Computer Communications 108 (2017) 27–35

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

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

Salah Eddine Elayoubi [a], Younes Khadraoui [a], Bruno Baynat [b,*], Taoufik En-Najjary [a]

[a] Orange Labs, Issy-les-Moulineaux, France
[b] Sorbonne Université, UPMC University Paris 06, CNRS, LIP6 Laboratory, Paris, France

ARTICLE INFO

Article history:
Received 17 June 2016
Revised 18 April 2017
Accepted 28 April 2017
Available online 5 May 2017

Keywords:
Mobile networks
Analytical modeling
Performance evaluation
Empirical validation
QoS and capacity estimation

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.

1. 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 Bonald and Proutière [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 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

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

* Corresponding author.
E-mail address: bruno.baynat@lip6.fr (B. Baynat).

http://dx.doi.org/10.1016/j.comcom.2017.04.011
0140-3664/

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 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 2, we present the infinite source queuing models. Section 3 deals with the finite source queuing models and extends the theory to the low mobility case. Section 4 shows how the analytical models can be adapted to take as inputs network measurements. Section 5 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 6 concludes the paper.

2. 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. Bonald and Proutière [8] and show how that can be applied to an arbitrary number of classes of radio conditions.

2.1. 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

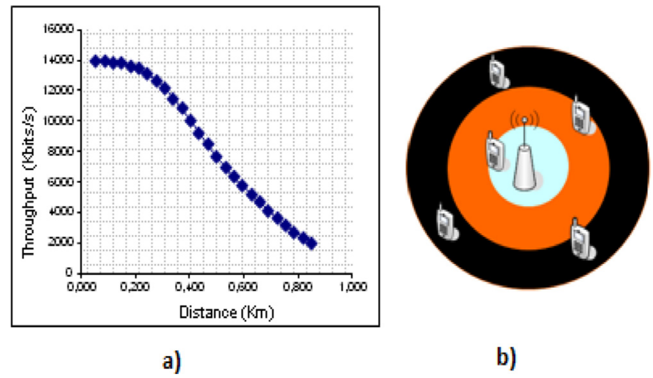

Fig. 1. a) Example of simulated achievable throughput over an HSPA cell: only one user is in the cell but is simulated at different positions in the cell. b) Cell decomposition into zones of equal radio conditions.

to the case where he is far from it, as illustrated in Fig. 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, $i = 1, \ldots, K$. 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.

2.2. 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]:

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

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

$$\bar{C} = \frac{1}{\sum_{i=1}^{K} \frac{p_i}{C_i}} \tag{2}$$

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