A Markovian Approach to Radio Access Technology Selection in Heterogeneous Multiaccess/Multiservice Wireless Networks

Xavier Gelabert, *Student Member*, *IEEE*, Jordi Pérez-Romero, *Member*, *IEEE*, Oriol Sallent, and Ramon Agustí, *Member*, *IEEE*

Abstract—This paper addresses the problem of Radio Access Technology (RAT) selection in heterogeneous multiaccess/multiservice scenarios. For such purpose, a Markov model is proposed to compare the performance of various RAT selection policies within these scenarios. The novelty of the approach resides in the embedded definition of the aforementioned RAT selection policies within the Markov chain. In addition, the model also considers the constraints imposed by those users with terminals that only support a subset of all the available RATs (i.e., multimode terminal capabilities). Furthermore, several performance metrics may be measured to evaluate the behavior of the proposed RAT selection policies under varying offered traffic conditions. In order to illustrate the validation and suitability of the proposed model, some examples of operative radio access networks are provided, including the GSM/EDGE Radio Access Network (GERAN) and the UMTS Radio Access Network (UTRAN), as well as several service-based, load-balancing, and terminal-driven RAT selection strategies. The flexibility exhibited by the presented model enables to extend these RAT selection policies to others responding to diverse criteria. The model is successfully validated by means of comparing the Markov model results with those of system-level simulations.

Index Terms—Algorithm design and analysis, Markov processes, mobile communication systems, radio resource management.

1 Introduction

Today's wireless communications can be driven by a wide range of Radio Access Technology (RAT) standards. The success of second-generation (2G) cellular systems, e.g., Global System for Mobile Communications (GSM), cdmaOne, and Pacific Digital Cellular (PDC), along with the IP data support provided by 2.5G technologies, such as the General Packet Radio System (GPRS), paved the way toward evolved systems with higher data rate capabilities. In this sense, technologies like the Enhanced Data rates for GSM Evolution (EDGE) offer high data rates using inherited 2G network infrastructure and frequency spectrum. In order to supply even higher data rates, third-generation (3G) systems arose with new assigned frequency bands along with the deployment of new network elements, especially in the radio access part. 3G systems comprise several standards such as the Universal Mobile Telecommunications System (UMTS), the Freedom of Mobile Multimedia Access (FOMA), CDMA2000, and the Time Division-Synchronous Code Division Multiple Access (TD-SCDMA) among others. Moreover, in parallel with the evolution of cellular systems, a number of Wireless Local Area Networks (WLANs) like, e.g., the IEEE 802.11 standard families, have emerged and become profusely used in home environments. In addition, Wireless Metropolitan Area Networks (WMANs) like the Worldwide Interoperability for Microwave Access (WiMax/IEEE 802.16) standard will extend communication ranges beyond those covered by WLANs.

In this framework, the heterogeneous network notion arises in order to propose a flexible architecture capable of managing this large variety of wireless access technologies along with applications and services comprising different quality-of-service (QoS) demands and protocol stacks. The deployment of such heterogeneous network topologies requires, however, a degree of interworking between the different network entities which may lead to open, loose, tight, and very tight couplings [1]. In this way, heterogeneous networks may provide a larger set of available resources than individual networks, allowing users to seamlessly connect, at any time and any place, to the access technology that is most suitable according to some user/operator-specified criteria. This notion has been coined as the *Always Best Connected* concept [2].

In order to manage, in the most efficient way, the pool of existing resources provided by a heterogeneous network comprising several RATs, Common Radio Resource Management (CRRM) architectures and strategies are devised [3]. In this sense, CRRM has been identified as an important issue by the Third Generation Partnership Project (3GPP), which defines some recommendations and architectures for CRRM operation [4], [5]. Efficient CRRM will then exploit the trunking gain that results from the common management of all the available radio resources of all networks rather than managing those radio resources considering stand-alone networks [6]. Then, the tighter the coupling

• *The authors are with Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Jordi Girona 1-3 (Campus Nord), 08034 Barcelona, Spain.*
E-mail: {xavier.gelabert, jorperez, sallent, ramon}@tsc.upc.edu.

Manuscript received 26 Feb. 2007; revised 20 Sept. 2007; accepted 13 Mar. 2008; published online 1 Apr. 2008.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0062-0207.
Digital Object Identifier no. 10.1109/TMC.2008.50.

between these networks, the better the resources are being utilized leading to an improved performance. Consequently, efforts in the definition and implementation of required interfaces must be developed in this direction.

Inherent to heterogeneous networks, to select an appropriate RAT for an incoming user requesting a given service becomes a key CRRM issue. This RAT selection can be carried out considering different criteria (such as, e.g., service type and load conditions) with the final purpose of enhancing overall capacity, resource utilization, and service quality.

Although the RAT selection problem has been covered in a number of papers, see, for example, [7], [8], and [9], the proposed methodology usually relies on system-level simulations in order to extract some relevant performance metrics to compare different strategies. The analytical approach to the RAT selection problem, however, has been less addressed in the literature. To the authors' knowledge, only a few analytical proposals have been developed up to date, e.g., [10], [11], and [12]. In [10], Lincke-Salecker and Hood propose an analytical approach to the problem of traffic overflowing between several RATs using a multidimensional Markov model. However, in order to derive a closed form solution by means of applying independence between service types, Markov states in this model indicate the number of sessions of each service that are being allocated in whole composite network, but not on which RAT each session is being served. In [11], a near-optimum service allocation is proposed in order to maximize the combined multiservice capacity. The authors assumed an a priori knowledge of the services that need to be allocated, rather than modeling user arrival process. In [12], Koo et al. assess the separate and common Erlang capacity of a multiaccess/multiservice system. For this purpose, an Erlang Loss queuing approach is assumed and a closed product form expression for the equilibrium probability is provided. Nevertheless, this assumption implies that the fractional traffic loads of each service offered to each system are known, so the approach is only valid to evaluate some basic RAT selection policies.

In this paper, the proposed analytical model entails a more flexible framework by assuming that only the total offered traffic to the multi-RAT system for each service is known. Thus, fractional traffic arriving to each RAT will be dependent on the chosen RAT selection scheme which is fully embedded in the model. This feature constitutes the main innovative contribution of this article and differentiates it from previous approaches to the problem. In particular, the model describes the allocation of two service types onto two RATs, which allows the definition of a wide range of RAT selection policies taking into account several criteria such as service type, network conditions, and terminal types. Finally, the proposed model also captures the availability of multimode terminals, i.e., those that can operate on both RATs, so as to reflect a more realistic medium-term scenario considering the flexibility constraints of those terminals supporting one single RAT (i.e., single-mode terminals).

Multidimensional Markov models have been widely used in the field of networking to model the behavior of communication networks under variable traffic load conditions [13]. In the analysis of this paper, the focus is on two RATs with different underlying access methodologies: the Time-Division Multiple Access (TDMA) and the Wideband Code Division Multiple Access (WCDMA). These two access schemes may, e.g., represent 3GPP standardized technologies GSM/EDGE and UMTS, respectively [14], [15], although the model could be adapted to other standards by a proper change in the parameter values.

The article is organized as follows: Section 2 deals with the problem statement and the considered approach to solve it. Section 3 presents the analytical model and the notation that will be used throughout this paper. In Section 4, various RAT selection policies are described by means of the proposed model. Section 5 presents the performance metrics that will be used to evaluate the behavior of initial RAT selection policies in Section 6. Finally, Section 7 deals with the conclusions and future work.

2 Problem Statement

For the evaluation of the forthcoming RAT selection strategies, a scenario is assumed where a TDMA-based and a WCDMA-based technology coexist and provide coverage over a same area. Generically, one can characterize this scenario by means of a Markov chain represented by a $(N+M)$-dimensional state, denoted as $S_{(t_1,t_2,...,t_N,w_1,w_2,...,w_M)}$, where t_n $(1 \leq n \leq N)$ and w_m $(1 \leq m \leq M)$ relate to the N and M dimensions corresponding to TDMA and WCDMA RATs, respectively. Each of these dimensions can represent a single or a combination of communication characteristics, such as service type (e.g., voice or data), user communication status (e.g., active or queued users), transmission rate, and amount of allocated resources.

In this paper, a 4D Markov chain is considered accounting for two service types, generically voice and data, being served over the aforementioned RATs, TDMA and WCDMA, in order to model the system behavior. Therefore, let $S_{(i,j,k,l)}$ represent the state in which i voice users and j data users are being served through TDMA; and k voice users and l data users are being served through WCDMA. These indices represent the number of active simultaneous voice calls and data sessions being carried out at a given time.

Transitions between states within the Markov chain will occur due to call/session arrivals or due to call/session departures. Regarding traffic patterns, it is supposed that voice calls and data sessions are generated according to Poisson processes with rates λ_v and λ_d, respectively. As for voice-call holding time and data session time, they follow exponential distributions with means $1/\mu_v$ and $1/\mu_d$ correspondingly. It is assumed that only transitions between neighboring states (those that only differ in a single increment/decrement in a sole state dimension) are allowed. This prevents situations where more than one call/session arrives or departs from a given state at the same time.

Within the set of CRRM functions devoted to efficiently manage the available resources in a heterogeneous network, the RAT selection plays a key role in deciding the most appropriate RAT for a given service at a given time. In that sense, the algorithm operation might then respond to specific policies taking into account both technical and/or economical aspects (e.g., operator or user preferences). In

Fig. 1. Mapping of total-to-fractional arrival rates given by initial RAT selection.

the context of the proposed Markov framework, it is important to notice that, given the total voice call and data session arrival rates, λ_v and λ_d, respectively, the adopted RAT selection policy will determine the arrival rates into each RAT (see Fig. 1). Consequently, the RAT selection policy will modulate the transition rates between the states $S_{(i,j,k,l)}$ in the Markov chain according to a predefined RAT selection policy. Mathematically, given a generic RAT selection policy denoted as $\pi_{(i,j,k,l)}$, we may introduce the following function:

$$\begin{aligned} \pi_{(i,j,k,l)}: \quad \mathbb{R}^2 &\longrightarrow \mathbb{R}^4 \\ (\lambda_v, \lambda_d) &\longrightarrow \left(\lambda_v^T, \lambda_d^T, \lambda_v^W, \lambda_d^W\right), \end{aligned} \tag{1}$$

where λ_v^T, λ_d^T, λ_v^W, λ_d^W represent the fractional arrival rates of each service into to each of the available RATs given by policy $\pi_{(i,j,k,l)}$. In this way, by an appropriate definition of the RAT selection policies, it is possible to embed those into the Markov chain and evaluate the performance of the system by considering that only the total voice and data offered traffic, i.e., λ_v and λ_d, are known parameters. This approach differentiates our work from previous mentioned studies [10], [11], [12] and constitutes the main innovative contribution of this work which will be fully developed in the next sections.

3 The 4D Markov Model State Space

In the following, the Markov model state space containing the total set of feasible states is presented. Clearly, if the capacity in terms of number of supported users in each RAT is assumed to be fixed, a finite number of states $S_{(i,j,k,l)}$ (called feasible states) limited by the number of allowable users of each service in each RAT must exist.

This limit is usually set by the RAT-specific Call Admission Control (CAC) procedures, that determine if a given user should be admitted or not, so as to guarantee some minimum QoS requirements to users already admitted in the system. Because CAC is dependant on the underlying technology, the set of feasible states in TDMA, S^T, and WCDMA, S^W, can be individually defined as

$$S^T = \left\{ S_{(i,j,k,l)} | 0 \le f_{(i,j)}^T \le 1, \forall k, l \right\}, \tag{2}$$

$$S^W = \left\{ S_{(i,j,k,l)} | 0 \le f_{(k,l)}^W \le 1, \forall i, j \right\}, \tag{3}$$

where $f_{(i,j)}^T$ and $f_{(k,l)}^W$ are defined as the feasibility conditions which account for the CAC procedures in TDMA and WCDMA correspondingly.

Consequently, we can define the set of feasible states, S, which include all states $S_{(i,j,k,l)}$ that satisfy the CAC procedures in each of the systems. Then, a given state $S_{(i,j,k,l)}$ is said to be feasible, if it satisfies that $S_{(i,j,k,l)} \in S$ with $S = S^T \cap S^W$, i.e., a state is only feasible if it is feasible in both TDMA and WCDMA systems.

In the following sections, the state feasibilities for TDMA and WCDMA are presented. In this paper, CAC procedures are based on the reverse link (uplink, UL) in order to determine the number of allowable users in each RAT, which it is assumed to be the most restricting case.

3.1 TDMA State Feasibility

The resource allocation for voice and data services in a TDMA-based technology, such as, e.g., GSM/EDGE, relies on the *capacity on demand* principle [16]. Briefly, a data user can transmit data over a number of simultaneous channels or timeslots (TSLs). Moreover, several data users can be multiplexed over the same TSL for coordinated data transmission by means of an efficient scheduling mechanism. Given that voice and data users can demand different amounts of resources and that these resources are shared between them, mechanisms to referee the sharing among voice and data traffic are needed [17]. In this paper, and for the sake of simplicity, it is assumed that the total available capacity is shared between voice and data traffic on a first-come-first-served basis with no service priority.

If C is the total number of available channels (TSLs) available for voice and data services in the cell, the maximum number of voice users being served through TDMA, i, is upper-bounded by $i \le C$. Considering the UL direction, assuming voice and data users are granted with a single channel for each connection,[1] and that a maximum number of n_C data users are allowed to share the same TSL, the maximum number of simultaneous data users being served through TDMA must satisfy $j \le n_C C$. Since voice and data services share the total amount of resources, the previous conditions may be expressed jointly as

$$0 \le i/C + j/n_C C \le 1, \tag{4}$$

which implicitly defines the state feasibility condition for the TDMA system, i.e., $f_{(i,j)}^T = i/C + j/n_C C$.

3.2 WCDMA State Feasibility

In WCDMA-based systems, the UL load factor $(L_{(k,l)}^W)$ condition must hold in order to ensure that users are granted the desired capacity for their demanding services. Considering k voice users and l data users being served in WCDMA, the UL load factor condition for a single cell may be expressed as [15]

$$0 \le L_{(k,l)}^W \le \eta_{max}, \tag{5}$$

where

$$L_{(k,l)}^W = k\left[\frac{W/R_{b,v}}{(E_b/N_0)_v} + 1\right]^{-1} + l\left[\frac{W/R_{b,d}}{(E_b/N_0)_d} + 1\right]^{-1}, \tag{6}$$

1. Although the consideration of multislot capabilities in the model would be feasible, this would complicate the algebra and the model while not bringing substantial added value on the methodology and approach of this paper. Thus, a single TSL is allocated to data users.

with W as the chip rate; $R_{b,v}$ and $R_{b,d}$ are the bit rates granted to voice and data services; $(E_b/N_0)_v$ along with $(E_b/N_0)_d$ are the target bit-energy-to-noise-density ratio after de-spreading and decoding for voice and data users; and η_{max} is the admission threshold. By choosing an appropriate value for η_{max}, quality requirements of admitted users (e.g., in terms of bit error rate) depending on the coverage conditions can be ensured [15]. From (5), the state feasibility condition of WCDMA system, $f^W_{(k,l)}$, is easily identified as $f^W_{(k,l)} = L^W_{(k,l)}/\eta_{max}$.

3.3 Call Admission Control and Blocking States

Once the state space has been defined by means of the feasibility conditions in each RAT, let us define, for the sake of convenience, the set of states in which the acceptance of a new user would force a transition to an unfeasible state $S_{(i,j,k,l)} \notin S$. Under these circumstances, the RAT in question is said to be in a blocking state. Let $S^\rho_{b,\sigma}$ denote the set of feasible states where the acceptance of a service type σ user in RAT ρ forces the state to move to an unfeasible state. Then, the fractional per-service/per-RAT blocking states for voice and data services, i.e., $\sigma = \{v, d\}$, in TDMA and WCDMA RATs, $\rho = \{T, W\}$, are defined as

$$\begin{aligned} S^T_{b,v} &= \big\{S_{(i,j,k,l)} \in S | S_{(i+1,j,k,l)} \notin S\big\},\\ S^T_{b,d} &= \big\{S_{(i,j,k,l)} \in S | S_{(i,j+1,k,l)} \notin S\big\},\\ S^W_{b,v} &= \big\{S_{(i,j,k,l)} \in S | S_{(i,j,k+1,l)} \notin S\big\},\\ S^W_{b,d} &= \big\{S_{(i,j,k,l)} \in S | S_{(i,j,k,l+1)} \notin S\big\}. \end{aligned} \tag{7}$$

If S^ρ_b denotes the set of feasible states where the acceptance of any service type user in RAT ρ forces the state to move to an unfeasible state, we have

$$S^\rho_b = S^\rho_{b,v} \cap S^\rho_{b,d}. \tag{8}$$

Assuming that a given service type user can be allocated in either of the existing RATs provided the one selected by the RAT selection policy is blocked, we can define service blocking states where the acceptance of a given service type user $\sigma = \{v, d\}$ forces the current state to move to an unfeasible state in each of the considered RATs $\rho = \{T, W\}$. Bearing in mind (7), the per-service blocking set $S_{b,\sigma}$ can be defined as

$$S_{b,\sigma} = S^T_{b,\sigma} \cap S^W_{b,\sigma}. \tag{9}$$

Finally, if S_b defines the set of states where the acceptance of any service type user in any of the available RATs forces the state to move to an unfeasible state, then the total blocking states are defined as

$$S_b = S^T_b \cap S^W_b = S^T_{b,v} \cap S^T_{b,d} \cap S^W_{b,v} \cap S^W_{b,d}. \tag{10}$$

4 Radio Access Technology Selection Policies and State Transitions

Based on the relation provided by (1), we can conveniently define the RAT selection policies as functions that map the total arrival rates λ_v and λ_d into fractional arrival rates of each service into each system (i.e., λ^T_v, λ^T_d, λ^W_v, and λ^W_d) depending on the current state information. Then, in a given state $S_{(i,j,k,l)}$, relation (1) can be rewritten as

$$\begin{aligned} \pi_{(i,j,k,l)}: \quad \mathbb{R}^2 &\longrightarrow \mathbb{R}^4\\ \begin{pmatrix}\lambda_v\\ \lambda_d\end{pmatrix}^T &\longrightarrow \begin{pmatrix} \alpha_{(i,j,k,l)}\lambda_v\delta_{(i+1,j,k,l)}\\ \beta_{(i,j,k,l)}\lambda_d\delta_{(i,j+1,k,l)}\\ \bar{\alpha}_{(i,j,k,l)}\lambda_v\delta_{(i,j,k+1,l)}\\ \bar{\beta}_{(i,j,k,l)}\lambda_d\delta_{(i,j,k,l+1)}\end{pmatrix}^T, \end{aligned} \tag{11}$$

where, given RAT selection policy $\pi_{(i,j,k,l)}$, $\alpha_{(i,j,k,l)}$, and $\bar{\alpha}_{(i,j,k,l)} = (1 - \alpha_{(i,j,k,l)})$ are the functions determining the fraction of voice users into TDMA and WCDMA, respectively, and $\beta_{(i,j,k,l)}$ along with $\bar{\beta}_{(i,j,k,l)} = (1 - \beta_{(i,j,k,l)})$ are the functions governing the fractional data arrival rates into TDMA and WCDMA, respectively. Furthermore, function $\delta_{(i,j,k,l)}$ is an indicator function which will guarantee that nonfeasible states, i.e., $S_{(i,j,k,l)} \notin S$, are not taken into account in the transitions, thus

$$\delta_{(i,j,k,l)} = \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \in S,\\ 0 & \text{otherwise.}\end{cases} \tag{12}$$

The proposed model also allows us to take into consideration scenarios in which not all terminals have multimode capabilities. In that respect, assume that the availability of terminals that support both RATs (multimode) is given by p which is defined as the fraction of terminals with multimode capabilities. Accordingly, the ratio of terminals that only support TDMA (single-mode) is given by $\bar{p} = 1 - p$. The rationale behind this assignment resides in the fact that terminals supporting more recent technologies, such as WCDMA, will most probably support preceding technologies like TDMA. The converse, however, will be less usual. Then, the fractional traffic derived into each RAT stated in (11) may be rewritten as

$$\begin{aligned} \pi_{(i,j,k,l)}: \quad \mathbb{R}^2 &\longrightarrow \mathbb{R}^4\\ \begin{pmatrix}\lambda_v\\ \lambda_d\end{pmatrix}^T &\longrightarrow \begin{pmatrix} \alpha^p_{(i,j,k,l)}\lambda_v\delta_{(i+1,j,k,l)}\\ \beta^p_{(i,j,k,l)}\lambda_d\delta_{(i,j+1,k,l)}\\ \bar{\alpha}^p_{(i,j,k,l)}\lambda_v\delta_{(i,j,k+1,l)}\\ \bar{\beta}^p_{(i,j,k,l)}\lambda_d\delta_{(i,j,k,l+1)}\end{pmatrix}^T, \end{aligned} \tag{13}$$

where $\alpha^p_{(i,j,k,l)}$ and $\beta^p_{(i,j,k,l)}$ relate to the RAT selection policy assignment considering the presence of both multimode and single-mode terminals. In particular, voice and data traffic offered to TDMA will consist of not only the traffic allocated by means of the applied RAT selection policy but also by the traffic that does not support WCDMA. This can be expressed as

$$\begin{aligned} \alpha^p_{(i,j,k,l)}\lambda_v &= \big[\alpha_{(i,j,k,l)} + (1 - \alpha_{(i,j,k,l)})(1 - p)\big]\lambda_v,\\ &= \big(\bar{p} + \alpha_{(i,j,k,l)}p\big)\lambda_v,\\ \beta^p_{(i,j,k,l)}\lambda_d &= \big[\beta_{(i,j,k,l)} + (1 - \beta_{(i,j,k,l)})(1 - p)\big]\lambda_d,\\ &= \big(\bar{p} + \beta_{(i,j,k,l)}p\big)\lambda_d, \end{aligned} \tag{14}$$

where $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ relate to the policy decision (note that if $p = 1$, i.e., all terminals are multimode, thus expression (13) becomes expression (11)).

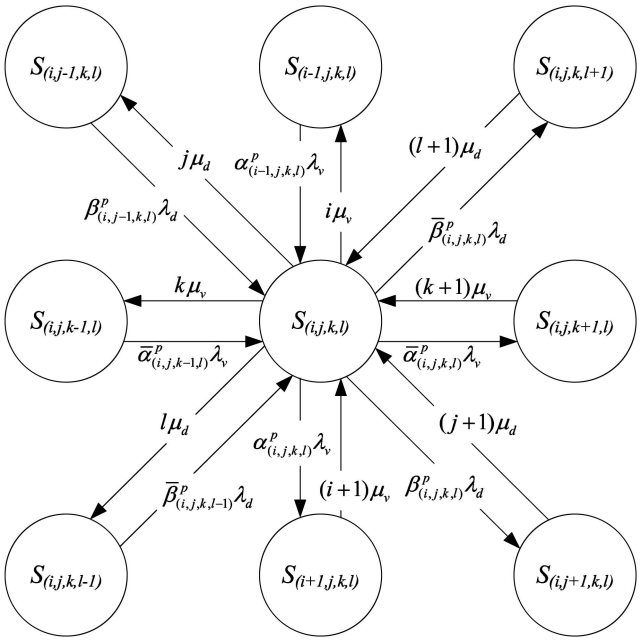

Fig. 2. State transition diagram for a general state.

Given the fractional arrival rates provided in (13), the state transition diagram at a particular nonboundary state $S_{(i,j,k,l)}$ may be built (Fig. 2). By inspection of Fig. 2, we may deduce the Steady-State Balance Equation (SSBE) for a given state $S_{(i,j,k,l)}$ as

$$\begin{aligned}
&P_{(i,j,k,l)}\Big[\alpha^p_{(i,j,k,l)}\lambda_v\delta_{(i+1,j,k,l)} + i\mu_v\delta_{(i-1,j,k,l)} \\
&\quad + \bar{\alpha}^p_{(i,j,k,l)}\lambda_v\delta_{(i,j,k+1,l)} + k\mu_v\delta_{(i,j,k-1,l)} \\
&\quad + \beta^p_{(i,j,k,l)}\lambda_d\delta_{(i,j+1,k,l)} + j\mu_d\delta_{(i,j-1,k,l)} \\
&\quad + \bar{\beta}^p_{(i,j,k,l)}\lambda_d\delta_{(i,j,k,l+1)} + l\mu_d\delta_{(i,j,k,l-1)}\Big] \\
&= \alpha^p_{(i-1,j,k,l)}\lambda_v P_{(i-1,j,k,l)}\delta_{(i-1,j,k,l)} \\
&\quad + (i+1)\mu_v P_{(i+1,j,k,l)}\delta_{(i+1,j,k,l)} \\
&\quad + \bar{\alpha}^p_{(i,j,k-1,l)}\lambda_v P_{(i,j,k-1,l)}\delta_{(i,j,k-1,l)} \\
&\quad + (k+1)\mu_v P_{(i,j,k+1,l)}\delta_{(i,j,k+1,l)} \\
&\quad + \beta^p_{(i,j-1,k,l)}\lambda_d P_{(i,j-1,k,l)}\delta_{(i,j-1,k,l)} \\
&\quad + (j+1)\mu_d P_{(i,j+1,k,l)}\delta_{(i,j+1,k,l)} \\
&\quad + \bar{\beta}^p_{(i,j,k,l-1)}\lambda_d P_{(i,j,k,l-1)}\delta_{(i,j,k,l-1)} \\
&\quad + (l+1)\mu_d P_{(i,j,k,l+1)}\delta_{(i,j,k,l+1),}
\end{aligned} \tag{15}$$

where $P_{(i,j,k,l)}$ is the steady-state probability of being in state $S_{(i,j,k,l)}$.

Once the SSBEs are obtained for all states $S_{(i,j,k,l)} \in S$, numerical methods may be used to solve the system of equations given by the SSBEs plus the normalization constraint

$$\sum_{S_{(i,j,k,l)}\in S} P_{(i,j,k,l)} = 1. \tag{16}$$

The proposed analytical approach allows us to define a wide range of RAT selection policies taking into account several allocation criteria, such as service type, load, and network conditions. In particular, some of the policies presented in [7] and [18] will be adapted to our Markov model in the following sections.

4.1 Random (RND) RAT Selection Policy

For illustrative purposes, this policy randomly selects the RAT on which the call/session will be carried out. Assume TDMA is selected randomly for voice and data users with a probability of α and β, respectively. In the same way, WCDMA is selected with a probability $(1-\alpha)$ and $(1-\beta)$ for voice and data users correspondingly.

If the system is in a voice blocking state, i.e., $S_{(i,j,k,l)} \in S^T_{b,v}$ or $S_{(i,j,k,l)} \in S^W_{b,v}$, or in a data blocking state, i.e., $S_{(i,j,k,l)} \in S^T_{b,d}$ or $S_{(i,j,k,l)} \in S^W_{b,d}$, then the arrival rates of voice and data users to nonblocked states happen with probability equal to the unity. This ensures that a call/session will not be dropped due to the random allocation policy if resources exist in the opposite RAT to the one chosen by the policy. Then, the values of $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (11) may be written as

$$\begin{aligned}
\alpha_{(i,j,k,l)} &= \begin{cases} \alpha & \text{if } S_{(i,j,k,l)} \notin (S^T_{b,v} \cup S^W_{b,v}), \\ 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,v}, \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,v}, \end{cases} \\
\beta_{(i,j,k,l)} &= \begin{cases} \beta & \text{if } S_{(i,j,k,l)} \notin (S^T_{b,d} \cup S^W_{b,d}), \\ 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,d}, \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,d}. \end{cases}
\end{aligned} \tag{17}$$

4.2 Service-Based #1 (SB#1) RAT Selection Policy

This policy intends to allocate voice users to TDMA and data users to WCDMA. If the assignment is not possible, i.e., the chosen RATs are at full capacity, the voice users are directed to WCDMA and data users to TDMA.

Bearing this in mind, a voice arrival is not allowed in WCDMA, i.e., the transition $S_{(i,j,k,l)} \to S_{(i,j,k+1,l)}$ is not allowed, unless we are in a TDMA voice blocking state ($S_{(i,j,k,l)} \in S^T_{b,v}$). Moreover, a data session arrival will not be accommodated in TDMA, i.e., the transition $S_{(i,j,k,l)} \to S_{(i,j+1,k,l)}$ is not allowed, unless we are in a WCDMA data blocking state, i.e., $S_{(i,j,k,l)} \in S^W_{b,d}$. In order to take these restrictions into account in the global balance equations, the functions $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (11) which define the feasibility of a voice arrival in WCDMA and the feasibility of a data arrival in TDMA can be defined as

$$\begin{aligned}
\alpha_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^T_{b,v}, \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,v}, \end{cases} \\
\beta_{(i,j,k,l)} &= \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,d}, \\ 0 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,d}. \end{cases}
\end{aligned} \tag{18}$$

4.3 Service-Based #2 (SB#2) RAT Selection Policy

This policy, acting as opposite to the SB#1 policy, intends to allocate voice users to WCDMA and data users to TDMA. If the assignment is not possible, i.e., the chosen RATs are at full capacity, the voice users are directed to TDMA and data users to WCDMA.

Keep in mind that a voice call will be not admitted in TDMA, i.e., the transition $S_{(i,j,k,l)} \to S_{(i+1,j,k,l)}$ will be not

allowed, unless no capacity is left for voice users in WCDMA, that is $S_{(i,j,k,l)} \in S^W_{b,v}$. Similarly, data users will be admitted in WCDMA, i.e., the transition $S_{(i,j,k,l)} \rightarrow S_{(i,j,k,l+1)}$, only if no capacity is left in TDMA to accommodate the data session, i.e., $S_{(i,j,k,l)} \in S^T_{b,d}$. In order to account for these limitations in the arrival rates, functions $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (11) denoting the feasibility of data arrival rates in TDMA and of voice arrival rates in WCDMA can be expressed as

$$\alpha_{(i,j,k,l)} = \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \in S^W_{b,v}, \\ 0 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,v}, \end{cases} \qquad \beta_{(i,j,k,l)} = \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^T_{b,d}, \\ 0 & \text{if } S_{(i,j,k,l)} \in S^T_{b,d}. \end{cases} \tag{19}$$

4.4 Load Balancing (LB) RAT Selection Policy

The LB policy intends to allocate users to the RAT that undergoes a lower load situation at a given time. In particular, transitions between a source state and possible destination states will depend on the measured load at each destination state.

Before expressing this notion in terms of transition rates in our Markov model, it is convenient to define the load metrics in both RATs.

In TDMA-based GSM/EDGE, the TSL utilization factor, initially defined in [14], may be used to measure the load in a given state $S_{(i,j,k,l)} \in S$ as

$$L^T_{(i,j)} = n_{(i,j)}/C, \tag{20}$$

where C is the total number of available channels (TSLs) in the cell devoted to voice and data traffic services, and $n_{(i,j)}$ is the number of occupied channels (TSLs) when i voice users and j data users are currently being served in TDMA. For the case of data users requiring a single slot for their UL connection, $n_{(i,j)} = \min(C, i+j)$. Note that this definition of load will not account for multiple users sharing a same TSL nor users using multiple TSLs.

On the other hand, the load in a WCDMA-based system may be calculated by means of the UL load factor $L^W_{(k,l)}$, defined in (6), scaled by $\eta_{\max}$.

In order to determine whether the incoming user demanding a given service should be allocated to TDMA or to WCDMA, functions $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (11) will take the following values:

$$\alpha_{(i,j,k,l)} = \begin{cases} 1 & \text{if } (L^T_{(i+1,j)} < L^W_{(k+1,l)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \notin S^T_{b,v} \wedge S_{(i,j,k,l)} \in S^W_{b,v}), \\ 0 & \text{if } (L^T_{(i+1,j)} > L^W_{(k+1,l)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \in S^T_{b,v} \wedge S_{(i,j,k,l)} \notin S^W_{b,v}), \\ 0.5 & \text{otherwise}, \end{cases}$$

$$\beta_{(i,j,k,l)} = \begin{cases} 1 & \text{if } (L^T_{(i,j+1)} < L^W_{(k,l+1)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \notin S^T_{b,d} \wedge S_{(i,j,k,l)} \in S^W_{b,d}), \\ 0 & \text{if } (L^T_{(i,j+1)} > L^W_{(k,l+1)}/\eta_{\max}) \text{ or} \\ & \text{if } (S_{(i,j,k,l)} \in S^T_{b,d} \wedge S_{(i,j,k,l)} \notin S^W_{b,d}), \\ 0.5 & \text{otherwise}, \end{cases} \tag{21}$$

which account for the load levels in each of the corresponding RATs given voice call and data session arrivals.

4.5 Multimode Terminal-Driven (MMTD) RAT Selection Policy

With the purpose of taking advantage of terminal availability characteristics, we may use this information to decide the most appropriate RAT for an incoming call/session. In this sense, we may attempt to allocate single-mode users to TDMA and multimode users to WCDMA. Multimode users would eventually be allocated to TDMA if no capacity was left in WCDMA. With this policy, we try to minimize the impact of single-mode terminals being served in TDMA given the higher allocation flexibility of multimode terminals. To account for the situations where no voice or data capacity is available in WCDMA and consequently multimode users are allocated, if possible, in TDMA, we define the following indicator functions, $\alpha_{(i,j,k,l)}$ and $\beta_{(i,j,k,l)}$ in (14), for each feasible state $S_{(i,j,k,l)}$:

$$\alpha_{(i,j,k,l)} = \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,v}, \\ 0 & \text{if } S_{(i,j,k,l)} \in S^W_{b,v}, \end{cases} \qquad \beta_{(i,j,k,l)} = \begin{cases} 1 & \text{if } S_{(i,j,k,l)} \notin S^W_{b,d}, \\ 0 & \text{if } S_{(i,j,k,l)} \in S^W_{b,d}. \end{cases} \tag{22}$$

5 Performance Metrics

In order to compute the steady-state probabilities $P_{(i,j,k,l)}$, we must solve the global SSBEs given by the application of the aforementioned RAT selection policies for all feasible states $S_{(i,j,k,l)} \in S$. This may be carried out using numerical methods; in particular, an iterative power procedure will be utilized for such task [19]. In general, the dimensionality of the Markov chain, M_d, can be computed as the product of K available RATs and J supported services (assuming all services are supported on all RATs). Obviously, the higher the number of services and/or RANs, the higher the dimensionality of our model. As a result, the computational complexity to numerically solve these systems is a well-known fact and increases with the state dimension. Nevertheless, typically up to 3 or 4 RANs are available and, although services are high in number, not all RATs support all services, which may lower the impact on the Markov dimensionality. In addition, the higher the number of states in the Markov model, N_s, the more computation resources are needed as explained in the following. For the particular case of two services, voice and data, along with two RATS, i.e., TDMA and WCDMA, the resulting number of states can be computed as

$$N_s = N^T_s \cdot N^W_s \tag{23}$$

with the number of states in TDMA, N^T_s, being

$$N^T_s = \frac{(C+1)(n_C C + 2)}{2} \tag{24}$$

and the number of states in WCDMA, N_s^W, yielding

$$N_s^W \approx \left\lfloor \frac{\left(\eta_{\max} \cdot \left[\frac{W/R_{b,v}}{(E_b/N_0)_v}+1\right]+1\right)\left(\eta_{\max} \cdot \left[\frac{W/R_{b,d}}{(E_b/N_0)_d}+1\right]+1\right)}{2} \right\rfloor, \quad (25)$$

where $\lfloor x \rfloor$ denotes the integer value of x.

In our case, the iterative power method is used to solve the system of equations provided in (15). Its operation is based on iteratively performing the product of a probability vector $\mathbf{p}$ (of dimension $N_s \times 1$) with the $N_s \times N_s$ transition probability matrix ($\boldsymbol{P}$). If k iterations are needed for convergence, then a total number of $k \times N_s^2$ multiplications are needed. The number of k iterations needed to satisfy convergence is based on the following relative measure [19]:

$$\max_i \left(\frac{\left| p_i^{(k)} - p_i^{(k-1)} \right|}{p_i^{(k)}} \right) < \varepsilon, \quad (26)$$

where $p_i^{(k)}$ are elements of vector $\mathbf{p^{(k)}}$, which denotes the probability distribution after the kth iteration, and ε is the required solution accuracy which is in our case set to 10^{-6}.

Fortunately, matrix $\boldsymbol{P}$ is usually sparse, i.e., it contains a large amount of zero entries. Then, if N_z is the total number of nonzero entries in matrix $\boldsymbol{P}$, a total of $k \times N_z$ multiplications are now required. In this sense, the limiting factor would be in terms of memory storage requirements rather than in terms of computational complexity of operation and solution convergence time. Nevertheless, state-of-the-art computers are able to support these high memory storage requirements.

Then, performance metrics may be directly derived from the steady-state probabilities, $P_{(i,j,k,l)}$, as described in the following.

5.1 Blocking Probabilities

Making use of the blocking state sets defined previously in Section 3.3, the generalized form of the blocking probability of a service type σ in a given RAT ρ may be expressed as

$$P_{b,\sigma}^{\rho} = \sum_{S_{(i,j,k,l)} \in S_{b,\sigma}^{\rho}} P_{(i,j,k,l)} \quad (27)$$

with $\sigma = \{v, d\}$ and $\rho = \{T, W\}$.

If we are interested in the blocking probability of a particular service type σ over all the possible RATs, this can be computed as

$$P_{b,\sigma} = \sum_{S_{(i,j,k,l)} \in S_{b,\sigma}} P_{(i,j,k,l)}. \quad (28)$$

Finally, the total blocking probability may be computed as

$$P_b = \sum_{S_{(i,j,k,l)} \in S_b} P_{(i,j,k,l)}. \quad (29)$$

5.2 Carried Traffic

The average carried traffic, or average number of users, may also be computed from the steady-state probabilities $P_{(i,j,k,l)}$. The fractional average number of users demanding a given service σ in a given RAT ρ can be derived numerically from

$$N_{\sigma}^{\rho} = E[x] \quad \text{with } x = \begin{cases} i & \text{if } \rho = T, \sigma = v, \\ j & \text{if } \rho = T, \sigma = d, \\ k & \text{if } \rho = W, \sigma = v, \\ l & \text{if } \rho = W, \sigma = d, \end{cases} \quad (30)$$

and $E[x]$ the expectation of x defined as

$$E[x] = \sum_{S_{(i,j,k,l)} \in S} x \cdot P_{(i,j,k,l)}. \quad (31)$$

Similarly, the average number of users in each RAT ρ is computed as

$$N^{\rho} = N_v^{\rho} + N_d^{\rho}. \quad (32)$$

The per-service average number of users in the system is defined by

$$N_{\sigma} = N_{\sigma}^T + N_{\sigma}^W. \quad (33)$$

Finally, the total average number of users in the system yields

$$N = N_v^T + N_d^T + N_v^W + N_d^W. \quad (34)$$

5.3 System Load

Load metrics are also key performance indicators which can be obtained from the steady-state probabilities. Bearing in mind the load definitions given in (20) and (6), the average TSL utilization factor in TDMA yields

$$L^T = E\left[L_{(i,j)}^T\right], \quad (35)$$

and the average UL load factor in WCDMA may be computed as

$$L^W = E\left[L_{(k,l)}^W\right]. \quad (36)$$

5.4 Peak Throughput

Throughput definitions are also intrinsic to the underlying access scheme and will be, consequently, defined individually for TDMA and WCDMA systems.

5.4.1 TDMA Throughput

The throughput in TDMA at a given state $S_{(i,j,k,l)}$ can be expressed as the sum of voice and data throughput contributions as

$$\Gamma_{(i,j)}^T = i \cdot \kappa_v + \min(C - i, j) \cdot \kappa_d, \quad (37)$$

where κ_v and κ_d are the voice and data TSL bit rates, respectively, and the term $\min(C-i, j)$ accounts for the number of data users transmitting at κ_d bits per second. If i voice users are being served in GSM/EDGE Radio Access Network (GERAN), and they require a whole TSL, then at most $(C-i)$ data users will be able to transmit at κ_d bits per second. If $j < (C-i)$, then j data users transmit at κ_d bits per second.

It is important to note that although the throughput per voice user will be κ_v, for data users, the effect of TSL sharing will contribute to a decrease in throughput per

TABLE 1
System Parameters for Numerical Evaluation

Parameter	Symbol	Value
Number of channels in GERAN	C	8
Maximum number of simultaneous users sharing a same TSL in GERAN	n_C	3
Bit rate for voice users in GERAN	κ_v	12.2 kbps
Bit rate for data users in GERAN	κ_d	44.8 kbps
Chip-rate in UTRAN	W	3.84 Mcps
Required bit-energy-to-noise-density ratio for voice traffic in UTRAN	$(E_b/N_0)_v$	6 dB
Required bit energy-to-noise-density ratio for data traffic in UTRAN	$(E_b/N_0)_d$	5 dB
Bit rate for voice users in UTRAN	$R_{b,v}$	12.2 kbps
Bit rate for data users in UTRAN	$R_{b,d}$	44.8 kbps
Maximum UL load factor	η_{max}	1
Multi-mode terminal availability (unless otherwise stated)	p	1

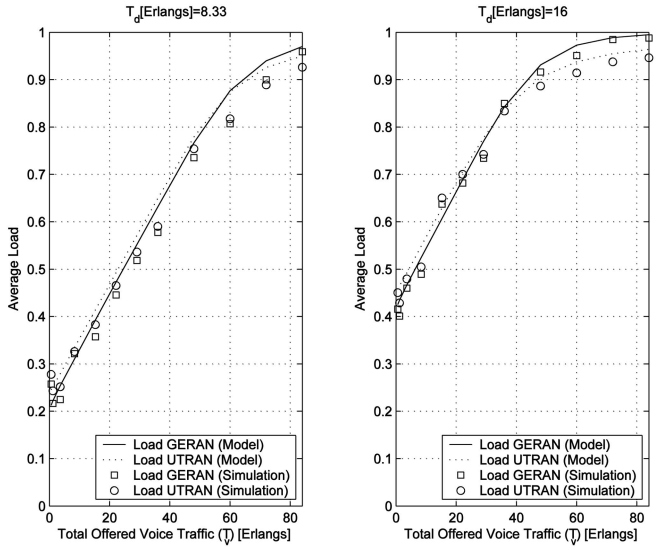

Fig. 3. Average load in each RAT under varying traffic.

data user as the number of multiplexed data TSLs increases. Actually, the throughput per data user will be equal to $\kappa_d \cdot \min(C-i,j)/j$.

Then, the total average throughput in TDMA becomes

$$\Gamma^T = E\left[\Gamma^T_{(i,j)}\right]. \tag{38}$$

5.4.2 WCDMA Throughput

Throughput delivered in WCDMA-based systems at a given state $S_{(i,j,k,l)}$ can be calculated as

$$\Gamma^W_{(k,l)} = k \cdot R_{b,v} + l \cdot R_{b,d}, \tag{39}$$

where $R_{b,\sigma}$ is the granted bit rate of a σ service type user.

Then, the average throughput in WCDMA is obtained as

$$\Gamma^W = E\left[\Gamma^W_{(k,l)}\right]. \tag{40}$$

5.4.3 Total Aggregate Throughput

Considering the combined throughput carried by both RATs, TDMA, and WCDMA, the total aggregate throughput, Γ_A, becomes

$$\Gamma_A = \Gamma^T + \Gamma^W. \tag{41}$$

6 Results

In order to illustrate the performance of the presented RAT selection policies, the GERAN and the UMTS Radio Access Network (UTRAN) will be used as representatives of TDMA and WCDMA technologies, respectively.

The performance of the system is evaluated under different offered voice and data traffic loads, T_v and T_d, where $T_v = \lambda_v/\mu_v$ and $T_d = \lambda_d/\mu_d$. The considered system parameters for numerical evaluation are represented in Table 1.

Under these assumptions, and as will be shown in the following numerical results, UTRAN exhibits a higher capacity in terms of maximum number of allowable voice and data users as compared to GERAN. Indeed, the $C = 8$ channels in GERAN correspond to a 200-kHz bandwidth single-carrier configuration, while for UTRAN a total bandwidth of 5 MHz is available [15].

6.1 Markov Model Validation

In order to validate the results provided by the Markov model, a system-level simulator has been developed. This simulator assumes a more realistic behavior than the model by considering that data users intend to transmit a particular amount of data (bits), which follows a Pareto distribution [20]. In this case, the data holding time will depend on the bit rate allocated to the user in the selected RAT rather than being modeled by an exponential distribution. RAT selection is performed and CAC procedures follow the same principles as the state feasibility conditions imposed for the Markov model. Once users are allocated in the appropriate RAT, statistics are measured on a discrete-time basis. In addition, another simulator considering the same assumptions as in the Markov model has been used to validate the correctness of the algebra, but not shown here due to lack of space.

In the following, we compare the results obtained via the Markov model with the results obtained through simulation using the LB criterion for the RAT selection policy. Fig. 3 shows the loads in GERAN and UTRAN considering an offered traffic load of 16 and 8.33 Erlangs and a range of voice traffic for the LB case. Loads in both RATs tend to follow each other as the total offered traffic increases. Note how the simulated values (represented as bullets) follow the same trend to those obtained via the Markov model and a good matching exists. Fig. 4 shows the number of served users of each service in UTRAN and GERAN. Bearing in mind that, with the current parameter setting (see Table 1), GERAN offers a much lower capacity than UTRAN (i.e., one single carrier of 200 kHz for GERAN as opposed to 5 MHz for UTRAN); many more users are needed in UTRAN in order to balance the loads. The number of data users in each RAT is kept constant while offered voice traffic varies between 1.2 and 36 Erlangs. For 84 Erlangs, data users are

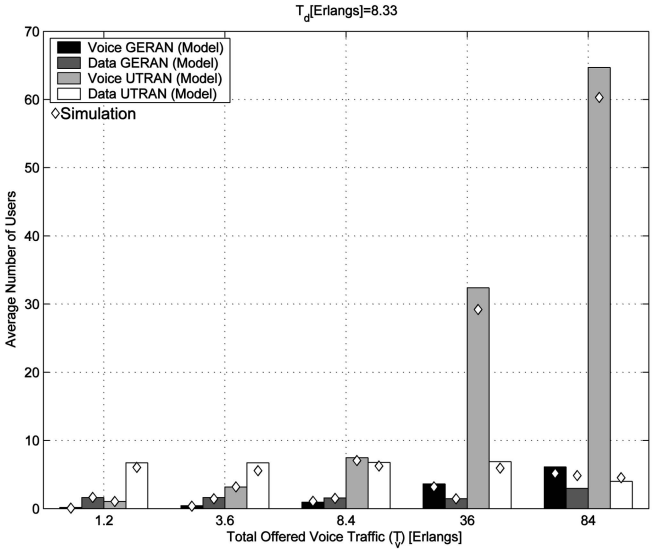

Fig. 4. Average number of served users in each RAT.

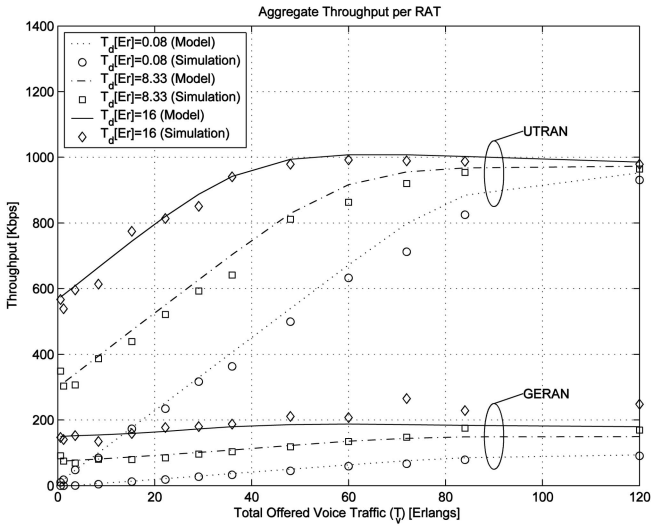

Fig. 6. Total throughput per RAT under varying traffic.

forced to share TSLs in GERAN, explaining the increase/decrease of data users in GERAN and UTRAN, respectively. Fig. 5 shows the total voice blocking probability in the combined GERAN/UTRAN system as defined in (28). Clearly, the higher the offered traffic the higher the blocking probability gets. Again, the simulated results (marked with bullets) match the Markov model behavior. Finally, Fig. 6 shows the throughput performance in each of the RATs for both the model and the simulated approaches under varying traffic conditions. At this point, the proposed Markov model has been validated, as shown in Figs. 3, 4, 5, and 6, and its suitability for testing different RAT selection policies confirmed. The comparison of such RAT selection policies is provided in the forthcoming sections.

6.2 RAT Selection Policies Comparison

This section provides illustrative results depicting the behavior of the presented RAT selection policies (except for policy MMTD, given that $p = 1$ and, thus, does not apply in this case). The focus is set on how the different policies allocate different services over the existing RATs by means of representing the probability, $P_{(i,j,k,l)}$, of having a given number of voice and data users in each RAT for several traffic mix conditions. In the following, statistical user distribution is represented with 2D discrete graphs with axis indicating the number of voice and data users, and probabilities depicted by gray-scaled shaded regions, where dark regions indicate high probability values and light regions low probability values. Stepwise admission limits are plotted for both systems (denoted as *feasibility region*) and, for the case of GERAN, the limit upper-bounding the region where data users are not sharing resources is also plotted (which is denoted as *data non-reuse region*).

In this sense, Fig. 7 shows the user distribution provided by policy SB#1. For a traffic mix of $T_v = 3.6$ and $T_d = 16$ Erlangs, GERAN is able to handle its share of

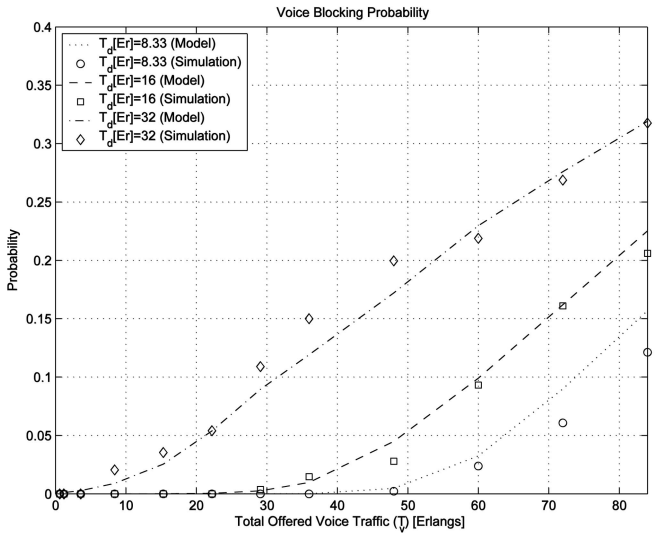

Fig. 5. Total voice blocking probabilities under varying traffic.

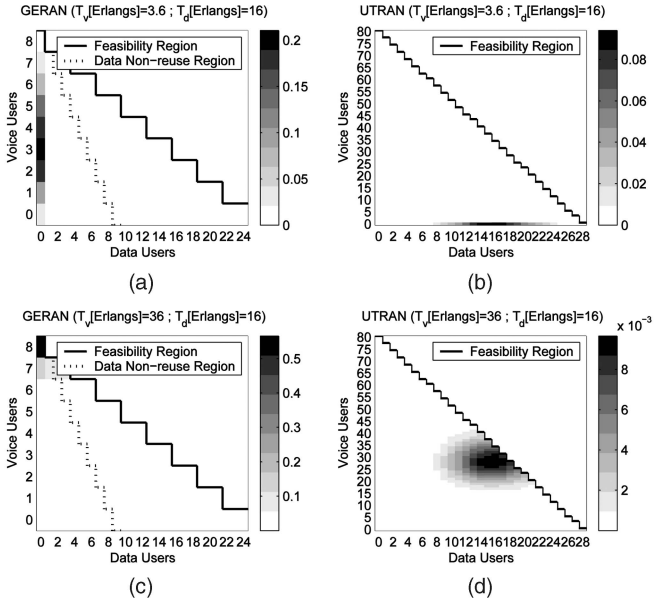

Fig. 7. Statistical user distributions in GERAN and UTRAN with policy SB#1 for two different service mixes.

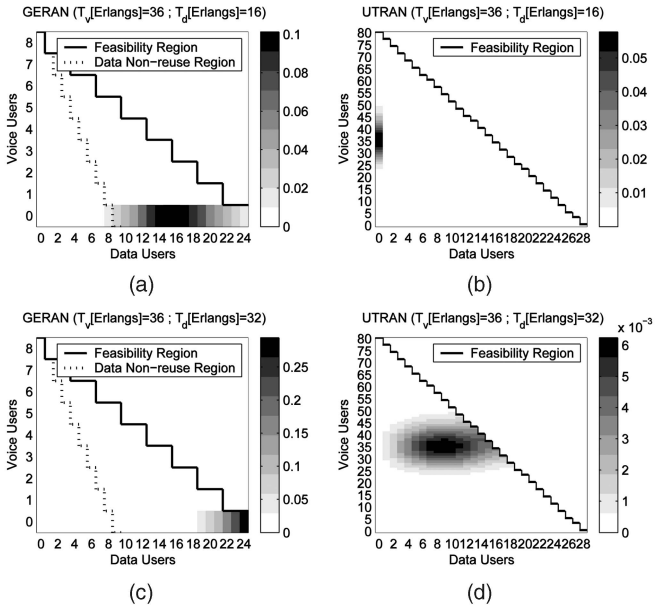

Fig. 8. Statistical user distributions in GERAN and UTRAN with policy SB#2 for two different service mixes.

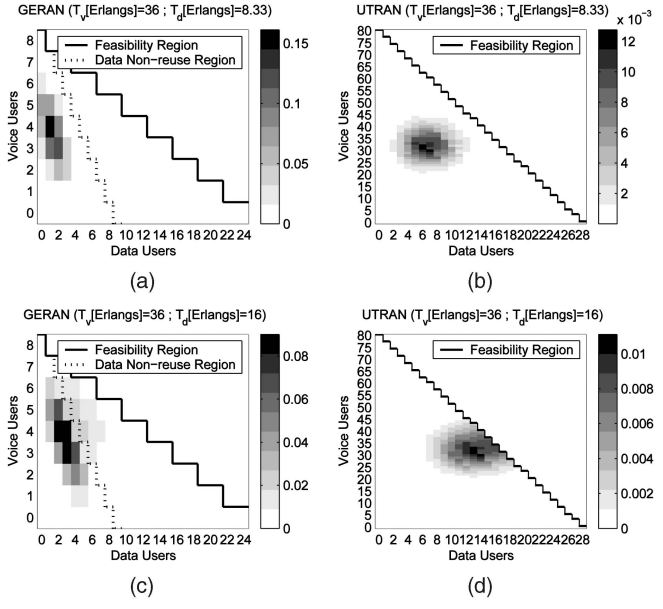

Fig. 9. Statistical user distributions in GERAN and UTRAN with policy LB for two different service mixes.

voice users, while UTRAN manages the offered data traffic. This causes users in GERAN being exclusively distributed over the voice user axis with no data users at all (Fig. 7a). Accordingly, in UTRAN, users are spanned over the data user axis with no voice user component (Fig. 7b). If voice traffic is increased so that GERAN is not able to handle all the requests, SB#1 policy will redirect voice traffic to UTRAN. Consequently, user distribution in GERAN is concentrated on the maximum number of allowed voice users (Fig. 7c). In UTRAN, users are now distributed over both axes due to overflowed voice traffic from GERAN (Fig. 7d). For the case of SB#2 policy, see Fig. 8, an analogous study may be made on the statistical user distributions. When both GERAN and UTRAN are able to manage their shares of data and voice users, respectively (Figs. 8a and 8b), data and voice user distributions lay on the corresponding axis in GERAN and UTRAN, respectively. If data traffic is increased so that GERAN is unable to handle all data traffic, UTRAN will have to manage with its share of voice users plus data users that could not be allocated in GERAN. This behavior can be observed in Figs. 8c and 8d. It is worth noting how SB#2 provides high reuse of data resources in GERAN. The rationale behind LB policy is to maintain both loads in GERAN and UTRAN at the same level. By doing so, and according to the load definitions presented earlier on, data users in GERAN will not be forced to share resources until UTRAN is fully loaded. This may be observed in Fig. 9. For offered traffic values of $T_v = 36$ and $T_d = 8.33$ Erlangs, UTRAN remains in a half-loaded situation given that user distribution remains far from the admission limit (see Fig. 9b). Consequently, in GERAN (Fig. 9a) data users do not share resources since user distribution lies below the data non-reuse region indicated by the dotted line. If traffic is increased such that UTRAN achieves fully loaded situations (Fig. 9d), data users in GERAN will then start to share resources which is indicated by user distribution falling above the data non-reuse region as observed in Fig. 9c. Finally, Fig. 10 shows the statistical user distribution when applying policy RND. As compared to LB policy, RND policy forces data users to share resources in GERAN even if they could be more efficiently managed by UTRAN system. Moreover, higher blocking situations in GERAN are achieved indicated by the proximity of user distribution to the admission limit.

6.3 Throughput Comparison

In the following, a comparison between the different RAT selection policies by evaluating a number of different performance measures is provided. Fig. 11 shows the performance in terms of aggregate throughput for the different presented RAT selection policies, i.e., SB#1, SB#2,

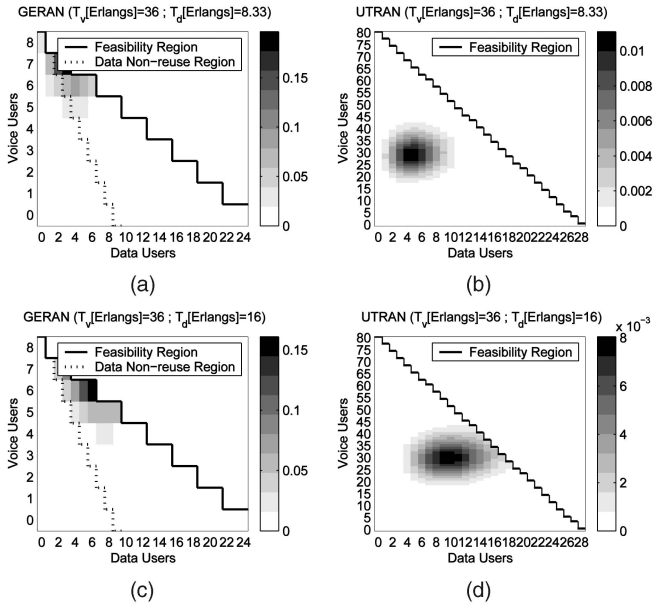

Fig. 10. Statistical user distributions in GERAN and UTRAN with policy RND for two different service mixes.

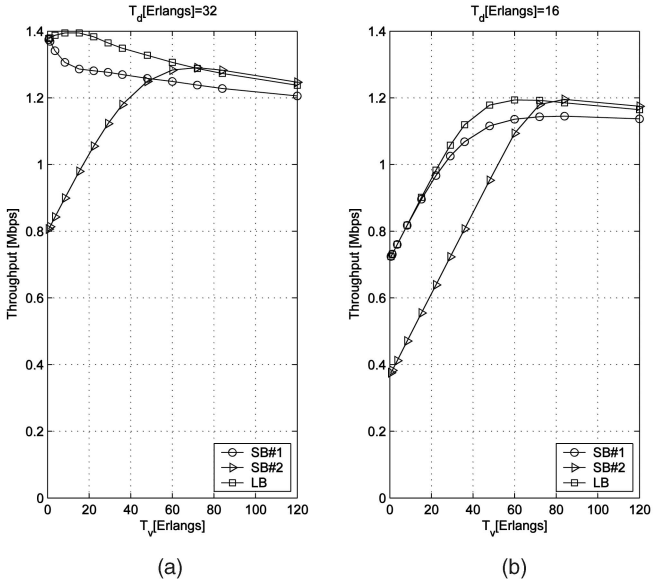

Fig. 11. Aggregate throughput values for policies SB#1, SB#2, and LB under varying traffic conditions.

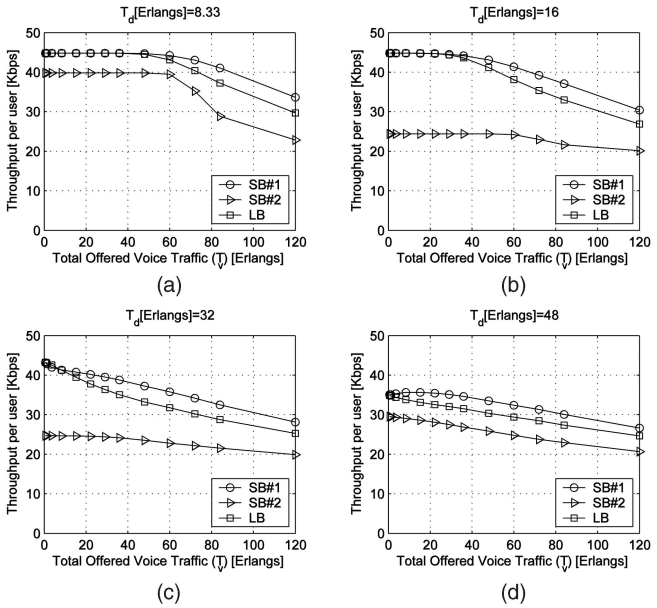

Fig. 12. Throughput per data user for different RAT selection policies and traffic mixes.

and LB (where RND policy is not shown given its poor performance and the lack of space). For medium data traffic loads (i.e., $T_v = 16$ and $T_d = 32$ Erlangs), we can clearly see how LB outperforms all other policies, with the worst overall behavior corresponding to SB#2. As it will be shown in the following, a major cause of throughput degradation is due to excessive TSL sharing in GERAN. In this sense, SB#1 allocates data users to UTRAN, so this problem is partially avoided. However, when voice load is increased, data users may be also allocated in GERAN causing throughput to exhibit similar performances than for other policies. In contrast, SB#2 allocates data users to GERAN causing high TSL reuse even for low voice loads thus negatively impacting on the total aggregate throughput. Notice that for the case of LB policy, load definitions provided in (20) and (6) prevent data users in GERAN to share TSL with other users unless UTRAN is fully loaded. Finally, an overall throughput degradation is noted when the offered voice traffic increases, in particular, when the offered data traffic is $T_d = 32$ Erlangs (Fig. 11a). This is caused by a major number of admitted voice users which contribute with lower throughput values (12.2 Kbps for a single voice user) than those offered by data users (a maximum of 44.8 Kbps for a single data user). As previously mentioned, a cause for aggregate throughput degradation may be found in the excessive reuse of data TSLs in GERAN. Fig. 12 shows, for several offered traffic configurations, the total throughput per data user when each of the considered RAT selection policies is applied. As it can be observed, SB#2 policy provides an excessive reuse of TSLs and thus throughput per data user is lower than that of SB#1 and LB policies. On the other hand, SB#1 policy will direct data users to UTRAN which exhibits a better throughput performance, and therefore, data users are less penalized by TSL reuse. Finally, LB policy will prevent data sharing in GERAN as long as UTRAN may handle offered traffic. In these cases, throughput per data user provided by LB policy is maximum, i.e., 44.8 Kbps (see Figs. 12a and 12b). However, if traffic increases, UTRAN will no longer be able to manage all traffic, and thus, LB will force data users in GERAN to share TSL which will consequently cause throughput per data user to decrease. Another important issue to take into account when designing RAT selection policies, is to provide a high ratio of admitted users in the system or, equivalently, to provide low blocking probabilities. Fig. 13 illustrates the blocking probability obtained by the presented policies when offered data traffic is $T_v = 16$ and $T_d = 32$ Erlangs. Clearly, SB#2 provides a lower blocking probability as compared to SB#1 and LB policies. In GERAN, the allocation of a voice user implies a TSL consumption of $1/C$ while as for a data user this consumption is $1/n_C C$. Therefore, it is more resource-consuming (n_C times more) to allocate voice users in GERAN than data users. On the other hand, the resource consumption in UTRAN may be quantified by means of the load factor definition, given in (6), with $[W/R_{b,v}(E_b/N_0)_v + 1]^{-1}$ and $[W/R_{b,d}(E_b/N_0)_d + 1]^{-1}$ the fractions of loads consumed by voice and data users,

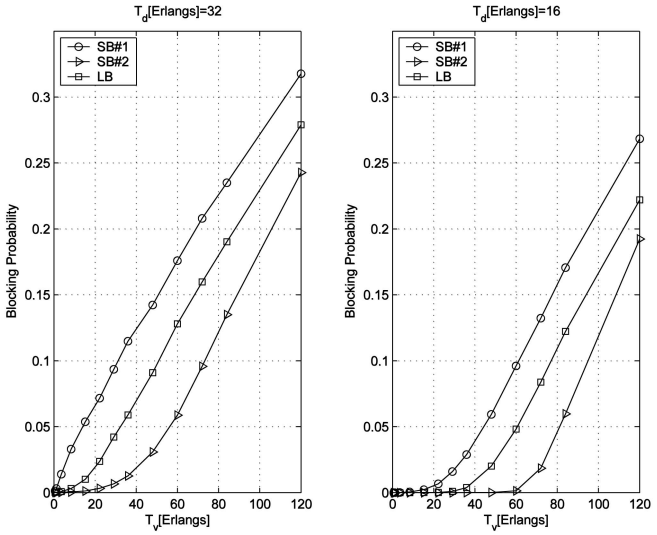

Fig. 13. Total blocking probabilities for policies SB#1, SB#2, and LB.

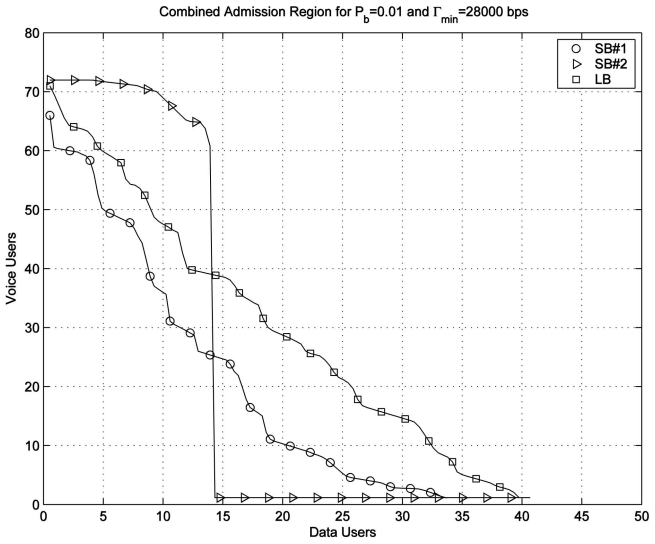

Fig. 14. Combined admission regions for policies SB#1, SB#2, and LB in a blocking probability limited scenario.

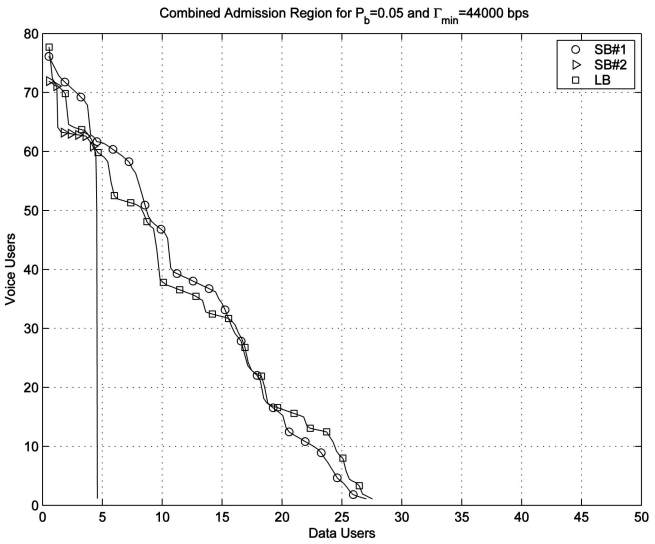

Fig. 15. Combined admission regions for policies SB#1, SB#2, and LB in a throughput limited scenario.

respectively. Bearing in mind the simulation parameters given in Table 1, it can be shown that a data user in UTRAN demands more resources than a voice user. In this sense, it is much more suitable in the considered scenario to allocate voice users in UTRAN and data users in GERAN, thus explaining the better behavior of SB#2 as opposed to SB#1. LB, on the other side, will provide a performance in-between SB#2 and SB#1.

As previously stated, to choose the most suitable RAT selection policy may depend on a number of quality requirements. In particular, it has been studied that both the data throughput per user and the blocking probability can decide the appropriateness of one RAT selection policy with respect to another. Therefore, in the following, the achievable capacity in terms of the maximum number of allowable voice and data users (admission regions) satisfying some blocking probability and data throughput per user requirements will be studied. Based on these measurements, the most appropriate RAT selection policy may be decided.

Fig. 14 shows the admission regions achieved by the different policies under equal traffic conditions. In this first case, a blocking probability limited scenario is considered where it is required a maximum blocking probability of $P_b = 0.01$ and a minimum data throughput per user of $\Gamma_{min} = 28\,\mathrm{Kbps}$ (recall that the maximum achievable throughput per data user is 44.8 Kbps). Results in Fig. 14 indicate that although SB#2 policy provides a larger admission region than SB#1, given by a better response in terms of blocking probability, the SB#2 policy is severely limited by the data throughput per user requirements for high data loads. This causes policy LB to better trade off both blocking probability and throughput requirements.

Fig. 15 considers a data throughput per user limited scenario, where the target blocking probability is set to 5 percent and the minimum required data throughput per user is $\Gamma_{min} = 44\,\mathrm{Kbps}$. In this case, SB#2 no longer provides a larger admission region than SB#1 due to the throughput requirements. As expected, SB#1 provides lower data resource utilization in GERAN thus being capable of allocating somewhat more users than policy LB.

Finally, if the system is limited by both blocking and throughput requirements, via setting $P_b = 0.01$ and $\Gamma_{min} = 44\,\mathrm{Kbps}$, the performance shown in Fig. 16 is obtained. Once again, throughput requirements are too severe for SB#2 and, thus, provide a smaller admission region than SB#1 and LB. LB, on the other hand, exhibits a better performance in terms of throughput and blocking probability than SB#1, therefore providing the largest admission region.

6.4 Multimode Terminal Availability Impact on Initial RAT Selection

Fig. 17 reflects the impact of multimode terminal availability over the performance in terms of total aggregate

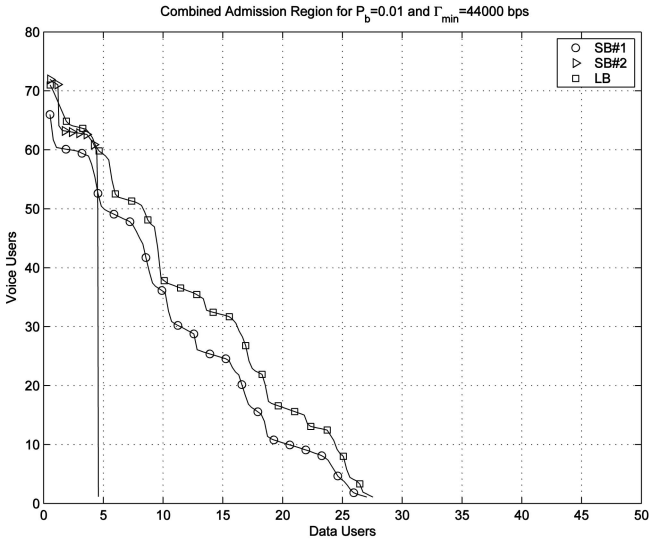

Fig. 16. Combined admission regions for policies SB#1, SB#2, and LB in a blocking probability and throughput limited scenario.

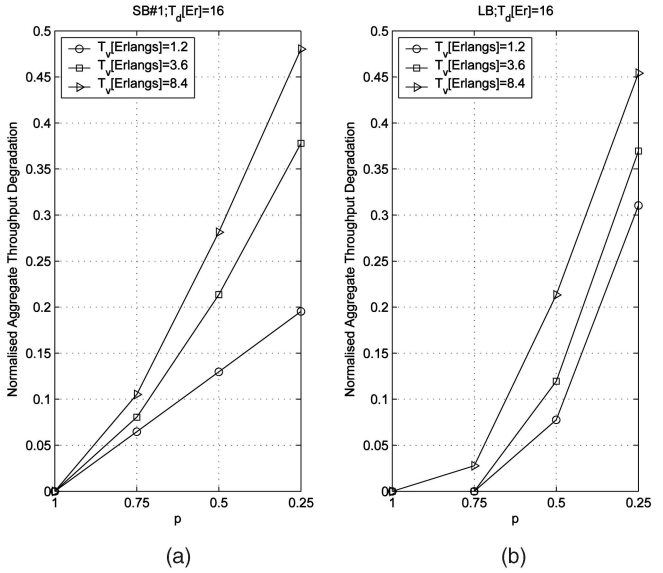

Fig. 17. Impact of multimode terminal probability (p) when applying (a) SB#1 and (b) LB policies.

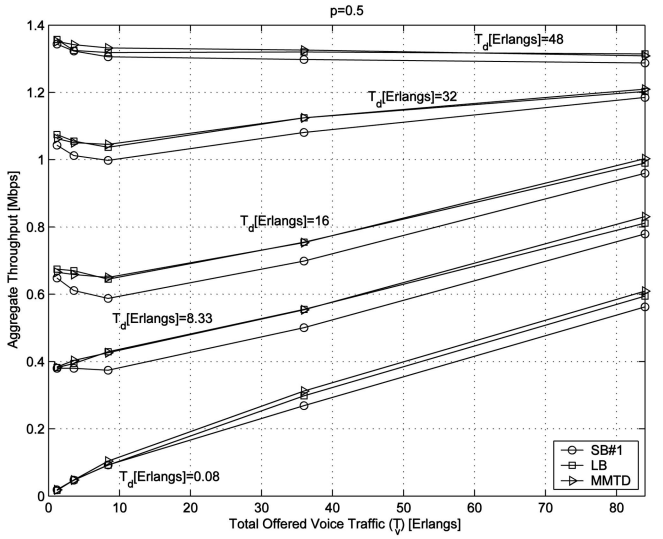

Fig. 18. Total aggregate throughput provided by SB#1, LB, and MMTD policies.

throughput when using policies SB#1 (Fig. 17a) and LB (Fig. 17b). Specifically, we plot the normalized throughput degradation measured as the relative difference in aggregate throughput when single-mode terminals are present with respect to the case of all terminals being multimode, mathematically expressed as

$$D^p = \left(\Gamma_T^1 - \Gamma_T^p\right)/\Gamma_T^1, \tag{42}$$

where Γ_T^p is the total aggregate throughput for a multimode terminal probability equal to p. The overall behavior of having low number of multimode terminals is translated into a higher throughput degradation which is represented in Fig. 17.

Fig. 18 compares, in terms of aggregate throughput, policies MMTD, SB#1, and LB in a scenario with 50 percent of terminals with multimode capabilities. With SB#1, GERAN handles voice users plus single-mode users, thus showing a poorer performance than MMTD which allocates voice multimode users to UTRAN. On the other hand, LB policy shows higher flexibility in allocating multimode users as compared to SB#1. As a result, similar performance is observed between LB and MMTD.

7 Concluding Remarks

It is widely established that RAT selection procedures in multiservice/multiaccess scenarios play a key role in the provision of CRRM functionalities. In this paper, a Markovian framework for the allocation of multiple services in multiple RATs is presented. It allows the evaluation of several RAT selection policies considering different allocation criteria which are fully embedded in the model. In addition, the model captures the availability of multimode terminals so as to consider the flexibility constraints of single-mode terminals. In particular, two different underlying radio access schemes are studied: TDMA and WCDMA. In this context, generic voice and data sessions are to be allocated to the aforementioned RATs given particular RAT selection policies which comprise: two service-based schemes, namely SB#1 and SB#2, along with a LB and a terminal-driven (MMTD) schemes, and finally, a random policy (RND). Results have confirmed the validity and suitability of the model which has been evaluated for the aforementioned RAT selection policies. Results indicate that a tradeoff between the average data throughput per user and the total blocking probability arises when comparing SB#1 and SB#2. As revealed, this tradeoff may be suitably managed by the appliance of the LB policy. Finally, RAT selection is also performed taking into account the multimode terminal availability information, indicating that this input must not be avoided to achieve a higher utilization of the offered resources. At this point, the authors believe that the proposed framework provides an appropriate platform for the development and evaluation of CRRM functions within the scope of multiservice/multiaccess networks.

Acknowledgments

This work has been supported by the Spanish Research Council under COGNOS Grant (ref. TEC2007-60985).

References

[1] A.K. Salkintzis, "Interworking Techniques and Architectures for WLAN/3G Integration Toward 4G Mobile Data Networks," *IEEE Wireless Comm.*, vol. 11, no. 3, pp. 50-61, June 2004.

[2] E. Gustafsson and A. Jonsson, "Always Best Connected," *IEEE Wireless Comm.*, vol. 10, no. 1, pp. 49-55, Feb. 2003.

[3] J. Pérez-Romero et al., "Common Radio Resource Management: Functional Models and Implementation Requirements," *Proc. 16th IEEE Int'l Symp. Personal, Indoor and Mobile Radio Comm. (PIMRC '05)*, vol. 3, pp. 2067-2071, Sept. 2005.

[4] *Improvement of RRM Across RNS and RNS/BSS*, 3GPP TR 25.881 v5.0.0, http://www.3gpp.org/, 2008.

[5] *Improvement of RRM Across RNS and RNS/BSS (Post Rel-5) (Release 6)*, 3GPP TR 25.891 v0.3.0, http://www.3gpp.org/, 2008.

[6] A. Tölli, P. Hakalin, and H. Holma, "Performance Evaluation of Common Radio Resource Management (CRRM)," *Proc. IEEE Int'l Conf. Comm. (ICC '02)*, vol. 5, pp. 3429-3433, Apr.-May 2002.

[7] J. Pérez-Romero, O. Sallent, and R. Agustí, "Policy-Based Initial RAT Selection Algorithms in Heterogeneous Networks," *Proc. Seventh IFIP Int'l Conf. Mobile and Wireless Comm. Networks (MWCN '05),* Sept. 2005.
[8] G. Fodor, A. Furuskär, and J. Lundsjo, "On Access Selection Techniques in Always Best Connected Networks," *ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems,* Aug. 2004.
[9] O. Yilmaz, A. Furuskär, J. Pettersson, and A. Simonsson, "Access Selection in WCDMA and WLAN Multi-Access Networks," *Proc. 61st IEEE Vehicular Technology Conf. (VTC '05),* Spring, vol. 4, pp. 2220-2224, May-June 2005.
[10] S. Lincke-Salecker and C.S. Hood, "Integrated Networks that Overflow Speech and Data between Component Networks," *Int'l J. Network Management,* vol. 12, no. 4, pp. 235-257, John Wiley & Sons, July-Aug. 2002.
[11] A. Furuskär and J. Zander, "Multiservice Allocation for Multiaccess Wireless Systems," *IEEE Trans. Wireless Comm.,* vol. 4, no. 1, pp. 174-184, Jan. 2005.
[12] I. Koo, A. Furuskär, J. Zander, and K. Kim, "Erlang Capacity of Multiaccess Systems with Service-Based Access Selection," *IEEE Comm. Letters,* vol. 8, no. 11, pp. 662-664, Nov. 2004.
[13] G. Bolch, S. Greiner, H. Meer, and K.S. Trivedi, *Queueing Networks and Markov Chains: Modelling and Performance Evaluation with Computer Science Applications.* John Wiley & Sons, 1998.
[14] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS.* John Wiley & Sons, 2002.
[15] J. Pérez-Romero, O. Sallent, R. Agustí, and M. Díaz-Guerra, *Radio Resource Management Strategies in UMTS.* John Wiley & Sons, 2005.
[16] C. Bettstetter, H.-J. Vogel, and J. Eberspacher, "GSM Phase 2+, General Packet Radio Service GPRS: Architecture, Protocols and Air Interface," *IEEE Comm. Surveys,* vol. 2, no. 3, 1999.
[17] M. Ermel, K. Begain, T. Muller, J. Schuler, and M. Schweigel, "Analytical Comparison of Different GPRS Introduction Strategies," *Proc. Third ACM Int'l Workshop Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '00),* pp. 3-10, 2000.
[18] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "On the Suitability of Load Balancing Principles in Heterogeneous Wireless Access Networks," *Proc. Int. Symp. Wireless Personal Multimedia Comm. (WPMC '05),* Sept. 2005.
[19] W.J. Stewart, *Introduction to the Numerical Solution of Markov Chains.* Princeton Univ. Press, 1994.
[20] *Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS,* Universal Mobile Telecommunications System (UMTS), TR 101 112 V3.2.0 (1998-04), UMTS 30.03 version 3.2.0.

Xavier Gelabert received the Telecommunications Engineering degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2004 and the MS degree in electrical engineering, with a major in wireless communications, from the Royal Institute of Technology (KTH), Stockholm, in 2003. In 2004, he joined the Radio Communication Research Group in the Department of Signal Theory and Communications, UPC, where he is currently a PhD candidate. His master thesis was performed at the Radio Communication Systems Lab (KTH) under the title "Synchronization Strategies in STDMA Tactical Radio Access Networks". His current research interests include the field of mobile radio communication systems, with a special emphasis on Common Radio Resource Management (CRRM) strategies in multiaccess networks, quality of service provisioning, and opportunistic/cognitive spectrum management. He has been actively involved in European-funded projects EVEREST, AROMA, and E3 along with Spanish projects COSMOS and COGNOS. He is a student member of the IEEE.

Jordi Pérez-Romero received the Telecommunications Engineering and PhD degrees from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 1997 and 2001, respectively. He is an associate professor in the Department of Signal Theory and Communications, UPC. He joined the Radio Communications Group of the department in 1998 and, since then, has been involved in different projects in the field of mobile communication systems, especially focusing on packet radio techniques, spread-spectrum systems, radio resource and QoS management, heterogeneous wireless networks, and spectrum management. He has been involved in different European Projects (WINEGLASS, ARROWS, EVEREST, E2R phases I and II, E3, NEWCOM, and AROMA) as well as in projects for private companies like Telefónica and Alcatel. He has published several papers in IEEE journals and conferences and serves as an associate editor in *IEEE Vehicular Technology Magazine*. He has also published the book "Radio Resource Management strategies in UMTS", from Ed. John Wiley & Sons. He is a member of the IEEE.

Oriol Sallent is an associate professor in the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona. He received the Telecommunications Engineering and PhD degrees from the UPC, Barcelona, in 1994 and 1997, respectively. His research interests include the field of radio resource and spectrum management for heterogeneous cognitive wireless networks, where he has published more than 100 papers in IEEE journals and conferences. He has participated in many research projects and consultancies funded by either public organizations or private companies. He is currently participating in E3 project within the seventh Framework Program of the European Commission.

Ramon Agustí received the Engineer of Telecommunications degree from the Universidad Politécnica de Madrid, Madrid, in 1973 and the PhD degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 1978. In 1973, he joined the Escola Tècnica Superior d'Enginyers de Telecomunicació de Barcelona, Barcelona, where he became a full professor in 1987. After graduation, he was working in the field of digital communications with particular emphasis on transmission and development aspects in fixed digital radio, both radio relay and mobile communications. For the last 15 years, he has been mainly concerned with the performance analysis and development of planning tools and equipment for mobile communication systems. He has published about 200 papers in those areas. He participated in the European program COST 231 and in the COST 259 as the Spanish representative delegate. He has also participated in the RACE, ACTS, IST European research programs as well as in many private and public funded projects. He is the recipient of the Catalonia Engineer of the Year in 1998 and the Narcís Monturiol Medal issued by the Government of Catalonia in 2002 for his research contributions to the mobile communications field. He is part of the editorial board of several scientific international journals. Since 1995, he has been conducting a postgraduate annual course on mobile communications. He coauthored two books on mobile communications. He is a member of the IEEE.