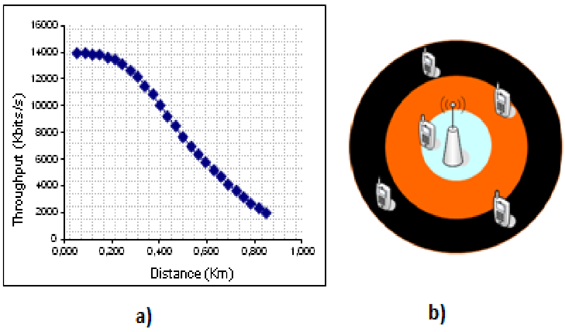

Fig. 1. a) Example of simulated achievable throughput over an HSPA cell: only one user is in the cell but is simulated at different positions in the cell. b) Cell decomposition into zones of equal radio conditions.

B. Processor sharing analysis

We consider here that connection demands arrive to the cell according to a Poisson process of intensity λ connections per second. We assume that a user that carries a new connection demand has a probability p_i to be off class i. In addition we suppose that users do not change their radio conditions during their communication and thus remain of the same class during their whole data transfer, i.e., no mobility is considered here. As a result, connection demands of class i arrive to the cell according to a Poisson process of rate $\lambda_i = \lambda p_i$. Each connection (regardless of its class) brings an average amount of data equal to F bits (a file size may follow an arbitrary distribution with average F), and it ends upon completion of this download. The amount of traffic volume offered to the cell per unit time is thus equal to $V = \lambda F$ (in bit/s). The load generated by users of class i is equal to $\rho_i = \frac{\lambda p_i F}{C_i}$, and the overall cell load is [8]:

$$\bar{\rho} = \sum_{i=1}^{K} \rho_i = \frac{\lambda F}{\bar{C}} \quad (1)$$

where $\bar{C}$ corresponds to the harmonic mean of achievable throughput, and can be regarded as the system capacity [8]:

$$\bar{C} = \frac{1}{\displaystyle\sum_{i=1}^{K} \frac{p_i}{C_i}} \quad (2)$$

The idea behind this is that, as far as we considered elastic services where the aim is to download a pre-determined amount of data, users at cell edge will have lower data rates and will stay longer in the cell and contribute more to the cell load. Based on this load expression, one can model the cell as a Processor Sharing (PS) queue of capacity $\bar{C}$, and the steady-state probability of having n active users in the cell are calculated by:

$$\bar{P}_{\mathrm{I}}(n) = \bar{\rho}^n(1-\bar{\rho}) = \left(\frac{V}{\bar{C}}\right)^n(1-\frac{V}{\bar{C}}) \tag{3}$$

the subscript I standing for “infinite source”.

Knowing the steady-state probabilities, one can derive several performance metrics. For instance, the average number of active flows is calculated by:

$$\bar{n} = \sum_{n=1}^{\infty} n\bar{P}_{\mathrm{I}}(n) = \frac{\bar{\rho}}{1-\bar{\rho}} \tag{4}$$

The flow throughput of users of class i is calculated by:

$$\bar{\gamma}_i = C_i(1-\bar{\rho}) \tag{5}$$

and the average flow throughput over the cell is:

$$\bar{\gamma} = \bar{C}(1-\bar{\rho}) \tag{6}$$

C. Harmonic mean versus arithmetic mean for the cell capacity

Some works in the literature suppose that the cell capacity is given by the arithmetic mean of achievable throughputs $\hat{C} = \sum_{i=1}^{K} p_i C_i$, instead of the harmonic mean given in relation 2. Obviously if the probabilities p_i involved in averaging $\hat{C}$ are the same as those appearing in $\bar{C}$, $\hat{C}$ will be different from $\bar{C}$, and modeling the cell by a PS queue of capacity $\hat{C}$ would provide different and thus false results if the aforementioned assumptions (e.g., no mobility of users) are satisfied. However, it is interesting to highlight that an arithmetic mean can provide an equivalent calculation of the cell capacity, if the probabilities involved in the summation have a different interpretation. Let us consider the following expression for $\hat{C}$:

$$\hat{C} = \sum_{i=1}^{K} p'_i C_i \tag{7}$$

where p'_i is not anymore the probability for a user of class i to generate a new connection demand (like for p_i), but the probability that an ongoing transfer in the cell is of class i, i.e., uses throughput C_i. p'_i can alternately be seen as the proportion of ongoing transfers that are of class i. Where p_i was only characterized by the ratio $\frac{\lambda_i}{\lambda}$, p'_i must now be expressed as the ratio $\frac{\rho_i}{\rho}$, where $\rho_i = \frac{\lambda_i F}{C_i}$ and $\rho = \sum_{i=1}^{K} \rho_i$. Indeed the more connection demands of class i, the more likely an ongoing transfer is of class i (like in p_i), but in addition, the higher the throughput C_i, the faster class i users finish their download and the less likely an ongoing transfer is of class i. Now if we replace in $\hat{C}$ the probabilities p'_i by their expressions, we obtain that $\hat{C} = \bar{C}$:

$$\hat{C} = \sum_{i=1}^{K} \frac{\rho_i C_i}{\rho} = \sum_{i=1}^{K} \frac{\lambda_i F}{\rho} = \frac{\lambda F}{\displaystyle\sum_{i=1}^{K} \rho_i} = \frac{1}{\displaystyle\sum_{i=1}^{K} \frac{p_i}{C_i}} = \bar{C} \tag{8}$$

As a conclusion of this subsection, one must be careful when choosing the averaging method, depending on the traffic inputs that are available (volume per radio condition or corresponding load). We will describe the traffic inputs in the numerical evaluation section. And it is of high importance to precisely characterize the probabilities that are involved in the averaging.

D. Introducing opportunistic scheduling gain

The above analysis supposes that the throughput achieved by a user of class i is equal to the throughput when he is alone in the cell divided by the number of active users. This makes the cell capacity $\bar{C}$ independent of the number of active users. This is a realistic assumption as long as a round robin scheduling is considered. However, opportunistic schedulers like Proportional Fair Scheduler (PFS) are usually activated on the networks and bring a multi-user diversity gain that has to be taken into account in the capacity analysis. The scheduling decision is in this case based on the instantaneous throughput $r_u(t)$ of user u at time t (which is zero if the user is not scheduled) and on the average throughput $R_u(t)$ of user u at time t, evaluated through the moving average:

$$R_u(t) = (1-\theta)R_u(t-1) + \theta r_u(t),$$

for some fixed time parameter θ (t is discrete and corresponds to scheduling time slots of the user). The PFS schedules the user with the best relative radio conditions, that is with the maximum ratio:

$$\frac{r_u(t)}{R_u(t)}.$$

The scheduling gain compared to a blind algorithm like round-robin depends on the number of users n and on the channel model. If we consider Rayleigh fast fading, the gain $G(n)$ is approximately equal to the harmonic sum $1 + \ldots + \frac{1}{n}$ for n users [7]. Note that the gain can be obtained numerically for more realistic channels like PA3 or VA3 channels (as defined in the 3GPP release 12 specifications [23]). In this paper, we introduce a scheduling gain calculated following the methodology of [14] and that corresponds to a PA3 channel that is usually reported to give realistic results in an urban environment [2]. The scheduling gains that are used are presented in Table I below.

TABLE I
SCHEDULING GAIN

Nb. of users	1	2	3	4	5	6	7	8	9
Scheduler gain	1	1.24	1.35	1.4	1.44	1.46	1.47	1,48	1.5

When opportunistic gain is considered, the system can be modeled as a State Dependent Processor Sharing queue, and the steady state probabilities for the no mobility case will still have a product form given by:

$$\bar{P}_{\mathrm{I}}^{\mathrm{PFS}}(n) = \bar{P}_{\mathrm{I}}^{\mathrm{PFS}}(0)\frac{\bar{\rho}^n}{\prod_{m=1}^{n} G(m)} \tag{9}$$

where $\bar{P}_{\mathrm{I}}^{\mathrm{PFS}}(0)$ is a normalizing constant. Note that an underlying assumption on users' channels for the product-form expression 9 to hold is that the rate variations are symmetric around the mean rate [11]. A similar expression can be found for the high mobility system, by replacing the harmonic average by an arithmetic one in the load expression.

III. Finite source queuing models

These models suppose that a finite number of users are physically present in the cell. This implies that, unlike the infinite source model, the arrival rate of connections depends on the number of users that are currently being served. Indeed, the more users are in service the less are likely to come. The other key assumption is that, for each user the generated traffic is an ON/OFF process, where the ON period corresponds to an activity session characterized by the average size of the transfered data $X_{\rm on}$ (note that for comparing infinite source and finite source models, we will take $X_{\rm on} = F$), while the OFF period corresponds to a reading time characterized by its average duration $T_{\rm off}$. We consider the same radio model as for the infinite source case, i.e., the cell is characterized by a set of K radio conditions, each one corresponding to a throughput C_i.

A. High mobility model

In [5] and [6], authors assume that there is a constant number N of users generating an ON/OFF trafic (as described above), and that, at any time slot, any user that is currently in ON period has a probability p_i to be of class i, i.e., to use a throughput C_i. In a sense, this corresponds to a high mobility pattern. It has been shown that the cell capacity can be expressed as an arithmetic mean of all achievable throughputs: $\hat{C} = \sum_i p_i C_i$, and the steady state probability of having $n \le N$ active users is given by relation (10) [5]. Note that, as discussed in Section II-C, the probabilities p_i involved in the arithmetic mean of $\hat{C}$ does not have the same meaning as those involved in the harmonic mean of $\bar{C}$ for the infinite source PS queue model (relation (2)).

$$\hat{P}_{\rm F}(n) = \hat{P}_{\rm F}(0)\left(\frac{X_{\rm on}}{T_{\rm off}\hat{C}}\right)^n \prod_{i=1}^{n}(N-i+1) \qquad (10)$$

the constant $\hat{P}_{\rm F}(0)$ can be obtained by normalization as:

$$\sum_{n=0}^{N} \hat{P}_{\rm F}(n) = 1$$

We can finally calculate the cell load as: $\hat{\rho}_{\rm F} = 1 - \hat{P}_{\rm F}(0)$.

B. No mobility model

We now consider the other extreme case where users cannot change their radio conditions at each time slot, but are rather fixed. To the best of our knowledge, this system has not been modeled in the literature. The system corresponds to the case where there is a fixed number of users for each class i, denoted by $N_1, N_2, ..., N_K$. We model this system with a network of K pairs of queues where each pair represents a class of users, be they in an OFF or an ON state, as illustrated in Figure 2.

Proposition 1: The probability of having the distribution $\vec{n} = (n_1, n_2, ...n_K)$ is given by:

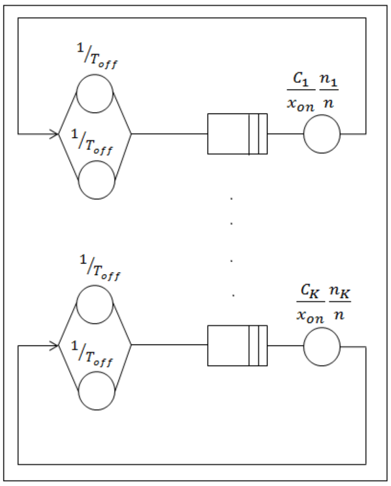
$^1/_{T_{off}}$
$^1/_{T_{off}}$
$\frac{C_1}{x_{on}}\frac{n_1}{n}$
$^1/_{T_{off}}$
$^1/_{T_{off}}$
$\frac{C_K}{x_{on}}\frac{n_K}{n}$

Fig. 2. No mobility system: each tandem of queues represents users of a given class. When a user achieves his service (the ON period) he returns to the OFF mode until the next activity period.

$$Pr(\vec{n}) = Pr[0]\frac{n!}{n_1!...n_K!}\prod_{i=1}^{K}\frac{N_i!}{(N_i-n_i)!}\alpha_i^{n_i} \qquad (11)$$

where $\alpha_i = \frac{X_{\rm on}}{T_{\rm off}C_i}$.

Proof: The service rate of queue i is equal to:

$$\mu_i = \frac{C_i}{X_{\rm on}}\frac{n_i}{n}$$

where $n = n_1 + ... + n_K$ is the total number of users. These service rates can be easily shown to verify the symmetric balance equations:

$$\mu_i(\vec{n})\mu_j(\vec{n}-\vec{e}_i) = \mu_j(\vec{n})\mu_i(\vec{n}-\vec{e}_j) = \frac{n_i n_j}{n(n-1)}$$

where $\vec{e}_k$ is a vector with 1 in position k and 0 elsewhere.

The system of $2K$ queues represented in Figure 2 corresponds thus to a closed Whittle network [10]. It has been shown in [10] that the embedded Markov chain in such a system is reversible and the system has a product-form solution for the steady-state probabilities. These latter can thus be shown, after some calculations, to be equal to that of expression (11).

C. Alternative simple model

The problem of the model with no mobility is that the steady-state probabilities of equation (11) cannot be reduced to a single dimension as for equations (3) and (10). The computation time for such models becomes prohibitive, especially for network dimensioning tools that deal with a large number of cells. In order to cope with this issue, we consider an alternative system that corresponds to a theoretical mobility case, by supposing that users can change their radio condition only after an ON session. This is obviously a simplifying assumption, which makes this model an intermediate case between the high mobility and the no mobility model. We still assume that there is a constant number N of users generating ON/OFF trafic in the cell, and we define p_i as the probability that a user that ends an OFF period is of class i, i.e., uses a throughput C_i, during the whole duration of its next ON period. Note that the probabilities p_i have, once again, a different interpretation than the probabilities used in the expression of the harmonic mean of $\bar{C}$ (used in the infinite source PS model), or the arithmetic mean of $\hat{C}$ (used in the finite source high mobility model).

This system can be modeled by a network of $K+1$ queues, K of them representing the ON states of the K classes, and the remaining one the OFF state, as illustrated in Figure 3.

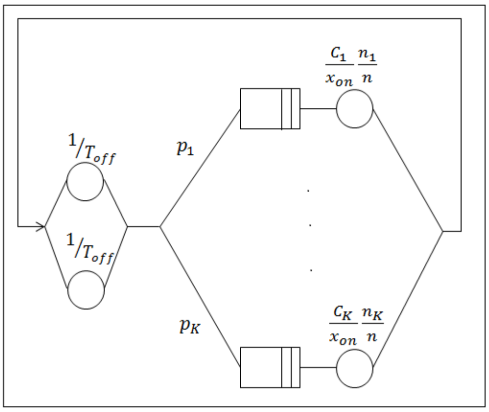

Fig. 3. Athernative system with a theoretical low mobility pattern: When a user achieves his OFF period, he choses following probability p_i to join queue i during his ON session; he then returns to the OFF queue until his next activity period.

Proposition 2: The probability of having the distribution $\vec{n} = (n_1, n_2, ...n_K)$ is given by:

$$Pr(\vec{n}) = Pr[0]\frac{N!}{(N-n)!}\frac{n!}{n_1!...n_K!}\alpha_1^{n_1}\alpha_2^{n_2}...\alpha_K^{n_K} \qquad (12)$$

Proof: The service rate of each queue remains the same as before; the system is thus still a Whittle network [10]. The same analysis as that of Proposition 1 leads to the expression (12).

Knowing these steady-state probabilities, we show in the following proposition that the system evolution can be described by the one-dimensional state $n = n_1 + ... + n_K$ instead of the K-dimensional state $\vec{n}$.

Proposition 3: The probability of having n users in the cell can be calculated by:

$$\bar{P}_{\mathrm{F}}(n) = \bar{P}_{\mathrm{F}}(0)\left(\frac{X_{\mathrm{on}}}{T_{\mathrm{off}}\bar{C}}\right)^n \prod_{i=1}^{n}(N-i+1) \qquad (13)$$

where $\bar{C}$ is the harmonic average of the throughputs over the cell.

Proof: The probability of having n users in the cell is equal to the probability of all possible distributions of $\vec{n}$ such as $n_1 + n_2 + ...n_K = n$:

$$\bar{P}(n) = \sum_{\vec{n}|n_1+n_2+...n_K=n} Pr(\vec{n}) =$$

$$Pr[0]\frac{N!}{(N-n)!}\sum_{\vec{n}|n_1+n_2+...n_K=n}\frac{n!}{n_1!...n_K!}\alpha_1^{n_1}\alpha_2^{n_2}...\alpha_K^{n_K}$$

In order to prove equation (13), it is sufficient to use the fact that :

$$\sum_{\vec{n}|n_1+n_2+...n_K=n}\frac{n!}{n_1!...n_K!}\left(\frac{p_1}{C_1}\right)^{n_1}...\left(\frac{p_K}{C_K}\right)^{n_K} = \left(\sum_{i=1}^{K}\frac{p_i}{C_i}\right)^n$$

Note that this expression corresponds to the model with high mobility (equation (10)), by replacing the arithmetic average by a harmonic one, and using probabilities p_i that have a different interpretation.

The system of Figure 3 can be regarded as an approximation of the system in Figure 2 when the maximal number of users in the two cases verify $N = \sum_{i=1}^{K} N_i$ and $N_i \approx p_i N$. Indeed, as a user during its activity period in Figure 3 does not change its radio class, the phenomenon of accumulation of users with bad radio conditions, characterizing the system of Figure 2 will still happen. In order to illustrate this, we plot in Figure 4 the cell load obtained for different values of X_{on} for the approximate and exact finite source models with static users (equations (13) and (11)), taking the same simple example of the previous section, with $T_{\mathrm{off}} = 40$ seconds and $N = 10$ users in the cell. We observe that the harmonic average method is a good approximation for the system with no mobility. We thus use, in the remainder of this paper, the harmonic average method for modeling the no mobility case.

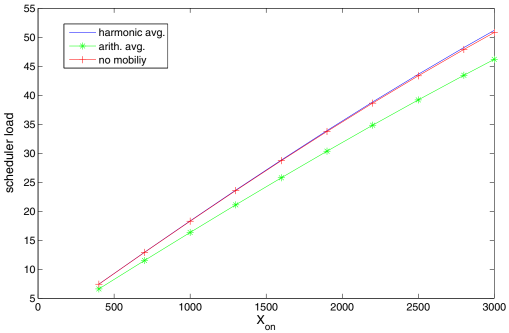

Fig. 4. Cell load vs. X_{on}

D. Impact of opportunistic scheduling

As explained for the infinite source models, multi-user diversity gains are to be taken into account in the performance calculations.

Proposition 4: The steady-state expressions (10) and (13) can be extended for considering a scheduling gain $G(n)$ that depends on the number of active users, as follows:

$$\hat{P}_{\mathrm{F}}^{\mathrm{PFS}}(n) = \hat{P}_{\mathrm{F}}^{\mathrm{PFS}}(0)\left(\frac{\lambda_u X_{\mathrm{on}}}{\hat{C}}\right)^n \frac{1}{\prod_{m=1}^{n} G(m)}\prod_{i=1}^{n}(N-i+1) \qquad (14)$$

and

$$\bar{P}_{\mathrm{F}}^{\mathrm{PFS}}(n) = \hat{P}_{\mathrm{F}}^{\mathrm{PFS}}(0)\left(\frac{\lambda_u X_{\mathrm{on}}}{\bar{C}}\right)^n \frac{1}{\prod_{m=1}^{n} G(m)} \prod_{i=1}^{n}(N-i+1) \tag{15}$$

Proof: The Whittle network condition still holds when the service rate of users of class i is equal to:

$$\mu_i = \frac{C_i}{X_{\mathrm{on}}} \frac{n_i}{n} G(n)$$

Expressions (10) and (13) can be derived as for the case of a constant capacity.

IV. Empirical data description

The multiplicity of analytical models that can be used for performance evaluation in cellular networks shows the need for an empirical validation and comparison of these models. In this section, we describe the measurements that have been collected from the field for being used in the validation process. We consider a live HSPA network in a major European city, focus on the downlink and make use of two sources of data:

- Operation and Maintenance Center (OMC) counters: these counters aggregate the information reported by user equipments to base stations, in addition to the data measured directly by the base stations. They include traffic volume information, radio conditions and performance metrics (loads, number of active users, etc).
- Probes data: Detailed traffic information that describe the behavior of users (number of connections per user, traffic volume per connection, etc.) cannot be obtained from OMC counters, but rather from traffic probes installed on the interface between the access and the core network. Our dataset consists of two captures of data call records of one hour each, the first one between 9 am and 10 am and the second trace collected between 7 pm and 8 pm. For each radio connection, the probe collects several pieces of information, namely: customer id (assigned anonymously by the probe, the probe keeps the same id inside the area covered by the RNC), the start and end time of the connection, DL/UL volumes in bytes, and cell identifier (assigned anonymously by the probe).

A. Radio measurements

We first begin by describing the radio-related parameters used in the validation process.

- CQI distribution: The main signal quality measure in HSPA is the so-called Channel Quality Indicator (CQI). This measurement is a quantified Signal -to-Interference-plus-Noise-Ratio (SINR) reported from UEs to their serving base station, available later at the OMC level. Using these field measurements, a CQI distribution can be constructed. An example CQI distribution obtained from the field is given in Figure 5.
- HSPA Device mix: This distribution indicates the proportion of HSPA device categories in terms of traffic volume. The construction of this distribution is similar to that stated for CQI distribution, with the appropriate field counters. Note that HSPA devices that are currently present on the networks range from category 6 (16 QAM modulation with a peak throughput of 3.6 Mbit/s) to Category 24 (dual cell with 64 QAM and a peak throughput of 42.2 MBit/s).

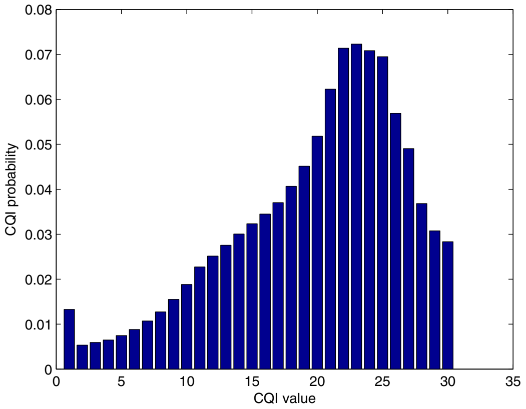

Fig. 5. An example field CQI distribution

B. Traffic measurements

The simplest traffic measure to obtain is the overall traffic intensity V in bit/s. It can be obtained by dividing the traffic volume, a KPI provided by OMC counters, by the observation time.

The other traffic measures that are used in this paper, especially for the finite source models, cannot be obtained from OMC counters, but from probes. Before estimating the average session volume X_{on} and average time between sessions T_{off} we need to take into account some characteristics of the new generation of smartphones. Indeed, new smartphones are designed to be always connected to Internet in order to receive notifications for the applications installed on them like social networks, email, news, etc. To receive notifications, smartphones either keep a continuous TCP connection with the servers by regularly sending keep-alive packets, or regularly initiate a new TCP connection toward the servers to check for new notifications. These background activities generate a huge number of very short radio connections with negligible amount of data traffic. Separating background activity from regular activity is not an easy problem. In this work, we assume all radio connections with less the 2 Kbytes (less than two MTU TCP packets) as background activity. These connections represent 62% of total number of radio connections but only 0.001% of transmitted volume. Keeping these connections while doing the estimation of X_{on} and T_{off} lead to a ridiculously small values that are not representative of the actual connections generating the observed traffic volumes. Therefore, we drop all these connections from our dataset.

In order to make the calibration process easier, we suppose that X_{on} and T_{off} are characteristics of customers behavior and are not dependent on the network cells. This of course a simplification (that will be relaxed in Section V-B), as user behavior is known to actually depend on the cell [21]. Therefore, we first estimate these values over the whole trace after cleaning out the background activity. We thus obtain, for each operating hour, specific values of X_{on} and T_{off}. We will check the validity of this assumption afterwards. We perform the following estimations:

- In order to determine if two consecutive packets having the same destination belong to the same downloading ON period, or to two consecutive ON periods separated by a reading OFF period, we considered a constant threshold (empirically set to 2s) under which the two packets are assumed to belong to the same ON period, and above which they are assumed to belong to two consecutive ON periods.
- X_{on} is estimated as the average of the volumes of all the radio connections in our dataset.
- For each customer in our dataset (we have more than 80 thousands customers per trace) we estimate the median interconnection time (MIT), and T_{off} is estimated as the median of all MITs of all customers.
- For each cell, the number of users physically present in the cell N is estimated as the total number of customers connected to this cell during our observation. We make use of two estimates: the first takes into account all the unique users observed during the considered hour on the cell; this gives an absolute maximum of the number of physically present users. The other estimate considers only users that generated at least two connections during the considered hour, in order to be sure that these users stayed in the cell for a significant amount of time.

C. Performance metrics

In addition to the traffic inputs and radio parameters, OMC data give some performance metrics, for instance:

- Scheduler load: this OMC metric corresponds to the proportion of time slots that are used for data transmission.
- Average number of active users: this OMC metric gives the average number of users for which the base station has some data to transmit (that are scheduled or wait to be scheduled).
- Average user throughput: this is not a metric that is directly given by counters, but can be estimated as the ratio between the traffic volume and the average number of active users.

D. Taking into account voice traffic

The models presented in this paper are specific to elastic data traffic. However, HSPA networks usually share the same spectrum resources with circuit switched services (R99 voice service) and some cells have a significant volume of voice traffic. As the circuit-switched traffic has priority in power resource allocation, the remaining power resource is assigned to HSPA, which reduces the throughput of data users compared to a carrier that is dedicated to HSPA. In order to estimate the amount of power consumed by voice users, we make use of multi-Erlang theory, as in [3]. Indeed, if we keep the same classification of the cell into zones of equal radio conditions, let P_i^{v} be the power consumed by a voice user at position i, the amount of power consumed by voice users when there is a vector of $\vec{n}^{\text{v}} = (n_1^{\text{v}}, ..., n_K^{\text{v}})$ voice users in the cell is equal to:

$$P_{\text{v}}(\vec{n}^{\text{v}}) = \sum_{i=1}^{K} n_i^{\text{v}} P_i^{\text{v}} \leq P_{\max} - P_{\text{cont}}$$

where $P_{\max}$ is the maximal transmitting power of the base station and P_{cont} is the power reserved for control channels. Multi-Erlang analysis (using Kaufman-Roberts algorithm [17][20]) allows computing the blocking rate for voice users and their average consumed power $\bar{P}_{\text{v}}$. The remaining available power for HSPA traffic is thus $P_{\max} - P_{\text{cont}} - \bar{P}_{\text{v}}$. This has an impact on the throughput that can be achieved by HSPA users (the achievable throughput is lower than the case of a dedicated carrier for HSPA), and this reduction can be taken into account by a shift of the CQI distribution by a number of dBs equivalent to the power reduction.

In order to take into account the impact of voice traffic on the performance, we make use of the following OMC counters:

- The voice traffic volume (in Erlang).
- The maximal cell power.
- The common channels power.

As of the powers consumed per voice call in different positions in the cells, it is not given by any field counter. We thus make use of the simulation results reported in [3] for estimating them.

E. Adaptation of measurements to analytical models

After describing the data that is available in the field, we now describe how these measurements can be used as inputs for the analytical models of Sections II and III.

First, we note that the analytical models described in Section II take as input a set of achievable throughputs, $\{C_i\}$, with their associated weights p_i. However the inputs collected from measurements give a distribution of CQIs, as explained above. In order to obtain the achievable throughputs, we make use of Transport Block Size (TBS) tables provided by the 3GPP, that associate each CQI to a volume of traffic per TTI (or equivalently a throughput in bit/s), for each device category [24]. Combining the device mix observed on the field with the CQI mix, the achievable throughput distribution can be obtained. As of the p_i's, they correspond to the weight associated to each CQI and represent the percentage of users that see the different CQIs. They correspond thus to a proportion of offered volume per radio condition.

Table II shows the measurement inputs used by the different analytical models and the outputs that can be compared to the KPIs observed on the field. Both infinite and finite source models take as input the radio parameters (CQI and device mix, powers). However, only one HSPA traffic parameter – the traffic volume – is needed for the infinite source models, while three traffic parameters are needed for the finite source models (average session volume, average time between sessions and number of users that are present in the cell). Some other

TABLE II
INPUTS/OUTPUTS OF ANALYTICAL MODELS ADAPTED TO THE FIELD MEASUREMENTS

Field parameter	Input for	Output for
CQI distribution	Finite and infinite source	-
Device mix	Finite and infinite source	-
Traffic volume	Infinite source	Finite source
$X_{\rm on}$	Finite source	-
$T_{\rm off}$	Finite source	-
$N_{\max}$	Finite source	-
Cell load	-	Finite and infinite source
Avg. nb. of active users	-	Finite and infinite source
Avg. user throughput	-	Finite and infinite source
Voice traffic volume	Finite and infinite source	-
Cell powers	Finite and infinite source	-

KPIs observed on the field (average cell load, number of active users, average user throughput) are used to compare the models outputs to the field.

V. VALIDATION RESULTS AND DISCUSSIONS

A. Infinite source models

We begin by presenting the results obtained with infinite source models. The models are tested on a set of 7 cells, selected so that they represent the typical traffic load values observed in the network during the measurement hours. Figures 6 and 7 present the scheduler loads and the numbers of active users, respectively. The x axis of these figures represents a fictitious cell ID, where cells are sorted following their increasing observed loads. Three values are plotted for each of the cells corresponding to the field measurements, the result obtained using a harmonic average of throughputs as the cell capacity and the result obtained using an arithmetic average (with the same weighting probabilities, as suggested in [16]).

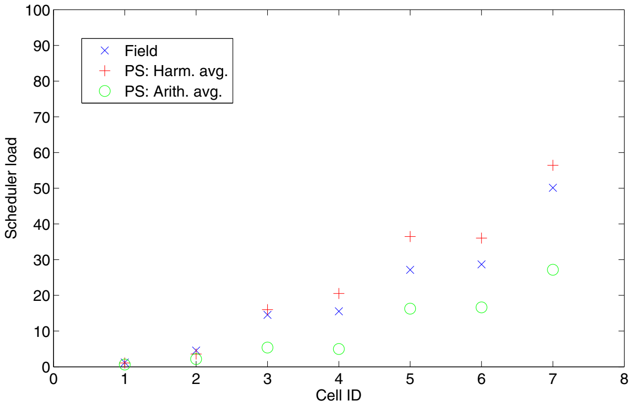

Fig. 6. Validation results for infinite source models: cell loads

The first observation is that the harmonic average gives in general acceptable results that are close to the field measurements, for both loads and average number of active users. The second observation is that the arithmetic average method gives optimistic results (low loads and small numbers of active users). We explain this by the fact that the majority of the traffic in the considered HSPA network (an in mobile networks in general) is generated by indoor users or by nomadic ones with low mobility. The no mobility assumption is thus closer to the reality than the high mobility assumption.

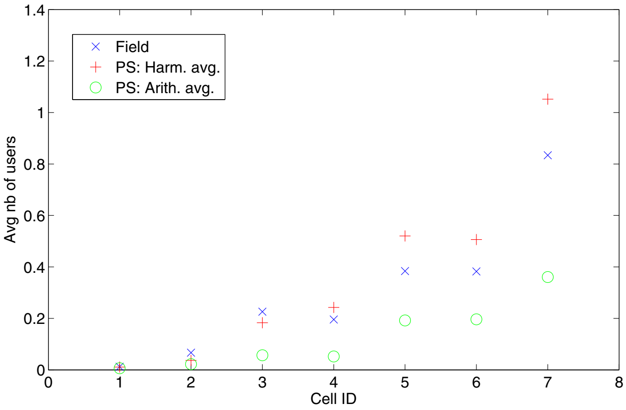

Fig. 7. Validation results for infinite source models: number of active users

However, there is still some gap between the field observations and the analytical results with a harmonic average; we can give some insight into the possible causes of this inaccuracy:

- Mobility: Even if the majority of traffic is generated by non moving users, some of the traffic is generated by users in mobility. Neither the harmonic average nor the arithmetic one are able to describe the performance. However, the results are closer to the harmonic average method as there are more static users than mobile ones.
- Inaccuracies in the inputs: Even if we tried to stick the best to the field measurements for the inputs of the analytical models, there are still some inputs values that we obtained from simulations as they cannot be directly observed on the field. This includes scheduling gains (Table I) and the power consumed by a voice user.
- Equipment specificities: Our models consider generic radio resource management (scheduling, throughput association for each of the reported CQIs, etc.) that does not take into account the specificities of the different vendor implementations.
- Devices with advanced receivers and receive diversity: The device mix reported by the OMC counters does not include the information about receivers and receive diversity. We thus make a conservative assumption of standard receivers with no receive diversity, which tends to give pessimistic results.

B. Finite source models

We now move to the finite source model. We consider the model presented in equation (13) for two values of the number of users N: the upper bound (all users) and the number of users that made at least two data connections during the observation period. Figures 8 and 9 present, respectively, the scheduler loads and the numbers of active users, and follow the same presentation as before.

We first observe that the gap between results obtained when considering all users versus the performance when filtering on

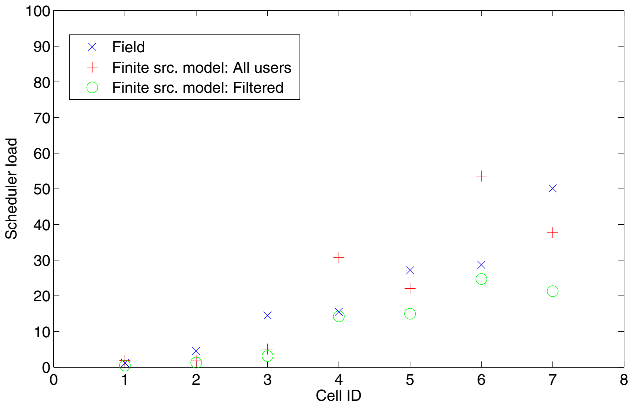

Fig. 8. Validation results for finite source models: cell loads (the same cells as before are used).

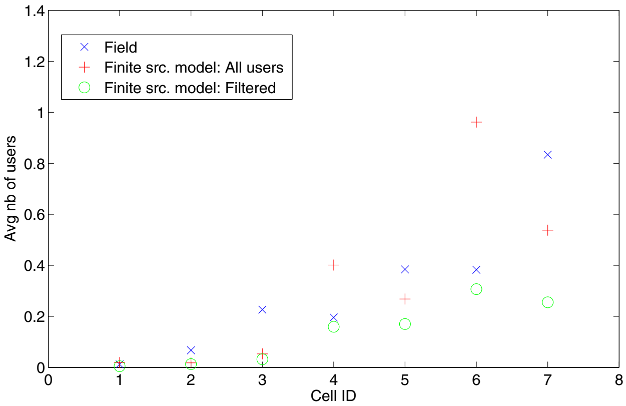

Fig. 9. Validation results for finite source models: number of active users

users with multiple connections is high: Taking all users, even those that make only one connection during the observation hour, increases the calculated cell load. The second observation is that when there is a match between the field observation and the prediction, the filtered case is the closest one; this is the case of cells 1, 4 and 6. For the other cells, we observe that both values (all users and filtered connections) give optimistic results compared to the field (lower loads and lower number of users). This may result from a sub-estimation of the other traffic parameters: the volume per connection X_{on} and the OFF duration T_{off}. We indeed based the X_{on} and T_{off} calculations on the RNC basis and not on a cell per cell basis, arguing that the behavior of users is not cell-dependent and that the differentiating parameter between cells is the number of users.

In order to investigate this issue of user's behavior that is different in some cells, we focus on cell 7 and extract the traffic characteristics of users that are connected to this cell (the results are not shown in Figures). We obtain an average connection size X_{on} of 315 Kbytes, compared to the RNC average of 243 Kbytes per connection, while the T_{off} value remains almost stable. Applying this new traffic characteristic to this cell, we obtain a load of 50%, very close to the 56% observed on the field (this is equivalent to the result obtained with the infinite source model using the harmonic average). This experiment illustrates the fact that the finite source models, if tuned on a cell per cell basis, can give good results that fit well to the field.

These results illustrate the difficulty of calibrating finite source models with field measurements. Indeed, while only the traffic volume V is needed to be extracted on a cell-by-cell basis in infinite source models, three parameters are needed in finite source models: X_{on}, T_{off} and N. Note however that analytical models with finite sources are only sensitive to the ratio $\frac{X_{\text{on}}}{T_{\text{off}}}$.

VI. Conclusion

The objective of this paper is twofold: modeling the cell capacity and comparing the two main sets of models for performance evaluation in mobile networks. We followed an empirical approach that consists of feeding the analytical models with inputs from a live network, and comparing the obtained performance metrics with observed performance on the field.

As of the first objective, we considered the two definitions of capacity in the literature that take into account the heterogeneity of radio conditions over the cell. We namely consider as capacity the arithmetic average of achievable throughputs, versus their harmonic average, and highlight the fact that, before comparing them, one must precisely characterize the probabilities that are involved in the averaging. We show that the harmonic average corresponds to the static case, while the arithmetic average is more related to the high mobility case. Our experiments show that the analytical results match better to field observations when the capacity is considered as the harmonic average of the achievable throughputs, as recommended in [8]. This is due to the fact that a large proportion of the traffic in mobile networks is generated by indoor or fixed users. For example, [25], Table 13, states that 80% of traffic in dense urban environments is generated by indoor users.

For the model itself, we studied the infinite source models where the arrival rate of connections is supposed to be independent of the network state, and the finite source models where the number of users that are physically present in the cell is limited so that the arrival rate reduces when the number of active users increases. We presented the models that are available in the literature for infinite source models and derived novel finite source queuing models that correspond to the case of a low mobility. We showed that finite source models are pretty difficult to calibrate with field measurements, due to a larger number of needed traffic inputs on a cell per cell basis, while infinite source models are simpler to calibrate because they only take into consideration the traffic volume. However, finite source models are able to capture the impact of user behavior and are useful, for instance, in capacity/technology planning based on marketing predictions. In this context, finite source models are very useful for predicting the impact on the performance of the introduction of new services/devices, as these latter usually introduce changes in user behaviors (activity patterns, connection volumes, etc.).

References

[1] N. Abbas, Y.-T. Lin and B. Sayrac, *Mobility-driven Scheduler for Mobile Networks Carrying Adaptive Streaming Traffic*, in 2016 IEEE PIMRC, Sep. 2016.

[2] K. L. H. Asplund and P. Okvist, *How typical is the typical urban channel model? Mobile-based delay spread and orthogonality measurements*, IEEE VTC Spring, 2008.

[3] A. Baroudy and S. Elayoubi, *HSUPA/HSDPA systems: capacity and dimensioning*, IEEE FGCN 2007.

[4] B. Baynat and N. Nya, *Performance Model for 4G/5G Networks Taking Into Account Intra- and Inter-Cell Mobility of Users*, 41st IEEE International Conference on Local Computer Networks (LCN 2016), Dubai, UAE, November 2016.

[5] B. Baynat, G. Nogueira, M. Maqbool and M. Coupechoux, *An efficient analytical model for the dimensioning of wimax networks*, IFIP Networking 2009.

[6] B. Baynat, *Analytical Models for Dimensioning of OFDMA-based Cellular Networks Carrying VoIP and Best-Effort Traffic*, International Journal of Computer Networks (IJCN), Volume (4): Issue (4): 2012.

[7] F. Berggren and R. Jantti, *Asymptotically fair transmission scheduling over fading channels*, IEEE Transactions on Wireless Communications, Jan. 2004.

[8] T. Bonald and A. Proutière, *Wireless Downlink Data Channels: User Performance and Cell Dimensioning*, ACM Mobicom'03, Sep. 2003.

[9] T. Bonald, S. Borst and A. Proutière, *How Mobility Impacts the Flow-Level Performance of Wireless Data Systems*, IEEE INFOCOM 2004.

[10] T. Bonald and M. Feuillet, *Network performance analysis*, Wiley, August 2011.

[11] S. Borst, *User-Level Performance of Channel-Aware Scheduling Algorithms in Wireless Data Networks*, IEEE/ACM Transactions on Networking (TON), Vol. 13, pp. 636-647, June 2005.

[12] E. Brockmeyer and H. L. Halstrom, *The Life of A. K. Erlang*, Transactions of the Danish Academy of Technical Sciences, 1948, No. 2, pp. 1618.

[13] J. W. Cohen, *The generalized Engset formulae*, Philips Telecommun. Rev., Vol. 18, pp. 158-170, 1957.

[14] R. Combes, S. Elayoubi and Z. Altman, *Cross-layer analysis of scheduling gains: Application to LMMSE receivers in frequency-selective Rayleigh-fading channels*, IEEE WiOpt 2011.

[15] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula and D. Estrin, *A first look at traffic on smartphones*, ACM SIGCOMM conference on Internet measurement, 2010.

[16] J. Gora and S. Redana, *Resource management issues for multi-carrier relay-enhanced systems*, EURASIP Journal on Wireless Communications and Networking , March 2012.

[17] J. Kaufman, *Blocking in a Shared Resource Environment*, IEEE Transactions on Communications, 1474-1481, 1981.

[18] A. Khlass, T.Bonald and S. Elayoubi, *Performance Evaluation of Intra-Site Coordination Schemes in HSPA+ Networks*, Performance Evaluation, Elsevier, February 2016.

[19] Y.-T. Lin, T. Bonald and S. Elayoubi, *Impact of Chunk Duration on Adaptive Streaming Performance in Mobile Networks*, IEEE WCNC 2016, April 2016.

[20] J.W. Roberts, *A service system with heterogeneous user requirements*, in: G. Pujolle (Ed.), Performance of Data Communications Systems and Their Applications, North-Holland, Amsterdam, 1981, pp. 423-431.

[21] I. Trestian, S. Ranjan, A. Kuzmanovic and O. A. Nucci, *Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network*, Internet Measurement Conference (IMC 2009), Chicago, USA, 2009.

[22] L. Rong, S. Elayoubi and O. Ben Haddada, *Performance Evaluation of Cellular Networks Offering TV Services*, IEEE Transactions on Vehicular Technology, 2010.

[23] 3GPP TR 25.700, *Study on Further Enhanced Uplink (EUL) enhancements*.

[24] 3GPP TR 25.855, *HSDPA; Overall UTRAN Description*.

[25] 5G PPP association, *5G-PPP use cases and performance evaluation models*, May 2016, available at https://5g-ppp.eu/.