Modelling and Simulation of Multi-rate Multi-service Traffic in the Presence of Mobility

Jesús M. Juárez, Rui R. Paulo, Eva Reguera San José, Fernando J. Velez, *Senior Member*
Instituto de Telecomunicações, DEM-University of Beira Interior
Calçada Fonte do Lameiro
6201-001 Covilhã, Portugal
jejuva@iespana.es, rfrpaulo@e-projects.ubi.pt, reguerasanjo@yahoo.com, fjv@ubi.pt

***Abstract*—A multi-service traffic model is proposed, and its validation is achieved by using event-based simulation results. In the single-service case, theoretical and experimental results for ON-OFF blocking probability are close to each other, and there is an almost perfect concordance between theoretical and simulation values when the average sojourn time in cells is equal to the average holding time. In the multi-service case, the behaviour is not exactly the same but a coherent behaviour is achieved for an average traffic per user up to 0.10 Erl.**

Keywords-multi-service traffic, simulation, handover, bursty behaviour, validation, mobility.

I. INTRODUCTION

Multi-rate multi-service traffic modelling is one of the most challenging issues involved in the dimensioning process of mobile multimedia communication systems, e.g., UMTS (Universal Mobile Telecommunications System) and its enhancements. On the one hand, models that consider queuing of data calls can be considered, which allows for obtaining results for delay. On the other hand, for real-time (and, in most of the cases, time-based applications), a simpler model can be used, which only models blocking and handover failure probabilities. This is the case of the Bernoulli-Poisson-Pascal/Markov-modulated Poisson Processes model from [1]. From previous work from the authors, besides considering multi-service, it represents the bursty behaviour of applications through the consideration of service components [2]. In this work, this multi-service traffic model is presented, and its validation is achieved by using event-based simulation results [3].

In Section II, after presenting the basis of the model, the user model and equivalent user are described, and details on how to compute the arrival rate are given. Section III presents the physical and mobility scenarios considered in the simulations, the concepts associated with the generation and termination of new and handover sessions, and the definitions of the quality of service parameters. Section IV presents results for blocking and handover failure probabilities, and also for the ON-OFF blocking probability, a measure for the blocking of bursts of traffic during multimedia traffic sessions. A comparison between theoretical and simulation results is performed, and conclusions are extracted for the multi-service traffic model validation. Finally, conclusions are drawn in section V.

II. TRAFFIC MODEL

A. Initial Considerations

To make an analysis of multi-service traffic in an advanced mobile communication network, e.g., Enhanced UMTS, a traffic model has to be defined. In the general model of a loss system with R_e=1 type of resources shared by J classes (i.e., service components), a customer arrival at the resources follows a specific random process. Each customer, i.e., service components users, requests a fixed number of resource units, i.e., channels, which are granted if available. If not, the request is cleared and the customer is blocked. The classification of customers is done on the basis of their arrival process, capacity requirement and mean holding time [1]. In this work, the performance measure that one is interested in is the probability that an arriving customer is blocked, i.e., the customer or connection blocking probability, P_b. Besides, this problem involves bursty traffic, with active and inactive, i.e., ON and OFF, periods, one needs to address ON-OFF blocking probability. Because of terminal mobility and the resulting handovers, one is also interested in the handover failure probability, whose limitation directly results from the existence of a threshold for the call-dropping probability.

The BPP (Bernoulli-Poisson-Pascal) model for the superposition of various types of traffic sources is being used here [1].

B. Basis of the Model

The capacity of the resource facilities is partitioned into capacity units. A customer is assumed to need a given number of units of each facility, and the demand is granted on a first come first served basis. If a customer demand cannot be granted, it is cleared and the new customer is blocked.

One considers J customer classes, each with different spatial and temporal requirements, and c is the total of available channels. The resource capacity vector is defined as $c_v = [c_1, \ldots, c_{Re}]$ but, since $R_e = 1$, one has $c_1 = c$. The class j (the term 'class' referring, in the context of this work, to 'service component') capacity demand of channels per customer, $a_j, j \in \vartheta$, where $\vartheta = \{1, \ldots, J\}$, and $a_j \in$ IN. Besides, the time that these channels, once granted, will be held by the service component (or class) j customer is i.i.d., it being specified by its mean value, whose specific distribution has no influence on the calculations that one is pursuing [1], [4]. Thus, given these considerations, the capacity vector, **A**, is a

vector of the following type

$$\mathbf{A} = \lfloor a_j \rfloor, j = 1,...,J \ . \quad (1)$$

Let the number of class j active customers, i.e., that hold their a_j resources at time t, be represented by the random variable $N_j(t)$. One can then express the state of the system by

$$\boldsymbol{N}(t) = (N_1(t), ..., N_J(t)) \quad (2)$$

and $Y(t)$, the current resource occupancy vector as a function of the system state variable

$$Y(t) = \boldsymbol{N}(t) \cdot \mathbf{A} \ . \quad (3)$$

The set of possible states N is bounded as a result of the finite resource capacity

$$\mathcal{N} = \left\{ \boldsymbol{n} \in \mathrm{IN}^J : [n_1,...,n_J] \bullet \begin{bmatrix} a_1 \\ ... \\ a_J \end{bmatrix} \le c \right\} \quad (4)$$

where $\boldsymbol{n}=(n_1, n_2, ..., n_J)$ is the state of the system (defining the number of each component active requests). In the limit, if there are more users from an application than from another, the other can use fewer channels. An example for $J = 3$ is given in Fig.1.

If the state vector $\boldsymbol{N}(t) \in \mathcal{N}$, then the service occupancy vector $Y(t) \in \mathcal{Y}$, where $\mathcal{Y}$ is simply defined by

$$\mathcal{Y} = \{ y \in \mathrm{IN} : y \le c \}. \quad (5)$$

The equilibrium *pmf* of the state $N(t)$ and the occupancy $Y(t)$ are defined as follows

$$p(\boldsymbol{n}) = \lim_{t \to \infty} \mathrm{Prob}\{\boldsymbol{N}(t) = \boldsymbol{n}\}, \quad (6)$$

$$q(y) = \lim_{t \to \infty} \mathrm{Prob}\{Y(t) = y\}. \quad (7)$$

When the system is in state $N(t) = \boldsymbol{n}$, the time until the next arrival of a class j customer's demand is exponentially distributed with parameter $\lambda_j(n_j)$. This parameter is normalized with respect to the average class j holding time, thus, a different time unit is introduced for each customer class. Blocking takes place if a request cannot be granted entirely, i.e., a class j request arrives when the system is in the set

$$\mathrm{B}_j = \{ \boldsymbol{n} \in \mathcal{N} : \boldsymbol{n} \cdot \mathbf{A} + a_j > c \}. \quad (8)$$

For exponential holding times, the BPP process can be modeled by the Markov chain of Fig. 2, although this model allows for considering more general distributions for the holding times.

While the equilibrium *pmf* of the state $N(t)$, $p(\boldsymbol{n})$, has a product form, in [1] an algorithm is proposed to compute the occupancy *pmf*, $q(y)$, that there is an algorithm that is economic in terms of computation time and storage space – as long as the number of resources is not too high. This algorithm assumes a BPP arrival process.

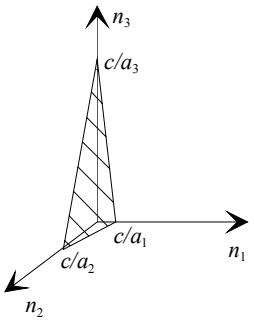

Fig. 1. Boundaries for resource usage for a complete sharing policy, $J = 3$.

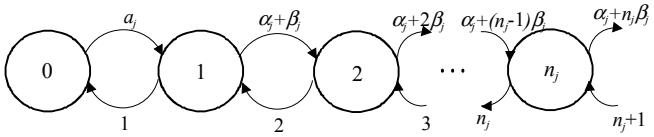

Fig. 2. Markov model of a BPP arrival process for exponential distributed holding times.

BPP processes are those whose arrival intensity (corresponding to an exponential distribution of the inter-arrival times), conditioned to n_j customers being in the system, is of the form

$$\lambda_j(n_j) = \alpha_j + n_j \cdot \beta_j, \text{ with } \alpha_j > 0 \quad (9)$$

where $(-\beta_j)$ is the activation rate and α_j is the arrival rate. In the Poisson case, as $\beta_j=0$, the *pmf* of the number of active customers in an infinite resource is [1] $\lambda_j(n_j) = \alpha_j$.

For exponential holding times, a BPP process can be modelled by the Markov chain from Fig. 2. Due to the normalization, mean holding times are unitary, thus, death rates are integer values.

The description and the pseudo-code for the algorithm for the computation of time and call blocking probabilities, P_{bt} and P_b, respectively, are presented in [1], [5].

A. User Model and Equivalent User

There are a total number of c available resources (or channels) in each cell, being used by a total number of equivalent users, M_T. Furthermore, one is considering two applications, voice (VOI) and Video-telephony (VTE), i.e., a total of $K_{app} = 2$ applications. The index k, k = VOI, VTE, refers to these applications. Given this traffic mixture, the model for applications activation by users is presented in Fig. 3. Each user can be either in an idle state or using one of the two applications, with generation rate, Λ_k, and total service rate, H_k, respectively.

Once application k is active, the service components are activated with rate $\Lambda_{j|k}$ and extinguished with total service rate $\mathrm{H}_{j|k}$, $j = 1, \ldots, J$, Fig. 4; they can be simultaneously active, or not, and some can even not be activated for a given application.

This is a loss system, whose performance can be measured by the blocking probability of each service component, which simplifies the analysis (because one only needs to consider the service components, and not each application).

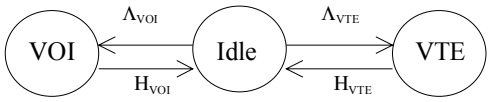

Fig. 3. Applications activation.

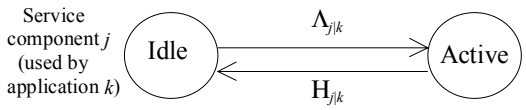

Fig. 4. Service components activation.

From Fig. 3 it is straightforward to derive the probability of a user having application k active

$$p_k = \frac{\Lambda_k / H_k}{1 + \sum_{i=1}^{K_{app}} \Lambda_i / H_i}. \tag{10}$$

The total traffic is

$$\rho = \sum_{k=1}^{K_{app}} (\Lambda_k / H_k). \tag{11}$$

Hence, the traffic generated by a specific application can be given by

$$\rho_k = \frac{\Lambda_k}{H_k} = prop_k \cdot \rho, \tag{12}$$

where $prop_k$ is the proportion of application k active users (numerically given by its usage).

A normalised generation rate is introduced

$$\Lambda_k^* = prop_k \cdot H_k, \tag{13}$$

such that

$$\sum_{k=1}^{K_{app}} \Lambda_k^* / H_k = 1 \tag{14}$$

(one has $\Lambda_k = \rho \cdot \Lambda_k^*$), reducing (10) to

$$p_k = \frac{\rho}{1+\rho} \cdot \left(\frac{\Lambda_k^*}{H_k} \right) = \frac{\rho}{1+\rho} \cdot prop_k. \tag{15}$$

The fraction of active users is

$$f = \frac{\rho}{1+\rho}. \tag{16}$$

The Poisson case of the BPP model is used [1]. In the context of this model, each equivalent user (of the applications) origins a given number of actual users, one for each service component, using resources with data rates B_{sj}, during a time exponentially distributed with average $1/H_j$, j=1,…, J (J is the total number of service components; for the service component j, one has a total service rate $H_j = \mu_j$ in the static case, and $H_j = \mu_j + \eta_j$, otherwise, where μ_j is the service rate, and η_j is the cross-over rate); so the number of users accessing each individual service component is also M_T.

The activation rate of each component j, given an application k, is defined by [5]

$$\Lambda_{j|k} = \frac{E[\text{number of service component } j \text{ requests}]}{E[\text{duration of application } k]} = \\ = \frac{n_{j|k}}{1/\mu_k} = n_{j|k} \cdot \mu_k \tag{17}$$

where $n_{j|k}$ is the number of times the service component j is activated during application k, it being unitary for permanent service components. When the application is static (due to absence of mobility, or because it is not considered), $H_{j|k}$ is given by $\mu_j = 1/\overline{\tau}$ if the service component is permanent ($\overline{\tau}$ being application k average duration), or by $1/\overline{\tau_s}$ if the application is not permanent, $\overline{\tau_s}$ being the average service component duration. Although the values of $\Lambda_{j|k}$ and $H_{j|k}$ change when the influence of mobility of terminals is considered (by a factor associated with the service and the cross-over rates of application k), their ratio is maintained constant, and no change will exist for traffic analysis purposes, except for the blocking/ handover failure thresholds.

The number of active users of component j using their resources at time t, a_j, are represented by the random variable $N_j(t)$. As it was already pointed out, when the system is in state $N(t) = \boldsymbol{n}$, the time until next user class j arrival is exponentially distributed with parameter $\lambda_j(n_j)$ (9).

B. Arrival Rate

The arrival rate is normalised with respect to the total service rate of service component j

$$\alpha_j^{norm} = \alpha_j / H_j, \tag{18}$$

meaning that different time scales are introduced for each service component.

The symmetric of the arrival rate of each service component is given by multiplying the expectation of $\Lambda_{j|k}$ by N_j, leading to [5], [6],

$$(-\alpha_j) = N_j \cdot \sum_{k=1}^{K_{app}} \Lambda_{j|k} \cdot p_k = N_j \cdot \frac{\rho}{1+\rho} \cdot \sum_{k=1}^{K_{app}} \Lambda_{j|k} \cdot prop_k. \tag{19}$$

If the system is stationary, the average occupancy of component j multiplied by N_j is given by the following ratio

$$(-\alpha_j^{norm}) = \frac{-\alpha_j}{H_j} = N_j \cdot \sum_{k=1}^{K_{app}} \frac{\Lambda_{j|k}}{H_{j|k}} \cdot p_k, \tag{20}$$

here called (the symmetric of) the normalised arrival rate, meaning that the service rate of service component j is

$$\mathrm{H}_j = \sum_{k=1}^{K_{app}} \Lambda_{j|k} \cdot \frac{\Lambda_k^*}{\mathrm{H}_k} \Bigg/ \left(\sum_{k=1}^{K_{app}} \frac{\Lambda_{j|k}}{\mathrm{H}_{j|k}} \cdot \frac{\Lambda_k^*}{\mathrm{H}_k} \right). \quad (21)$$

This does depend on mobility, because of the dependence of the numerator on it. If terminal mobility is considered, $\Lambda_{j|k}$ has to be replaced by $\Lambda_{j|k}$ times a factor $(\mu_k+\eta_k)/\mu_k$, where μ_k and η_k are the service and the cross-over rates associated with application k. The holding times for every service component should be i.i.d.. An example is the particular case of having exponential distributed holding times.

The data rate associated to each application is

$$b_k = \sum_{j=1}^{J} \frac{n_{j|k} \cdot 1/\mathrm{H}_{j|k}}{1/\mathrm{H}_k} \cdot B_{sj}, \quad (22)$$

where B_{sj} is the data rate associated to service component j (note that $B_{sj} = a_j \cdot B_{s1}$, where B_{s1} is the system basic data rate). Its value does not change with the consideration of mobility, because when it is considered, $\mathrm{H}_{j|k}$ and H_k are both affected by the factor $(\eta_k+\mu_k)/\mu_k$, the simultaneous change being cancelled by the division.

II. Physical and Mobility Scenario

The physical scenario has a cellular architecture composed by three cells (or ten) with the shape of a roundabout. The cellular architecture consists of a backbone network which interconnects fixed base stations, and mobile units communicating with the base stations via wireless links. Each cell has access to $c=N$ channels.

The call holding time is the average call duration if the call is not prematurely dropped, and it is assumed to be exponentially distributed with average $1/\mu$, where μ is the service rate. The transference of a mobile communication from one cell to another, while a call is in progress, is called handover. If there are not enough channels available in the new cell this call will be dropped, this phenomenon is known as handover failure. The sojourn time is the time that each user stays in a cell, and it follows an exponential distribution with average $1/\eta$, where η is the cross-over rate.

The handover rate γ is given by $\gamma=\eta/\mu$, and the channel occupancy time, τ_c, is given by the minimum between the call holding time and the sojourn time. As the minimum of two variables exponentially distributed is also exponentially distributed, τ_c is exponential [7].

In a roundabout scenario, the traffic is homogeneous, Fig. 5. As a consequence, there is a homogeneous probability of generating new and handovers calls in the three cells with rates λ_i and η_i, respectively. Hence, $\lambda_i = \lambda \forall i$, $\eta_i = \eta \forall i$, and $\sum_{k=1}^{Ncells} p_{ki} = 1 \forall i$ where p_{ki} is the probability that a call may attempt a handover from cell k to cell i, and *Ncells* is the total number cells in the geometry.

In the simulation model one uses three call generators, one for each cell, working simultaneously. Each generator models the calls of one third of the users in the entire roundabout [3].

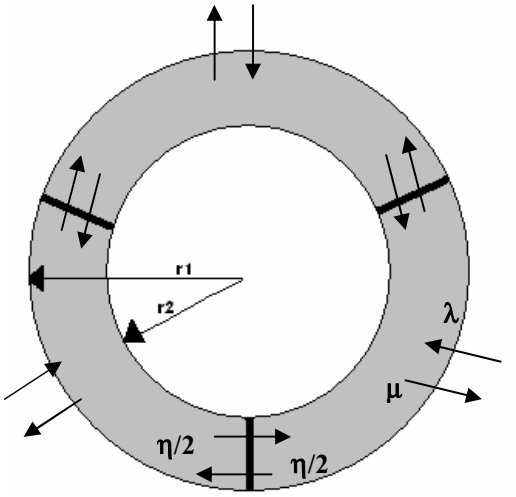

Fig. 5. Physical roundabout scenario.

The new calls are generated following a Poisson distribution with rate λ. The offered load per application is defined as $\rho=\lambda/\mu$. Packet switched traffic is commonly modelled as ON-OFF processes. Our simulator models the ON-OFF behaviour by using active/inactive time periods, according to [4]. A special model is used for real-time video-based applications like VTE due to the high level of burstiness introduced by compression techniques like MPEG-4. In simulations, however, it was considered as having continuous occupation of channels along all the call duration.

The definition of the main concepts and parameters enables the discussion of simulation results and their comparison with other simulation results. As these parameters are general the same formulas are used for different services. The call blocking is the ratio between the number of new calls that are rejected in the process of trying to obtain channels and the total number of new calls generated. The handover failure is the ratio between the number of handovers that are rejected at the new cell in the process of trying to obtain channels, and the total number of handovers produced.

When the traffic is being modelled by ON-OFF periods, the definitions of these call level parameters will be maintained. However, new parameters are needed at the burst level. In this case, the ON-OFF blocking probability, $P_{b\ ONOFF}$, is the ratio between the number of calls that are rejected at the beginning of ON periods, in the process of trying to obtain channels, and the total number of generated ON periods [8].

III. Results

Our simulator can be used for the validation of traffic models. One performs a comparison between the theoretical values obtained by considering the Bernoulli-Poisson-Pascal model for multi-service traffic [6], [8], and the results obtained by using the simulator developed in Visual SLAM with AweSim [8]. Results for bursty VOI are presented in Fig. 6, where a comparison of theoretical and simulation results for $P_{b\ ONOFF}$ is performed for different ρs, with γ as a parameter (VOI,

c=4). Exponential distributions are considered for the active/inactive periods, an average session duration of 60 s is assumed, and the time intervals between arrivals are the ones presented in Table I.

For the VOI application, the theoretical and the experimental values of $P_{b\ ONOFF}$ are close to each other, Figs. 6-7 (example for ρ=0.2 Erl for the latter), and there is an almost perfect concordance between theoretical and simulation values for γ=1, i.e, when the average sojourn time in cells is equal to the average holding time. The curves for P_b and P_{hf} follow a similar behaviour but P_{hf} takes lower values.

TABLE I. TIME INTERVALS BETWEEN ARRIVALS FOR SINGLE-SERVICE

ρ	Time between calls [s]
	VOI
0.05	257.14
0.10	128.57
0.15	85.71
0.20	64.29

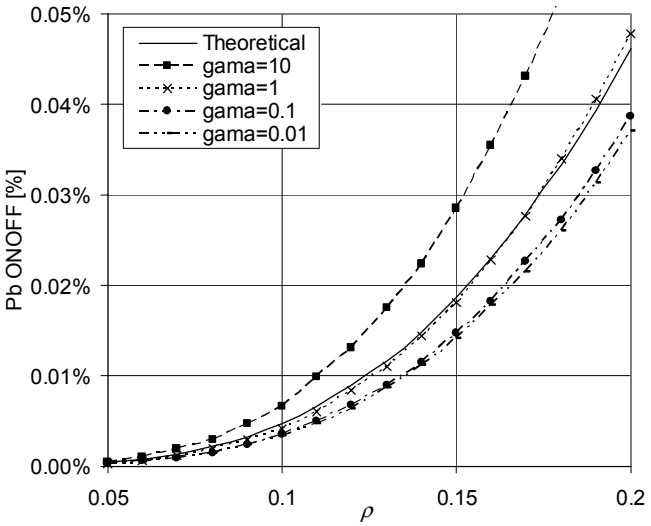

Fig. 6. Comparison of theoretical and simulation results for $P_{b\ ONOFF}$ for different ρs, with γ as a parameter (VOI, c=4).

Besides the validation of the model for the bursty behaviour in the single-service case, it is a worth analysing the case of multi-service, and a mixture of VOI and VTE was chosen. When it is active (60s), VOI has a bursty behaviour, and ON and OFF periods have exponential distributions with average durations 1.4 and 1.7 s, respectively. However, VTE (60s) does not present a bursty behaviour and is permanently active. the time intervals between arrivals are the ones presented in Table II. While for VOI values of P_b are different from $P_{b\ ONOFF}$, for VTE the curves for P_b and $P_{b\ ONOFF}$ are coincident, Figs. 8-9.

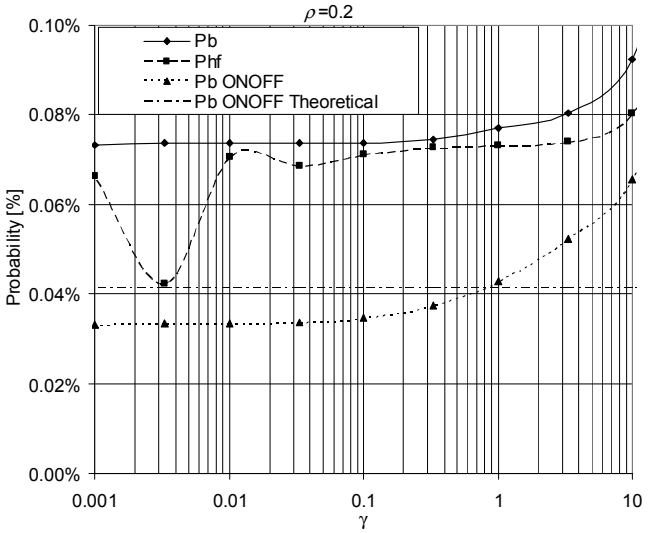

Fig. 7. P_b, P_{hf}, $P_{b\ ONOFF}$ and theoretical $P_{b\ ONOFF}$ as a function of γ for four channels and ρ=0.2 Erl and (VOI).

TABLE II. TIME INTERVALS BETWEEN ARRIVALS FOR MULTI-SERVICE

ρ	Time between calls [s]	
	VOI	VTE
0.05	149.17	391.30
0.10	74.59	195.65
0.15	49.72	130.43
0.20	37.29	97.83

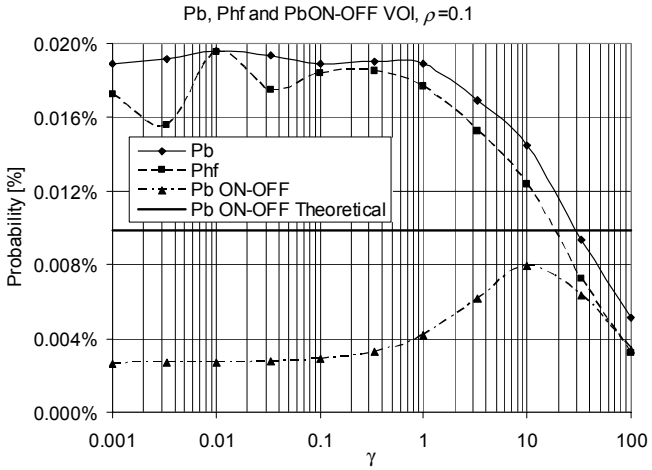

Fig. 8. P_b, P_{hf}, $P_{b\ ONOFF}$ and theoretical $P_{b\ ONOFF}$ for VOI in the multi-service case (c=48 and ρ=0.1 Erl).

In this case, while for VOI the theoretical $P_{b\ ONOFF}$ is always higher than the simulated values, for VTE, it takes values lower than the simulation ones for γs up to ~10. The theoretical and simulation values are close to each other for $\gamma \approx 10$ in both cases.

Figs. 10-11 presents the dependence of $P_{b\ ONOFF}$ on ρ.

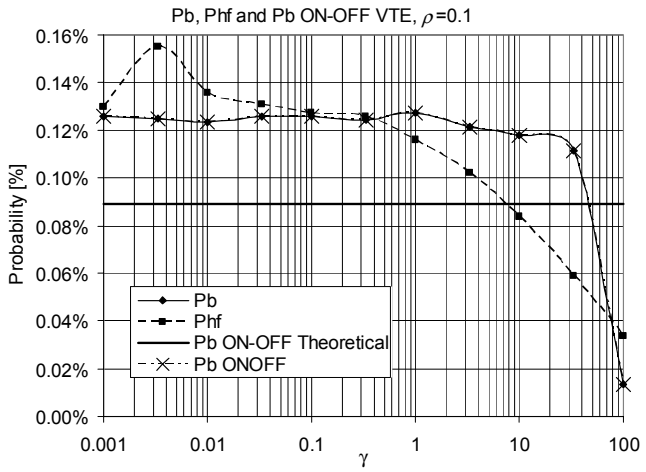

Fig. 9. P_b, P_{hf}, $P_{b\ ONOFF}$ and theoretical Pb ONOFF for VTE in the multi-service case (c=48 and ρ=0.1 Erl).

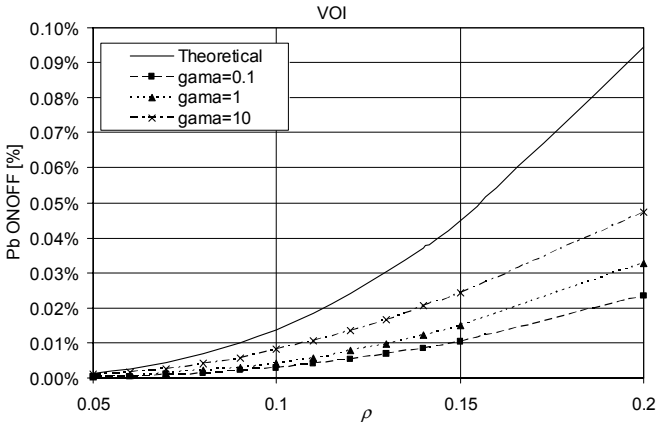

Fig. 10. Comparison of theoretical and simulation results for $P_{b\ ONOFF}$ for different ρs, with γ as a parameter, in the multi-service case (VOI, c=48).

One concludes that the simulation values for γ=10 agree with the theoretical ones only for the lowest values of ρ, i.e., $0.05 \leq \rho \leq 0.10$ Erl.

IV. Conclusions

A model was proposed for multi-rate multi-service traffic engineering purposes, which is based in the BPP model. Simulations were ran to obtain results for multi-service QoS measures, like blocking, handover, and ON-OFF blocking probabilities. By comparing simulation results with the theoretical ones, a perfect validation was achieved in the single-service case when the sojourn time in cells is equal to the average duration of voice calls. In the multi-service case, the behaviour is not exactly the same but a coherent behaviour is achieved. For γ=10 simulation values agree with the theoretical ones for only the lowest values of ρ, i.e., $0.05 \leq \rho \leq 0.10$ Erl.

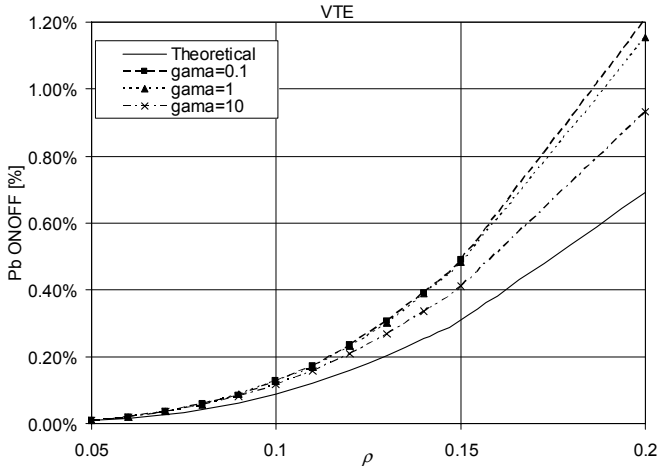

Fig. 11. Comparison of theoretical and simulation results for $P_{b\ ONOFF}$ for different ρs, with γ as a parameter, in the multi-service case (VTE, c=48).

Acknowledgment

This work was partially funded by MULTIPLAN and CROSSNET (Portuguese Foundation for Science and Technology POSI and POSC projects with FEDER funding), and by "Projecto de Re-equipamento Científico" REEQ/1201/EEI/2005 (a Portuguese Foundation for Science and Technology project).

References

[1] G. A. Awater and H. A. van de Vlag, "Exact computation of Time and Call Blocking Probabilities in Large, Multi-traffic, Multi-resource Loss Systems," *Performance Evaluation*, Vol. 25, No. 1, Mar. 1996, pp. 41-58.

[2] F. J. Velez and Rui R. Paulo, "High Capacity Wideband Traffic in Enhanced UMTS: a Step Towards 4G", in *Proc. of 3G 2004 – 5th IEE International Conference on 3G Mobile Communication Technologies*, London, UK, Oct. 2004.

[3] Jesús M. J. Valero, Rui R. Paulo and Fernando J. Velez, "Tele-Traffic Simulation for Mobile Communication Systems Beyond 3G," in *Proc. of AICT' 06 - The 2nd Advanced International Conference on Telecommunications*, Guadeloupe, French Caribbean, Feb. 2006.

[4] J. S. Kaufman, "Blocking in a Shared Resource Environment," *IEEE Transactions on Communications*, Vol. COM-29, No. 10, Oct. 1981, pp. 1474-1481.

[5] R. M. Carvalho and J. M. Brázio, "Multi-service Traffic Model for the Sharing of a Resource by a Homogeneous Population of Users with Stochastically Heterogeneous Demands" (in portuguese), in *Proc. of 6th Conference of the Portuguese Statistic Society*, Tomar, Portugal, June1998.

[6] R. M. Carvalho, *Multi-service Traffic Models for Cellular Mobile and Personal Communication Systems* (in portuguese), Graduation Report, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, Jan. 1998.

[7] M. Sidi and D. Starobinski, "New Call Blocking versus Handoff Blocking in Cellular Networks", *wireless Networks*, vol.3, No 1, Feb. 1997, pp. 15-27.

[8] Jesús M. J. Valero, *Tele-Traffic Simulation for Mobile Communication Systems Beyond 3G*, Graduation Thesis, University of Beira Interior, Covilhã, Portugal, Sep. 2005.