A novel hybrid simulation methodology for capacity estimation in mobile networks

S.E. Elayoubi [a], M.K. Karray [a], Y. Khan [a], S. Jeux [a]

[a] *Orange Labs*
38, rue du Général Leclerc
92794 Issy les Moulineaux, France
salaheddine.elayoubi@orange.com, Mohamed.karray@orange.com,
yasir.khan@orange.com, sebastien.jeux@orange.com

Abstract

Radio capacity simulation tools are gaining a large importance with the development of mobile networks. System radio simulators that are currently used in standardization bodies are becoming increasingly complex as they have to work on two time scales: the scale of "milliseconds" for modeling the behavior of schedulers and the scale of "tens of seconds" for modeling the dynamic behavior of arrivals and departures of users. In this paper, we propose a hybrid system simulation methodology that combines the advantages of system simulators in accurately modeling the physical/MAC interfaces, with the robustness of queuing theory analysis that catches the flow dynamics. We validate our simulation methodology versus complete system level simulators in representative scenarios and show an excellent match between both methodologies. We then show how to extend our simulation methodology for including a mix of services and how to incorporate network measurement results within the proposed methodology.

Key words: radio capacity, system simulations, queuing theory.

1 Introduction

The expansion of mobile networks accelerated in the last few years with the evolution of 3G networks towards 3G+, LTE and LTE-Advanced [1]. In this context of burdening standardization, robust radio capacity simulators are needed in order to assess the performance of proposed features and systems. Common simulation methodology is system simulation that consists in simulating a fixed number of users under an almost complete modeling of the

physical and MAC layers (see [1], section A2.1.3). This simulation methodology, called the full buffer traffic model, has the advantage of being highly accurate when modeling the lower layers as it takes into account a complete channel model including path loss, shadowing and fast fading. However, as it considers full buffer traffic, i.e. users that are continuously present in the network and have always traffic to send, this model fails to consider the dynamic behavior of users. Indeed, in realistic settings, users are not continuously transmitting; they arrive to the system, download a file, make a phone call or watch a video before leaving the system. It is thus essential to model this dynamic behavior for users if we want to model realistic services.

As the primary aim of operators is to ensure a good Quality of Service (QoS) for their clients, a dynamic simulation methodology taking into account realistic services is needed. The Next Generation Mobile Networks (NGMN) alliance, leaded by operators, pushed 3GPP to adopt a new evaluation model that extends the full buffer simulation model, by considering users that arrive to the system with the aim of downloading a file of fixed size and leave it after achieving this task. The system is thus simulated for different traffic loads, defined by increasing the (Poisson) arrival rate of users. This model has been adopted by the 3GPP as a complementary method for evaluating the system, starting from Release 9 (see FTP simulation model in [1], section A2.1.3). This new simulation methodology is suitable for data services, as it models accurately the system starting from the physical layer, including MAC layer (HARQ, scheduling and even admission control) and reaching the higher layers (session layer).

Although the FTP simulation methodology is the most suitable for estimating the QoS for data flows, it is complex to implement and long to converge. Indeed, this methodology operates at two different time scales: the scale of milliseconds for PHY/MAC layer mechanisms (coherence time of the channel), and the scale of tens of seconds for the session level (dynamic behavior of users). In this paper, we propose a new hybrid simulation methodology that is as accurate as system simulations but is faster and more flexible. The idea resides in relying on system simulations for generating the distribution of radio conditions over the network under realistic channel models, and using these distributions as inputs for a queuing theory analysis that takes into account the session (flow) level. The advantages of our proposed methodology are the following:

(1) Accuracy: the proposed methodology is accurate when modeling the PHY/MAC layers as it relies on extensive system simulations.
(2) Quick simulations: the use of queuing theory for modeling the flow level makes the simulations extremely rapid as many closed form expressions and efficient analysis algorithms are available in the queuing theory literature [2][3].

(3) Easy implementation: The proposed methodology can be easily implemented on top of existing system level simulators.
(4) Mix of services: Our methodology is able to integrate constant bit-rate (CBR) and variable bit-rate (VBR) services, based on recent advances in queuing theory that allow integrating multimedia services [5][4].
(5) Mix of devices and environments: in a realistic environment where users have different device capabilities (device category, receivers, etc.) and may be outdoors or indoors, our methodology allows to take into account this heterogeneity in a flexible way.

The remainder of this paper is organized as follows. In Section II, we recall the system simulation methodology developed in 3GPP and NGMN bodies. Section III presents our approach for simulating data traffic, that is validated with system simulations. Section IV shows how to extend the proposed hybrid methodology to a mix of best effort (variable bit rate) and streaming (constant bit rate) services and how to integrate field measurements in it. Section V eventually concludes the paper.

2 Preliminaries on system simulators

In this section, we survey the system simulation methodologies proposed in 3GPP and NGMN.

2.1 Static simulators

Classical full buffer simulations consist in the following steps:

(1) Network setting: The simulated network is in general a homogeneous network composed of hexagonal tri-sectored cells, as shown in Figure 1. The traffic is generated everywhere in the network. 57 cells are at least simulated in order to accurately model the interactions between cells, and the network is virtually reproduced at the edge in order to avoid the border effects.
(2) Traffic generation: In full buffer simulations, a constant number of users is simulated in each of the cells. This is done by generating a number of users (typically 10 users by cell) and associating to each of them a path loss (calculated following a path loss model as defined in [1], section A2.1.1) and a log-normal shadowing variable [6].
(3) Resource allocation: For a given configuration of users (number and positions), a large number of TTIs (Transmission Time Intervals) is simulated. At each TTI, the channel is generated for each of the users, using

standard fast fading models (for IMT-Advanced models, see [1] section B). This channel, combined to the path loss and shadowing values and the interference calculated from other cells gives the SINR (Signal to Interference plus Noise Ratio), used as a basis to estimate the scheduling metric (e.g. using proportional fair scheduler) and to allocate resources to users. This operation is repeated at each TTI.

(4) Convergence analysis: steps 2 (traffic generation) and 3 (resource allocation) are repeated a large number of times until the throughputs converge, ensuring that all possible positions in the cell are covered.

(5) Throughput calculations: Once the convergence is ensured, the throughput CDF (Cumulative Distribution Function) is constructed over the network. In addition to the average throughput, two values from the CDF are of particular interest: the cell center throughput (95% percentile) and the cell edge throughput (5% percentile).

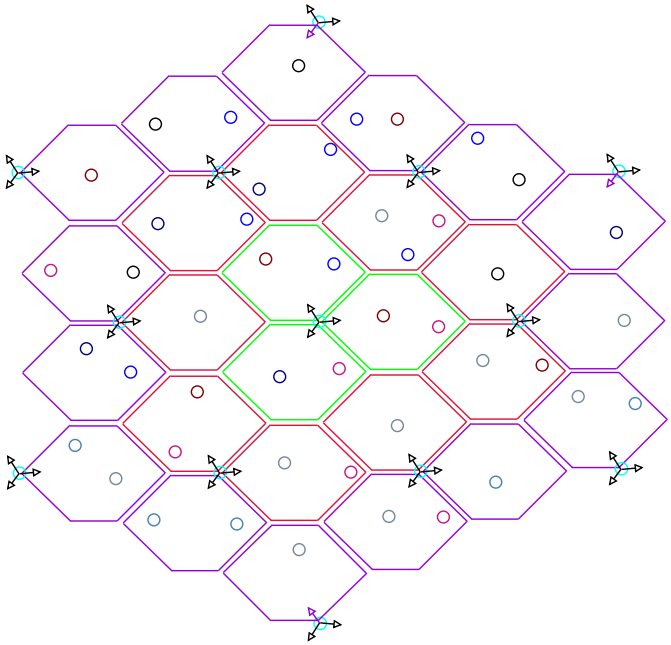

Fig. 1. Network layout.

2.2 Dynamic simulations

2.2.1 Description

When dealing with VBR data traffic, simulating a fixed number of users in a uniform way over the cell is not realistic. Indeed, users that have bad radio conditions (low throughputs) will stay longer in the network, leading to a larger number of users at cell edge. 3GPP FTP model fills this gap by simulating uniform Poisson arrivals to the cell, but with users staying in the system

until they transmit a file of predetermined size [1]. Users at cell edge will naturally stay longer in the system and contribute more to the cell load. By reference to the steps describing full buffer simulations in section 2.1, traffic generation in step 2 is not performed for a fixed number of users, but users are progressively introduced to the system with an exponential inter-arrival time whose parameter is inversely proportional to the traffic load. Once a user finishes transmitting his file, he quits the system. Note that, even for high traffic loads, a large number of TTIs is present between two flow level events (arrival or departure of a call), as a TTI (1 or 2 ms) is very small compared to inter-arrival times and file transfer times. Two different flavors are proposed:

- FTP model 1 considers Poisson arrivals with a predetermined file size of 0.5 or 2 Mbytes.
- FTP model 2 simulates a constant number of users in parallel, each generating an http-like traffic (files of 0.5 Mbytes, with an average reading time of 5 seconds). The number of users is thus increases to simulate a higher load.

Note that the main difference between dynamic (FTP) and full buffer simulation methodology is that the former simulates static uncorrelated snapshots of the systems, while the former simulates the dynamic behavior of the system by correlated snapshots taking into account the actual workload brought by each user.

2.2.2 Capacity definition

As discussed above, FTP simulation model calculates the throughput average and percentiles for different offered traffics. The question is thus how to define the capacity of the system based on the different simulation loads. Two ways are possible:

(1) Capacity for a target load: The first way is to define capacity as the offered traffic that generates a target load, e.g. 70% resource utilization. This definition is useful for comparing two systems: System A is more efficient that system B if the former reaches the target load with a higher traffic than that of the latter.
(2) Capacity for a target QoS: A more classical capacity definition (compliant with Erlang-B law) is the traffic that achieves a target QoS. In this case, multiple objectives can be set up, for instance the average user throughput is larger than 1 Mbps and the cell edge throughput is larger than 0.5 Mbps.

3 Proposed hybrid methodology

3.1 Methodology description

As explained above, FTP simulations are necessary for catching the dynamic behavior of users. However, the complexity of this approach makes simulations too time consuming (in the order of several days for obtaining one simulation point on a standard computer). Relying on these simulations for evaluating capacities for mixes of mobile categories becomes quickly unfeasible. Our hybrid simulation methodology relies on queuing theory analysis in order to reproduce the dynamic behavior of users. This is summarized in Figure 2 and consists in the following steps:

(1) Using full buffer simulations, construct a database of peak throughput distributions for the different devices (mobile categories, receivers, receiver diversity, etc.) and for the different possible situations (indoor/outdoor/incar). No device/environment mixes are needed for this step.
(2) For a given mix of devices and situations, use these databases as an input for the queuing theory analysis, as explained in the next section 3.2.

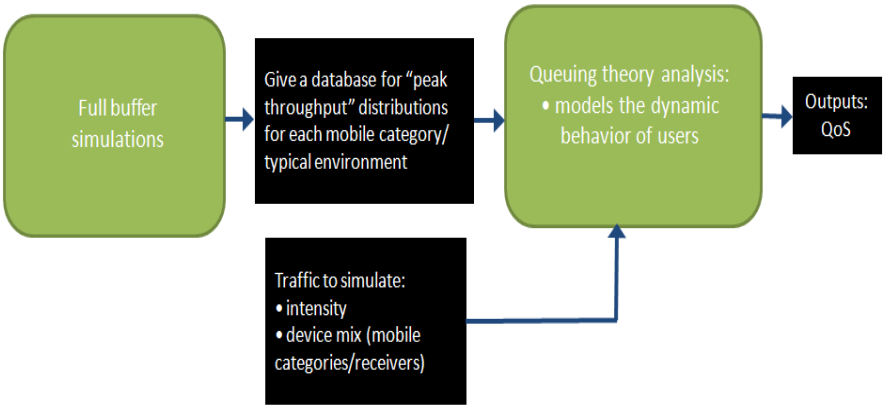

Fig. 2. Hybrid simulation methodology.

3.2 Queuing theory analysis

We aim now to study the relation between the peak bit-rate distribution and the performance of the system in a dynamic context (i.e., when users arrive and depart from the network), and specifically the QoS perceived by users.

We consider the so-called *variable bit-rate* (VBR) calls; i.e., connections aiming

to transmit some given volume of data at a rate that may be decided by the network (e.g. mail, http, ftp). The traffic demand in a given geographic zone may be defined as the average volume of data per call divided by the average duration between two successive call arrivals. When we account for calls arrivals and departures, the QoS perceived by VBR users is the *mean throughput* (or alternatively *delay*) averaged over the long run of the system. We shall give the expression of the user's *mean throughput* as function of the traffic demand per cell and the peak bit-rates distribution. For VBR calls, the main QoS indicator is the *outage probability*; i.e., the proportion of users in the network that do not get their required bit-rate [11].

The set of positions of all the users served at a given time is called *configuration of users*. If the process describing the evolution in time of the users configuration is not ergodic, then the mean number of users in the system grows unboundedly in the long run of the system (and consequently, the delay of each user goes to infinity and his throughput goes to zero). This *unstable* situation has to be avoided; that is the operator should guarantee that his network is *stable*. We shall also give the condition guaranteeing stability.

Let $(R_1, R_2, \ldots, R_J)$ be the set of peak bit-rates where a given $j \in \{1, 2, \ldots, J\}$ may be seen as a *location* in the cell. The peak bit-rate distribution is a vector $(p_1, p_2, \ldots, p_J)$ where p_j is the probability associated to R_j (of course $\sum_{j=1}^{J} p_j = 1$). The values of $\{(R_j, p_j)\}_{j=1,2,\ldots,J}$ are obtained from the system simulator.

Let ρ be the traffic demand (in bits/s) per cell. We distribute this cell traffic over the different locations by defining the traffic demand at location j as $\rho_j := p_j \rho$. We shall assume that the base station gives to each user an equal portion of time.

Let $X_j(t)$ be the number of user al location $j \in \{1, 2, \ldots, J\}$ at time t and $X(t) = (X_1(t), X_2(t), \ldots, X_J(t))$. The evolution of the system is described by the process $\{X(t)\}_{t \geq 0}$. We assume that the durations between succesive arrivals to the network are i.i.d. expenentially distributed random variables. The following proposition characterizes the evolution of the process $\{X(t)\}_{t \geq 0}$ in the long run. If the traffic demand is too high, then as time goes to infinity, the number of users keeps growing and the throughput of each user vanishes. We say then that the system is *unstable*. The following proposition gives the value of the traffic demand which should not be exceeded to assure stability, and the throughput per user in the cell at the stationary state in the stability case.

Proposition 1 *The cell is stable when the traffic demand per cell ρ does not*

exceed some critical *value:*

$$\rho < \rho_{\mathrm{c}} := \left[\sum_{j=1}^{J} p_j R_j^{-1} \right]^{-1}$$

In case of stability, the throughput per user in the cell at the stationary state is given by

$$\bar{r} = \rho_{\mathrm{c}} - \rho \tag{1}$$

PROOF. This results follows from known results on multi-class processor sharing queues; see e.g. [7], [3].

Note that the critical traffic demand ρ_{c} is the *harmonic* mean of the peak bit-rates. For the effect of the user's mobility on its QoS we refer the reader to [8] where it is observed that when users move, the critical traffic demand is the *arithmetic* mean of the peak bit-rates and the throughput per user is no longer a simple linear function of the traffic demand as in Equation (1).

We define the *queuing load* of the cell by

$$\theta = \frac{\rho}{\rho_{\mathrm{c}}}$$

We shall compare in the numerical section this queuing load to the *simulated load* defined by the system simulator as the portion of time when there is at least a user in the cell in which case the base station transmits at its maximal power (otherwise the base station's power vanishes). The validation consists of checking that the numerical values of these two loads are close.

3.3 Validation for VBR services

In order to validate our approach, we conduct system simulations for LTE in order to extract Resource Block (RB) throughput distribution [1]. We show in figure 3 an example RB throughput distribution for obtained in the band of 2.6 GHZ, with a MIMO 2*2 scheme and an LMMSE receiver. We then perform full FTP system simulations and compare the obtained capacity with the result of our hybrid approach. As explained in section 2.2.2, the notion of capacity we use is the maximum offered traffic that generates a load equal to $x\%$, with x varying from 5% to 70%. Figure 4 shows a perfect match between capacities obtained from full FTP and hybrid simulations. Figure 5 shows the

[1] By extracting resource block throughputs, we ensure more flexibility for our simulator, as we can model any carrier size by changing the number of RBs.

expected drop in throughput, as the traffic is increased. We show in Figure 6, the impact of increasing the expected target throughput on the offered traffic for a fixed outage: the offered traffic must be reduced to achieve the increase in the expected target throughput.

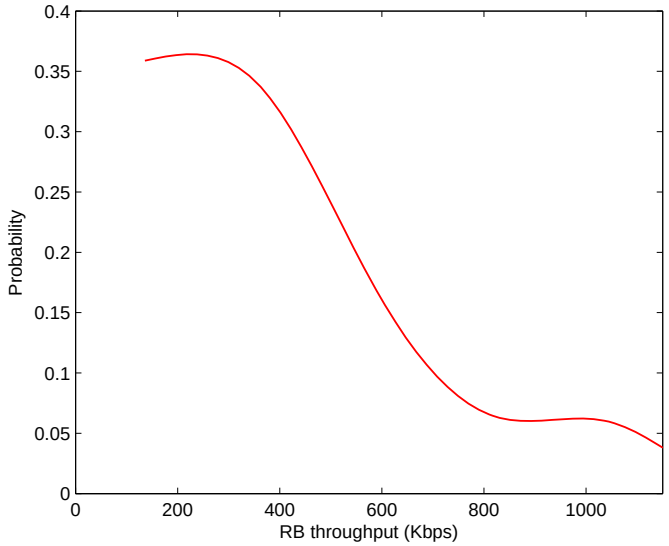

Fig. 3. RB throughput distribution over a typical cell.

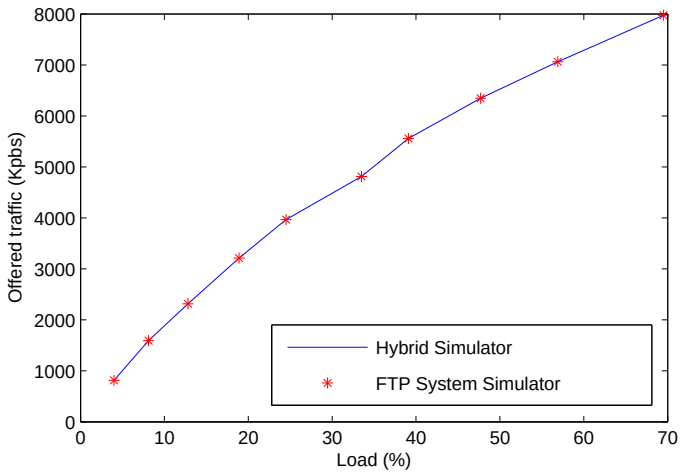

Fig. 4. Comparison of the capacities obtained by FTP full simulations and our hybrid methodology; a perfect match is observed.

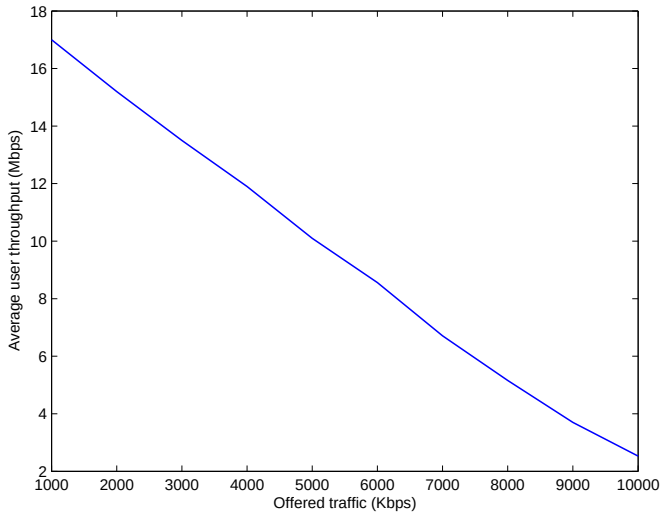

Fig. 5. Variation of the user throughput with offered traffic.

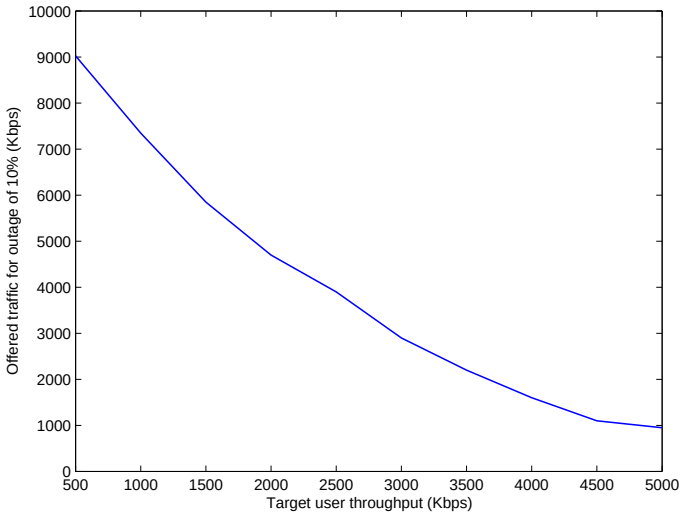

Fig. 6. Variation of the target user throughput with offered traffic for outage of 10%.

4 Extensions

4.1 CBR services

We consider now the so-called *constant bit-rate* (CBR) calls; e.g., voice, video conferencing, streaming. Users require some given (constant) bit-rate for some duration. In this case the requested bit-rates may sometimes exceed the available capacity, a situation usually called congestion. CBR services do not tolerate temporary interruptions of their transmissions. Consequently, if congestion

occurs, the network blocks (i.e., refuses the access to) new calls.

For CBR calls, the main QoS indicator is the *blocking probability*; i.e., the proportion of call arrivals which are denied from accessing the network. We shall describe how to calculate the blocking probability for a given traffic demand and peak bit-rate distribution.

Let $(x_1, x_2, \ldots, x_J)$ be the number of users at the different locations. Let r be the bit-rate requested by CBR calls. The base station should allocate to each user at location j a portion of time equal to

$$\frac{r}{R_j}$$

where R_j is the peak bit-rate at location j. Writing that the sum of these time portions should not exceed 1, we get

$$\sum_{j=1}^{J} x_j \frac{r}{R_j} \leq 1 \tag{2}$$

which is considered as the *admission condition*.

Let ρ be the traffic demand (in Erlang) per cell. In order to account for the peak bit-rate distribution $(p_1, p_2, \ldots, p_J)$, we distribute the cell traffic demand over the different locations by defining the traffic demand at location j as $\rho_j := p_j \rho$.

Is is shown in [10] that the blocking probability at a given location is equal to the probability that the free (without blocking) process with a user at this location does not satisfy (2), given that the process satisfies (2). Note that the admission condition (2) has the so-called *multi-Erlang* form, thus we may calculate the blocking probability using the *Kaufman-Roberts algorithm* [12,13]. Figure 7 shows the blocking probabilities of CBR (streaming calls with a target throughput of 256 Kbps) when the offered traffic increases. It can be observed that, for a target blocking probability of 5%, the system capacity is equal to 30 Erlang.

When a mix of CBR and VBR traffic is considered, priority is usually given to CBR calls so that VBR calls will share the capacity left by CBR ones. Queuing theory results are available for this case, as shown in [5,9]. We show in figure 8 the impact of increasing CBR traffic on the performance of VBR calls: when CBR traffic increases, QoS is severely degraded for VBR as the former has strict priority over the latter.

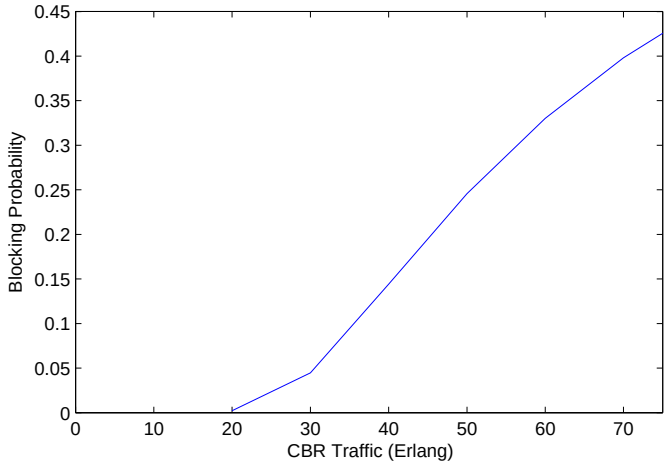

Fig. 7. CBR blocking probabilities.

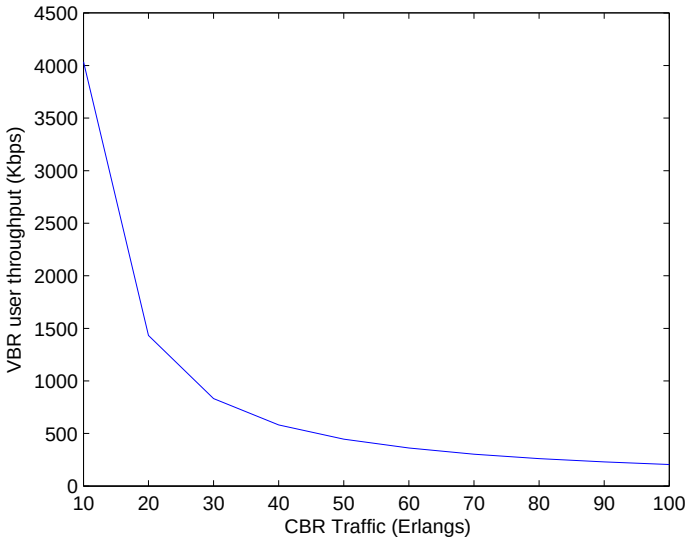

Fig. 8. VBR average user throughput for a mix of CBR and VBR calls: VBR traffic demand is fixed to 0.5 Mbps while CBR traffic increases.

4.2 Taking into account measurement data

The hybrid methodology presented till now relies on system level simulators to produce the distribution of radio conditions, that are used as inputs for the queuing theory analysis. However, for already-deployed systems, operators have an important source of information that is measurement data originating from live networks. Ideal simulation tools must thus be able to be calibrated with measurement data. This is the case of our proposed hybrid methodology where the queuing theory part takes as input the throughput distribution.

As example, we consider real network drive test measurements. The mea-

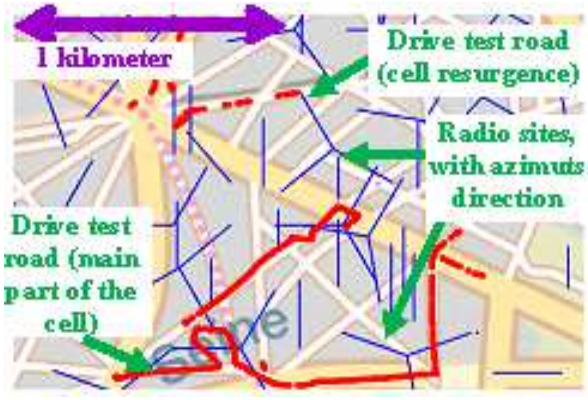

Fig. 9. Map of the drive test for the measured cell.

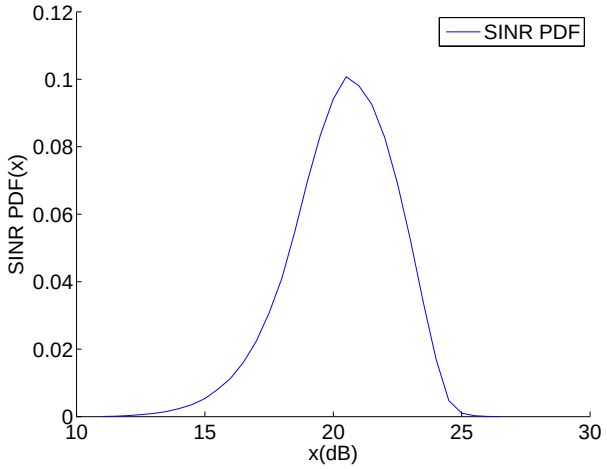

Fig. 10. distribution of the average SINR over the cell.

surement campaign has been run in the center of a major European town. Measurements are available for hundreds of cells, and one typical cell is selected for the capacity calculations (measurement samples are filtered with the scrambling code of the selected cell). This cell is composed of several (non contiguous) parts, which is common in live networks, especially in urban areas, because of irregular propagation from both the serving cell and the many interfering cells around, and because of the irregular cell planning. 1000 measurement samples are available for this cell. The average CPICH (Common PIlot CHannel) Ec/I0 on the cell is -7db, and the average CPICH RSCP (Received Signal Code Power) is -65dBm. The map of the drive test is shown in Figure 9.

We thus obtain the probability density function (p.d.f) of the average SINR over the cell, shown in Figure 10. We can then obtain throughput distribution by associating this SINR distribution with link level curves. These latter

consist in look up tables giving the throughput for each SINR value, obtained from link level simulations as indicated in [1], section A.1. Once these throughput distributions are obtained, we apply our hybrid methodology using them as radio inputs.

5 Conclusion

We describe in the present paper a novel hybrid simulation methodology for quality of service (QoS) and capacity estimation in mobile networks. It consists of taking as inputs the peak bit-rates at each location. These peak bit-rates may for example be the outputs of some system level simulator such as the FTP model proposed by 3GPP.

We show how to use the queuing theory tools to deduce from these peak bit-rates the QoS perceived by the users. In particular for variable bit-rate (VBR) users, we give the explicit expression of the critical traffic (which is the limit of the traffic demand over which the system becomes unstable) as well as the throughput per user. We compare the capacities obtained from our proposed approach and the FTP dynamic system simulation approach and show a perfect match between the results.

For constant bit-rate (CBR) users, we show how to calculate the blocking probability given the peak bit-rates and the traffic demand. We illustrate numerically the whole approach in the cases of pure CBR traffic and in the case of a mix between CBR and VBR traffics.

Our proposed approach has the advantage of being accurate while being quick, easy to implement and flexible enough to take into account several types of services and inputs from field measurements. We believe that this approach has a large potential for estimating QoS and capacity in mobile networks.

References

[1] 3GPP TR 36.814 V9.0.0 (2010-03). *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release 9)*, March 2010.

[2] 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.

[3] T. Bonald and A. Proutiere, *Wireless downlink data channels*, ACM Mobicom 2003.

[4] Sem C. Borst, Nidhi Hegde, *Integration of Streaming and Elastic Traffic in Wireless Networks*, IEEE INFOCOM 2007.

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

[6] I. Forkel, M. Schinnenburg, and M. Ang, *Generation of two-dimensional correlated shadowing for mobile radio network simulation*, IEEE WPMC 2004, Abano Terme (Padova), Italy, Sep. 2004.

[7] J. W. Cohen, *On regenerative processes in queuing theory*. Springer Verlag, 1976.

[8] M. K. Karray, *User's mobility effect on the performance of wireless cellular networks serving elastic traffic*, Wireless Networks (Springer), 17(1), 2011.

[9] M. K. Karray, *Analytical evaluation of QoS in the downlink of OFDMA wireless cellular networks serving streaming and elastic traffic* IEEE transactions on Wireless communications, 2010.

[10] F. Baccelli, B. Błaszczyszyn, and M.K. Karray. *Blocking Rates in Large CDMA Networks via Spatial Erlang Formula*, In Proc. of IEEE INFOCOM, 2005.

[11] M. K. Karray and Y. Khan. *Evaluation and comparison of resource allocation strategies for new streaming services in wireless cellular networks*, In Proc. of IEEE ComNet, 2012.

[12] J.S. Kaufman, *Blocking in a shared resource environment*. IEEE Trans. Commun., 29(10):1474–1481, 1981.

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