Radio Access Congestion in Multiaccess/Multiservice Wireless Networks

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

***Abstract*—This work addresses the problem of radio access congestion control and resource allocation in scenarios where multiple available radio access technologies (RATs) support a wide range of services over a given coverage area. A key issue in these networks is selecting the most appropriate RAT at a call/session establishment according to some specified user/operator criteria. In this sense, a wide range of high-level policies can be defined, providing the most favorable resource allocation. Regardless of having efficient RAT selection policies, which may ensure some initial quality of service (QoS) requirements, intrinsic network dynamics (e.g., mobility, user activity, and interference rise) can cause potential QoS failures, leading to a degraded network performance and, hence, radio access congestion. This work is devoted to the study of the impact of radio access congestion on a number of RAT-selection policies. Consequently, a congestion probability (CP) model is developed to capture the statistical behavior of radio access congestion events. In addition, a general Markovian framework is adopted to evaluate the allocation of multiple services into multiple RATs by means of high-level policy definitions. Specific RAT-selection policies are defined according to several criteria, and their performances are evaluated in a time-division multiple-access (TDMA)/wideband code-division multiple-access (WCDMA) multi-RAT scenario supporting voice and data services. Moreover, the use of CP information as a possible allocation principle for RAT selection is also evaluated, which, in the assumed scenario, results in the most favorable allocation policy.**

***Index Terms*—Beyond 3G (B3G), mobile communication systems, radio access congestion control, radio resource management (RRM).**

I. Introduction

COEXISTENCE, coordination, and cooperation are key features of actual wireless network deployments. Operators and manufacturers in the wireless arena are faced with the challenge of managing, in the most efficient way, multiple radio access technologies (RATs) supporting a wide range of services in a specific coverage area. A typical example of these scenarios comprises the coexistence of cellular-based systems such as the Global System for Mobile communications (GSM)/Enhanced Data Rates for GSM evolution (GSM/EDGE) and the Universal Mobile Telecommunications System (UMTS). In addition, non-cellular systems such as wireless local area networks and wireless metropolitan area networks may extend communication ranges to other domains, such as home or rural environments.

A coordinated use of the total amount of radio resources offered by several RATs leads to improved performances, as opposed to considering each RAT as a stand-alone system. This potential gain is commonly referred to as *trunking gain* and was evaluated in [1]. In addition, a flexible architecture that allows interworking and cooperation across the considered RATs is needed to enable such coordinated use of resources. The level of interworking between network entities and the considered deployment topologies will, in great measure, impact the overall performance, as revealed in, e.g., [2] and [3]. This concept of integrated network comprising several RATs is commonly referred in the literature as heterogeneous or Beyond 3G (B3G) networks.

In this context, the notion of common radio resource management (CRRM)[1] has been developed to provide architectures and strategies to achieve the highest and most efficient utilization of the scarce radio resources provided by B3G networks. Here, the term *common* emphasizes the fact that resources provided by the different RATs should be regarded as a whole. Consequently, well-known radio resource management (RRM) strategies [4] such as call admission control (CAC), congestion control (CC), and packet scheduling (PS) should be redefined in this new context, leading to common CAC, common CC (CCC), common PS, etc. In addition to these redefined CRRM functionalities, the problem of selecting the most appropriate RAT according to some specified user/operator criterion is nontrivial. As a consequence, a RAT-selection functionality appears, which can be carried out at call/session initiation, in which case, we will refer to it as initial RAT selection, and/or during the call/session lifetime, which is usually denoted as vertical handover (VHO). To a great extent, the objective of initial RAT selection and VHO algorithms is to enable the so-called *Always-Best-Connected* paradigm [5], which ensures that a user is connected to the best RAT anytime and anywhere.

In this paper, a framework for the evaluation of CCC in B3G scenarios is addressed, and the impact of initial RAT-selection mechanisms is assessed.

A. Motivation and Problem Statement

Among the aforementioned well-known RRM strategies, CC is the function devoted to overcoming potential quality of

Manuscript received July 29, 2008; revised February 19, 2009 and April 30, 2009. First published May 27, 2009; current version published October 2, 2009. This work was supported in part by COGNOS Project TEC2007-60985 through the Spanish Ministry of Science and Education and the European Regional Development Fund. The review of this paper was coordinated by Prof. Y. Cheng.

The authors are with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain (e-mail: xavier.gelabert@tsc.upc.edu; sallent@tsc.upc.edu; jorperez@tsc.upc.edu; ramon@tsc.upc.edu).

Digital Object Identifier 10.1109/TVT.2009.2023902

[1]Also denoted as joint or multiple RRM in some works.

service (QoS) failures at the radio interface layer due to the intrinsic dynamics of the network (e.g., mobility, interference rise, and traffic variability) [6]. Regardless of having a strict CAC mechanism, which may ensure some QoS requirements at call/session establishment, if the dynamics of certain network parameters suffer from high random behavior, the network may experience high-load/high-interference situations, which, in turn, may degrade the QoS perceived by users. Consequently, CC strategies are designed to minimize the negative impact of radio access congestion on the network performance. Typically [6], CC mechanisms involve three phases: 1) congestion detection (CD); 2) congestion resolution (CR); and 3) congestion recovery (CRV). CD is responsible for monitoring the network status to correctly identify a congestion situation by means of RAT-specific measurements. On the other hand, CR actuates over a set of RAT-specific parameters to reduce the load and, consequently, the congestion situation. Finally, CRV will attempt to restore the transmission parameters that were set before the congestion was triggered.

In a B3G scenario with multiple RATs and multiple services, the congestion situations in each RAT can statistically be characterized by means of a *congestion probability (CP)*, given the number of allocated users. Then, provided a set of initial RAT selection policies, the evaluation of the congestion impact of these policies is of special interest, especially because favorable user allocations according to the CP in each of the existing RATs can be found. Consequently, the adopted approach in this paper is *proactive* in the sense that potential congestion problems are anticipated by means of efficient initial RAT selection, as opposed to *reactive*, i.e., acting after congestion is detected, which would include CR mechanisms such as, e.g., those in [7] and [8]. In addition, if the CP is known *a priori* (through models, real measurement campaigns, etc.) for a given user distribution, an initial RAT-selection policy can be designed using such information so that the resulting CP is kept as low as possible.

Under the aforementioned framework, this paper intends to define and assess the impact of radio access congestion in RATs with different underlying technologies, such as time-division multiple access (TDMA) and wideband code-division multiple access (WCDMA), which are subject to different traffic load conditions. The main objective is to determine user allocation principles at call/session establishment, i.e., initial RAT-selection policies, which are favorable in terms of potential congestion eventualities. In addition to the valuable interest of determining such initial RAT-selection policies, this will provide a solid basis for prospective evaluations contemplating the support of inter-RAT handovers (or VHOs, as defined earlier), which has not been considered in this study for the sake of model tractability. In addition, this work proposes a novel initial RAT-selection policy, which allocates users to minimize the resulting CP. Further detail on the contributions of this work is presented in the next section.

B. Related Work and Main Contributions

The Third-Generation Partnership Project (3GPP) standardization body identified some CRRM architectures and procedures for integrated GSM/EDGE and UMTS network operation in [9] and [10].

Despite the large amount of work devoted to CRRM algorithms in recent years (especially in those that refer to RAT selection, see, e.g., [11]–[15]), to the best of the author's knowledge, very few contributions have been made toward radio access CCC strategies. The authors in [16] proposed several schemes for controlling and distributing the network traffic over two RATs. Cost metrics are assigned to each service in each RAT, and load control is evaluated by means of initial RAT selection and VHO. In [7] and [8], the authors presented a framework for managing congestion situations in heterogeneous networks. In particular, practical methodologies for CD in GSM/EDGE and UMTS networks are provided, along with VHO and bit-rate reduction techniques to lessen the congestion status of the network.

While the aforementioned approaches usually rely on extensive system-level simulations, a theoretical approach to the problem is seemingly uncovered and thus constitutes the main novelty of this paper. It could then be of great interest to gain insight into the general problem and inspire the definition of practical and efficient congestion control strategies. In this sense, the main contributions of this work are given here.

1) To propose a general Markovian framework for the allocation of multiple services into multiple RATs by means of simple initial RAT-selection policies. This is covered in Section II and, along with the inclusion of both uplink (UL) and downlink (DL) procedures, extends previous work done by the authors in [17] and [18].
2) To define a statistical model for the characterization of radio access congestion in B3G scenarios, which is provided in Section III. This model takes into account the congestion that may eventually arise in both the UL and DL, provided a particular distribution of services is known. Such service distribution will be provided by the Markov allocation framework given in Section II.
3) To describe several initial RAT-selection policies responding to different allocation principles within the specified multiservice/multi-RAT framework. In particular, a novel initial RAT-selection policy that takes advantage of congestion information is presented, which can be found in Section IV.
4) For a particular multi-RAT scenario, considering TDMA and WCDMA access networks, along with generic voice and data services, this work presents RAT-specific analytical expressions for computing the CP in the aforementioned RATs, which is detailed in Section V. These congestion-related expressions are formulated, considering well-known (see, e.g., [19] and [20]) TDMA and WCDMA expressions.

II. General Framework for Multiservice Allocation in Multiaccess Networks

The problem of resource allocation in multiaccess systems may be approached by means of multidimensional Markov models where each dimension corresponds to the allocation of a particular service type into a given RAT. In particular, the allocation framework presented in the following is based on previous work done by the authors in [17] and [18], to which interested readers can refer for model validation and further details. The model definition involves, in the first place, the identification of the state space, followed by the definition of the state transition rates, which will eventually lead to the steady-state balance equations provided at the end of this section. Given that the focus of this work is to evaluate and determine the initial RAT-selection policy impact on the radio access congestion, considering that the static case (i.e., no mobility) seems to be adequate for such evaluation while maintaining the model complexity at a tractable level. In addition, it will be assumed that the served users will not suffer from call/session disruptions due to the poor quality or link failures in the selected RAT.

A. Defining the Markov State Space

Let us consider a number of J different traffic classes. A total of K co-sited RATs are deployed, with each one of them supporting either all J traffic classes or a subset of them. To account for RATs that do not uphold particular traffic classes, a $K \times J$ compatibility matrix, which is denoted as $\mathbf{B}$, may be defined with elements $b_{kj} = 1$ if RAT k supports traffic type j and $b_{kj} = 0$ otherwise.

Based on matrix $\mathbf{B}$, the number of supported services J_k by a given RAT k can be computed as $J_k = \sum_{j=1}^{J} b_{kj}$. Therefore, the Markov state dimension M that accounts for the allocation of each supported service into each RAT may be computed as $M = \sum_{k=1}^{K} J_k$.

We may now define the row vector

$$\boldsymbol{N}_k = [N_{k,1}, N_{k,2}, \ldots, N_{k,l}, \ldots, N_{k,J_k}] \in \mathbb{Z}_+^{J_k} \tag{1}$$

with elements $N_{k,l}$ denoting the number of allocated users in RAT k with supported service l. Note that index l, with $l = 1, 2, \ldots, J_k$, corresponds to the lth supported service in RAT k, whereas j is the available service index. For convenience, let us define a *mapping* function g_k that returns the available traffic class index j, given the lth supported traffic class in RAT k, i.e., $j = g_k(l)$. Furthermore, the inverse mapping function is also defined, i.e., $l = g_k^{-1}(j)$, providing the lth supported traffic class in RAT k, given available traffic class index j. It is worth noticing that, if a RAT k supports all service types, then $l = j$.

Taking into account the number of available RATs, the number of users of each supported service in each RAT may be defined as a row vector

$$\boldsymbol{N} = [\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_k, \ldots, \boldsymbol{N}_K] \in \mathbb{Z}_+^{M} \tag{2}$$

where $\boldsymbol{N}$ may be used as the index to uniquely define each state, which is hereinafter denoted as $S_{\boldsymbol{N}}$, in the Markov chain model.

Assuming that the capacity of a particular RAT k, which is defined as the maximum allowable number of users of each service type it may handle, is upper bounded (i.e., *hard capacity* is assumed), a finite number of states $S_{\boldsymbol{N}}$ exists. This limit is usually set by RAT-specific CAC procedures that determine if a new user should be admitted or not in the system so that the minimum QoS requirements of already-accepted users are guaranteed. In terms of the number of states, we may define the set $\mathcal{S}^k$ of feasible states in RAT k as

$$\mathcal{S}^k = \left\{ S_{\boldsymbol{N}} : 0 \leq f_{\boldsymbol{N}_k}^k \leq 1 \right\} \tag{3}$$

where $f_{\boldsymbol{N}_k}^k$ is defined as the *feasibility condition*, which accounts for the CAC procedures in RAT k by defining a given state $S_{\boldsymbol{N}}$ as *feasible* in RAT k, provided $0 \leq f_{\boldsymbol{N}_k}^k \leq 1$.

Finally, a given state $S_{\boldsymbol{N}}$ is said to be feasible if it satisfies $S_{\boldsymbol{N}} \in \mathcal{S}$, with $\mathcal{S} = \bigcap_{k=1}^{K} \mathcal{S}^k$, i.e., if it is feasible in all RATs.

B. Defining State Transitions

Transitions between states $S_{\boldsymbol{N}} \in \mathcal{S}$ in the resulting M-dimensional Markov chain happen due to service arrival rates, i.e., $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_j, \ldots, \lambda_J]$ or due to service departure rates $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_j, \ldots, \mu_J]$. As widely assumed in the literature (see, e.g., [21]–[25]), arrival rates λ_j are Poisson distributed, and service times follow an exponential distribution with mean service time $1/\mu_j$ [26]. Since not all services may be supported by all RATs, we define the supported arrival rates $\boldsymbol{\lambda}_k$ into RAT k as

$$\boldsymbol{\lambda}_k = [\lambda_{k,1}, \lambda_{k,2}, \ldots, \lambda_{k,l}, \ldots, \lambda_{k,J_k}] \in \mathbb{R}_+^{J_k} \tag{4}$$

with $\lambda_{k,l}$ being the arrival rate of the lth supported service type in RAT k. Note that $\boldsymbol{\lambda}_k$ is a subset of $\boldsymbol{\lambda}$ determined by compatibility matrix $\mathbf{B}$. Finally, the supported arrival rates of each traffic class into each RAT is captured by row vector $\boldsymbol{\lambda}_u = [\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_k, \ldots, \boldsymbol{\lambda}_K] \in \mathbb{R}_+^{M}$.

A particular traffic allocation policy, which is referred to as $\pi_{\boldsymbol{N}}$, is then responsible for determining, in each state $S_{\boldsymbol{N}} \in \mathcal{S}$, the specific transition arrival rates of each service type into each of the available RATs $\boldsymbol{\lambda}_\pi$, thus defining the following function:

$$\begin{aligned} \pi_{\boldsymbol{N}} : \quad & \mathbb{R}_+^{J} \longrightarrow \mathbb{R}_+^{M} \\ & \boldsymbol{\lambda} \longrightarrow \boldsymbol{\lambda}_\pi \end{aligned} \tag{5}$$

where vector $\boldsymbol{\lambda}_\pi$ contains elements $\lambda_{(\pi,k,l)}$, which denotes the transition arrival rate of supported service l into RAT k due to policy $\pi_{\boldsymbol{N}}$ in state $S_{\boldsymbol{N}}$.

A specific policy π can be implemented by means of a *policy decision function* $\boldsymbol{\Theta}_{\boldsymbol{N}} \in \mathbb{R}^M$, with elements $\Theta_{(\boldsymbol{N},k,l)} \in [0, 1]$ (called *policy actions*) determining the fraction of supported traffic l into RAT k in a state $S_{\boldsymbol{N}}$, i.e.,

$$\lambda_{(\pi,k,l)} = \Theta_{(\boldsymbol{N},k,l)} \lambda_{k,l}. \tag{6}$$

C. Defining the Steady-State Balance Equations (SSBEs)

In equilibrium, the SSBE for state $S_{\boldsymbol{N}} \in \mathcal{S}$ results from equaling the inflow rate to state $S_{\boldsymbol{N}}$ to the outflow rate of state

$S_{\boldsymbol{N}}$ [26], i.e.,

$$P_{\boldsymbol{N}}\left[\sum_{k=1}^{K}\sum_{l=1}^{J_k}\lambda_{k,l}\Theta_{(\boldsymbol{N},k,l)}\delta_{(\boldsymbol{N}+\boldsymbol{a}_{k,l})}+N_{k,l}\mu_{k,l}\delta_{(\boldsymbol{N}-\boldsymbol{a}_{k,l})}\right]$$
$$=\sum_{k=1}^{K}\sum_{l=1}^{J_k}\lambda_{k,l}\Theta_{(\boldsymbol{N}-\boldsymbol{a}_{k,l},k,l)}P_{(\boldsymbol{N}-\boldsymbol{a}_{k,l})}\delta_{(\boldsymbol{N}-\boldsymbol{a}_{k,l})}+(N_{k,l}+1)\mu_{k,l}P_{(\boldsymbol{N}+\boldsymbol{a}_{k,l})}\delta_{(\boldsymbol{N}+\boldsymbol{a}_{k,l})} \tag{7}$$

where $P_{\boldsymbol{N}}$ is the probability of being in state $S_{\boldsymbol{N}}$, and $\boldsymbol{a}_{k,l}\in\mathbb{Z}_{+}^{M}$ is a row vector containing all zeros, except for the lth supported service in RAT k element, which is 1. In addition, $\delta_{\boldsymbol{N}}$ is an indicator function that will guarantee that nonfeasible states, i.e., $S_{\boldsymbol{N}}\notin\mathcal{S}$, are not taken into account; thus, $\delta_{(\boldsymbol{N})}=1$ if $S_{\boldsymbol{N}}\in\mathcal{S}$ and $\delta_{(\boldsymbol{N})}=0$ otherwise.

Once the SSBEs are determined for all states $S_{\boldsymbol{N}}\in\mathcal{S}$, numerical methods may be applied to solve the resulting system of equations given by the SSBEs plus the normalization constraint $\sum_{S_{\boldsymbol{N}}\in S}P_{\boldsymbol{N}}=1$.

See, e.g., [27] and [28] for further details on the numerical solution of Markov chains.

III. Congestion Probability Model

The Markov model proposed in the previous section provides, given the total traffic arrival rates $\boldsymbol{\lambda}$ and a given RAT-selection policy π, the probability $P_{\boldsymbol{N}}$ of having a particular number of admitted users denoted by vector $\boldsymbol{N}$. Then, the CP in each RAT k for a given RAT-selection policy π may be formulated as follows:

$$P_c^{k,\pi}=\sum_{S_{\boldsymbol{N}}\in\mathcal{S}}P_c^k(\boldsymbol{N}_k)\cdot P_{\boldsymbol{N}} \tag{8}$$

where $P_c^k(\boldsymbol{N}_k)$ is the CP in RAT k, given $\boldsymbol{N}_k$ allocated users, and is averaged over all state probabilities $P_{\boldsymbol{N}}$ in state space $\mathcal{S}$. A detailed explanation on how $P_c^k(\boldsymbol{N}_k)$ is computed will be provided in Section V-B for the case of TDMA and WCDMA technologies. At this point, it should be noted that $P_c^k(\boldsymbol{N}_k)$ highly depends on the underlying access scheme of the considered RAT k, which may lead to different definitions of radio access congestion and, thus, different definitions of CP. In this sense, a TDMA-based scheme can undergo congestion situations due to, e.g., excessive timeslot (TSL) reuse. On the contrary, a RAT based on WCDMA may be congested if, e.g., excessive interference is detected, hence degrading the system performance. Moreover, it can be foreseen that the higher the number of allocated users $\boldsymbol{N}_k$, the higher the CP; nevertheless, users demanding different services may have different impacts on the perceived congestion, as will be shown in Section V-B.

In general, the CP in RAT k $P_c^k(\boldsymbol{N}_k)$, given $\boldsymbol{N}_k$ allocated users, may arise independently due to congestion measured in the UL and/or due to congestion measured in the DL. Consequently, one can express this probability as

$$P_c^k(\boldsymbol{N}_k)=1-\left[1-P_c^{\mathrm{UL},k}(\boldsymbol{N}_k)\right]\left[1-P_c^{\mathrm{DL},k}(\boldsymbol{N}_k)\right] \tag{9}$$

where $P_c^{\mathrm{UL},k}(\boldsymbol{N}_k)$ and $P_c^{\mathrm{DL},k}(\boldsymbol{N}_k)$ indicate UL and DL congestion probabilities, respectively.

Provided the CP in RAT k, given policy π in (8), the overall CP in the multi-RAT scenario, i.e., the probability that there exists congestion in at least one RAT P_c^{π}, can then be written as

$$P_c^{\pi}=1-\prod_{k=1}^{K}\left(1-P_c^{k,\pi}\right) \tag{10}$$

where it is assumed that congestion probabilities arising in each RAT are independent.

IV. Defining RAT-Selection Policies

With previous definitions, a number of RAT-selection policies can be characterized by means of appropriately defining policy decision function $\boldsymbol{\Theta}_{\boldsymbol{N}}$ through its elements $\Theta_{(\boldsymbol{N},k,l)}$. The outcome of a particular RAT-selection policy π for the allocation of a j-service-type user at a given state $S_{\boldsymbol{N}}\in\mathcal{S}$ can be represented by the set $\mathcal{R}_j^{\pi}$ containing the selected RAT(s). Note that the elements in $\mathcal{R}_j^{\pi}$ are those RAT(s) selected by the policy, provided service class j is supported and the allocation of such service is possible, i.e., admission control (AC) is successful. Mathematically, it is required that $\mathcal{R}_j^{\pi}\in\mathcal{A}_j$ with

$$\mathcal{A}_j=\left\{k\in[1,K]:b_{kj}\neq0,S_{(\boldsymbol{N}_k+\boldsymbol{e}_{k,l})}\in\mathcal{S}\right\} \tag{11}$$

where $\boldsymbol{e}_{k,l}\in\mathbb{Z}^{J_k}$ is a vector containing all zeros, except for the lth component, which is 1. Moreover, referring to Section II-A, $b_{kj}\neq0$ indicates that service j is supported by RAT k.

Note that, if $|\mathcal{R}_j^{\pi}|$ denotes the cardinality of $\mathcal{R}_j^{\pi}$ (i.e., the number of elements in $\mathcal{R}_j^{\pi}$), then, in the singular case that set $\mathcal{R}_j^{\pi}$ contains more than one selected RAT, i.e., $|\mathcal{R}_j^{\pi}|>1$, it is assumed that a RAT is chosen among those according to some preestablished prioritization scheme, which is particular to each policy. In addition, it should be noticed that the RAT selection policy definition given by (11) implicitly states that, if the *preferred* RAT according to a given policy cannot admit further service-class-j users, then alternative RATs are sequentially selected according to a prioritized order determined by the policy.

Subsequently, some specific RAT-selection policies are presented and defined in terms of $\mathcal{R}_j^{\pi}$. Such RAT-selection policies are suggested and justified by the authors in [18], to which interested readers are referred for further details.

A. Load-Balancing (LB) RAT Selection Policy

This policy allocates a given user demanding a particular traffic class to the RAT that undergoes a lower load status. Naturally, this implies defining appropriate load metrics for each of the considered RATs. Let the load of a RAT k at a given state $S_{\boldsymbol{N}}\in\mathcal{S}$ be denoted as $L_{\boldsymbol{N}_k}^k$, where it explicitly states that the load value in RAT k will depend, among other RAT-specific parameters, on the number of users of each service class in RAT k, i.e., $\boldsymbol{N}_k$.

At a specified state $S_{\boldsymbol{N}}\in\mathcal{S}$, a service-class-$j$ user will be allocated to the RAT k that, including this new service-class-j

user, exhibits a lower load. Accordingly, we may define

$$\mathcal{R}_j^{\mathrm{LB}} = \arg\min_{k \in \mathcal{A}_j} \left\{ L^k_{(\boldsymbol{N}_k + \boldsymbol{e}_{k,l})} \right\} \tag{12}$$

where the elements in $\mathcal{R}_j^{\mathrm{LB}}$ are those RAT(s) with the lowest load, provided service class j is supported and the allocation of supported service class l is possible.

Then, the LB policy decision toward supported traffic l into RAT k in state $S_{\boldsymbol{N}}$ can be defined in terms of $\mathcal{R}_j^{\mathrm{LB}}$ as

$$\Theta_{(\boldsymbol{N},k,l)} = \begin{cases} \frac{1}{|\mathcal{R}_j^{\mathrm{LB}}|}, & \text{if } k \in \mathcal{R}_j^{\mathrm{LB}} \text{ with } j = g_k(l) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where it is implicitly assumed that, in the singular case of $|\mathcal{R}_j^{\mathrm{LB}}| > 1$, i.e., the minimum load is achieved in more than one RAT, selection is randomly done among those RATs, with probability $1/|\mathcal{R}_j^{\mathrm{LB}}|$.

B. Service-Based (SB) RAT-Selection Policy

This policy selects a particular RAT k based on the demanding user service type j according to some predefined preference scheme. Consider that, for each available service type j, we may build a row vector $\boldsymbol{s}_j = [s_{(1,j)}, s_{(2,j)}, \ldots, s_{(k,j)}, \ldots, s_{(K,j)}] \in \mathbb{Z}^K$. Each element $s_{(k,j)}$ in this vector takes a nonnegative integer value indicating the preference of service j on RAT k. The higher the value of $s_{(k,j)}$, the higher the priority.

At a specified state $S_{\boldsymbol{N}} \in \mathcal{S}$, a service-class-$j$ user will be allocated to RAT k according to

$$\mathcal{R}_j^{\mathrm{SB}} = \arg\max_{k \in \mathcal{A}_j} \left\{ s_{(k,j)} \right\} \tag{14}$$

where $\mathcal{R}_j^{\mathrm{SB}}$ contains the RAT with the highest priority value, provided it is capable of supporting and admitting the new service-class-j user. It is worthwhile mentioning that, for this particular policy, we will always have $|\mathcal{R}_j^{\mathrm{SB}}| = 1$, i.e., the policy outcome is always one RAT.

In this case, the SB policy decision with respect to the supported traffic l into RAT k at a given state $S_{\boldsymbol{N}}$ can be defined in terms of $\mathcal{R}_j^{\mathrm{SB}}$ as

$$\Theta_{(\boldsymbol{N},k,l)} = \begin{cases} 1, & \text{if } k \in \mathcal{R}_j^{\mathrm{SB}} \text{ with } j = g_k(l) \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

C. Congestion-Aware (CA) RAT-Selection Policy

The statistical model for CP presented in Section III can be used as a valuable input for deciding the most appropriate RAT for a specific service request. This way, a particular user may be allocated to the RAT that will less likely fall in a high congestion state. In this case, the set of selected RAT(s) is

$$\mathcal{R}_j^{\mathrm{CA}} = \arg\min_{k \in \mathcal{A}_j} \left\{ P_c^k(\boldsymbol{N}_k + \boldsymbol{e}_{k,l}) \right\} \tag{16}$$

where the RAT(s) with the lowest CP is selected, provided service class j is supported and the allocation of supported service class l is possible.

Similar to the LB case in (13), the CA policy decision with respect to supported traffic l into RAT k at a given state $S_{\boldsymbol{N}}$ can be defined in terms of $\mathcal{R}_j^{\mathrm{CA}}$ as

$$\Theta_{(\boldsymbol{N},k,l)} = \begin{cases} \frac{1}{|\mathcal{R}_j^{\mathrm{CA}}|}, & \text{if } k \in \mathcal{R}_j^{\mathrm{CA}} \text{ with } j = g_k(l) \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

where it is also assumed that, in the particular case of $|\mathcal{R}_j^{\mathrm{CA}}| > 1$, i.e., the minimum CP is achieved in more than one RAT, the RAT is randomly selected among those with probability $1/|\mathcal{R}_j^{\mathrm{CA}}|$.

According to the aforementioned definitions, the policy decisions in the SSBEs will be performed using (13), (15), and (17) into (7) to account for the different user-allocation strategies.

V. Case Study: Voice and Data Services in a TDMA/WCDMA Scenario

In this section, the state space and the congestion probabilities assuming two generic service types, i.e., voice and data, and two RATs, i.e., TDMA and WCDMA, are derived. Both RATs are considered to support both services in a single-cell scenario; thus, a 4-D ($M = 4$) Markov model arises. In the remainder, let v and d represent voice and data indexes, along with t and w, which represent TDMA and WCDMA indexes.

A. Markov State Space

A boundary on the number of states comprised by the Markov chain is set by imposing appropriate CAC mechanisms in each RAT.

1) TDMA Case: For TDMA-based systems, a total number of C channels (or TSLs) are to be shared among voice and data users within a time frame. It is further assumed that no service has priority over any other. Typically, but not necessarily, voice users occupy a whole TSL throughout the duration of a call, which will be assumed in this work. As for data users, a given TSL may be shared by up to n_C data transmissions corresponding to different users by efficient time scheduling. Moreover, a given data user may be granted several TSLs to increase its achievable bit rate. This feature is commonly referred to as multislot capability. To capture this multislot capability, it is assumed that the number of allocated TSLs to a particular data user is q.

The CAC mechanism may then be expressed in mathematical terms, considering the total amount of resources, i.e., C TSLs, to be shared among the simultaneous voice and data users in the system. This will define the set of feasible number of admitted voice and data users in the UL and DL, i.e., $\mathcal{S}^{\mathrm{UL},t}$ and $\mathcal{S}^{\mathrm{DL},t}$, respectively, as

$$\mathcal{S}^{\mathrm{UL},t} = \left\{ \boldsymbol{N}_t : 0 \leq \frac{N_{t,v}}{C} + \frac{N_{t,d}\alpha_d^{\mathrm{UL}} q^{\mathrm{UL}}}{n_C^{\mathrm{UL}} C} \leq 1 \right\}$$
$$\mathcal{S}^{\mathrm{DL},t} = \left\{ \boldsymbol{N}_t : 0 \leq \frac{N_{t,v}}{C} + \frac{N_{t,d}\alpha_d^{\mathrm{DL}} q^{\mathrm{DL}}}{n_C^{\mathrm{DL}} C} \leq 1 \right\} \tag{18}$$

where $N_{t,v}$ ($N_{t,d}$) is the number of admitted voice (data) users in TDMA, and α_d^{UL} (α_d^{DL}) is the UL (DL) activity factor of data

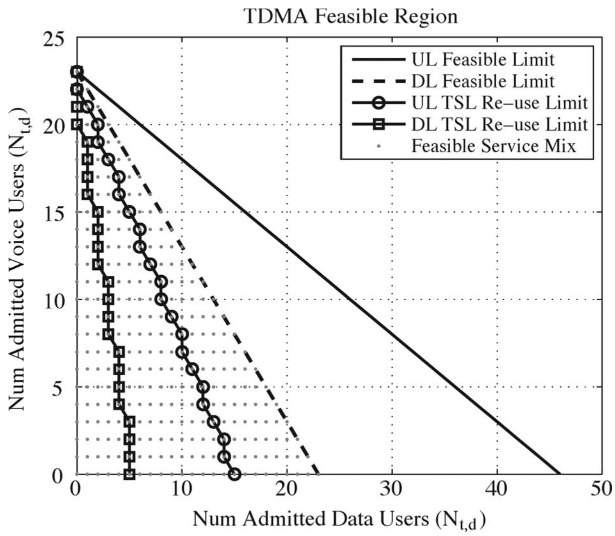

Fig. 1. TDMA feasible region for $C = 23$, $n_C^{\mathrm{UL}} = 3$, $n_C^{\mathrm{DL}} = 4$, $q^{\mathrm{UL}} = 3$, $q^{\mathrm{DL}} = 8$, and $\alpha_d^{\mathrm{UL}} = \alpha_d^{\mathrm{DL}} = 0.5$.

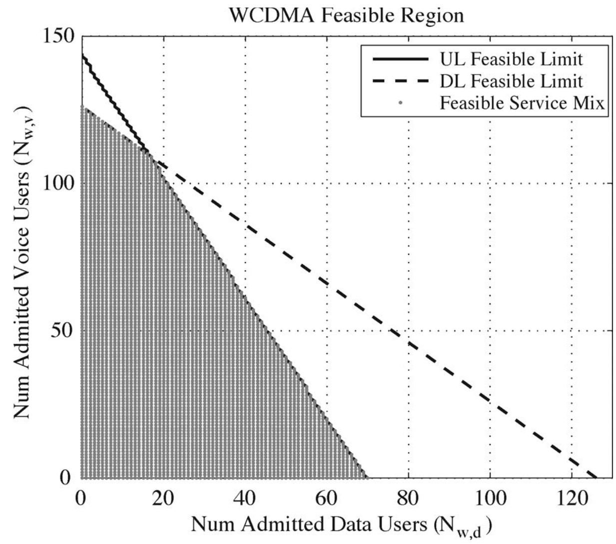

Fig. 2. WCDMA feasible region for $W = 3.84$ MHz, $R_{bv,i}^{\mathrm{UL}} = 12.2$ kb/s, $R_{bd,i}^{\mathrm{UL}} = 32$ kb/s, $\theta_{v,i}^{\mathrm{UL}} = 6$ dB, $\theta_{d,i}^{\mathrm{UL}} = 5$ dB, $\alpha_v^{\mathrm{UL}} = \alpha_d^{\mathrm{UL}} = \alpha_d^{\mathrm{DL}} = 0.5$, $\eta_{\max}^{\mathrm{UL}} = 0.9$, $\mathcal{C}_{\max} = 63/64$, $\mathrm{SF}_{v,i}^{\mathrm{DL}} = 128$, and $\mathrm{SF}_{d,i}^{\mathrm{DL}} = 64$.

users. In addition, the number of allocated TSLs to a particular data user in the UL (DL) is represented by q^{UL} (q^{DL}), and the maximum number of users sharing the same TSL in the UL (DL) is denoted by n_C^{UL} (n_C^{DL}). The set of feasible number of users satisfying both conditions in (18), i.e., $\mathcal{S}^t = \mathcal{S}^{\mathrm{UL},t} \cap \mathcal{S}^{\mathrm{DL},t}$, constitutes the so-called TDMA admission region. In Fig. 1, the admission limits imposed by the UL and DL are shown, along with the limits above which TSL reuse exists in the UL and DL directions.

2) WCDMA Case: As for WCDMA-based systems, CAC procedures also comprise both UL and DL mechanisms. In WCDMA, the *uplink load factor* η^{UL} can be used to determine if a new user should be admitted in the system by setting a maximum value $\eta_{\max}^{\mathrm{UL}}$, which cannot be exceeded. Similar to what was shown for TDMA, an UL WCDMA feasibility set can be defined as

$$\mathcal{S}^{\mathrm{UL},w} = \left\{ \boldsymbol{N}_w : 0 \leq \eta_{\boldsymbol{N}_w}^{\mathrm{UL}} \leq \eta_{\max}^{\mathrm{UL}} \right\} \tag{19}$$

where the UL load factor in WCDMA is defined as [20]

$$\eta_{\boldsymbol{N}_w}^{\mathrm{UL}} = \sum_{i=1}^{N_{w,v}} \alpha_v^{\mathrm{UL}} / A_{v,i}^{\mathrm{UL}} + \sum_{i=1}^{N_{w,d}} \alpha_d^{\mathrm{UL}} / A_{d,i}^{\mathrm{UL}} \tag{20}$$

with

$$A_{v,i}^{\mathrm{UL}} = \frac{W/R_{bv,i}^{\mathrm{UL}}}{\theta_{v,i}^{\mathrm{UL}}} + 1 \quad A_{d,i}^{\mathrm{UL}} = \frac{W/R_{bd,i}^{\mathrm{UL}}}{\theta_{d,i}^{\mathrm{UL}}} + 1 \tag{21}$$

where W is the chip rate, $R_{bv,i}^{\mathrm{UL}}$ ($R_{bd,i}^{\mathrm{UL}}$) is the UL bit rate granted to voice (data) users, and $\theta_{v,i}^{\mathrm{UL}}$ ($\theta_{d,i}^{\mathrm{UL}}$) is the UL target bit-energy-to-noise-density ratio after despreading and decoding for voice (data) users. Additionally, α_v^{UL} is the UL activity factor for voice users.

In the DL, a CAC algorithm based on the availability of orthogonal variable spreading factor (OVSF) codes is assumed as, e.g., in [20]. Then, assuming that the same OVSF code can be shared among different data users due to inactivity periods, the number of dedicated channels will have to fulfill the relationship [20]

$$\mathcal{C}_{\boldsymbol{N}_w} = \sum_{i=1}^{N_{w,v}} \frac{1}{SF_{v,i}^{\mathrm{DL}}} + \sum_{i=1}^{N_{w,d}} \alpha_d^{\mathrm{DL}} \frac{1}{SF_{d,i}^{\mathrm{DL}}} \leq \mathcal{C}_{\max} \tag{22}$$

with $\mathrm{SF}_{v,i}^{\mathrm{DL}}$ ($\mathrm{SF}_{d,i}^{\mathrm{DL}}$) being the DL spreading factor for voice (data) users, and $0 \leq \mathcal{C}_{\max} \leq 1$. Then, the feasible set of voice and data users in WCDMA DL is expressed as

$$\mathcal{S}^{\mathrm{DL},w} = \{ \boldsymbol{N}_w : 0 \leq \mathcal{C}_{\boldsymbol{N}_w} \leq \mathcal{C}_{\max} \}. \tag{23}$$

The resulting WCDMA total feasible set $\mathcal{S}^w$ must satisfy both the UL and DL feasibility conditions defined in (19) and (23), i.e., $\mathcal{S}^w = \mathcal{S}^{\mathrm{UL},w} \cap \mathcal{S}^{\mathrm{DL},w}$, which results in the admission region shown by the dotted area in Fig. 2.

Finally, the global Markov state feasibility space $\mathcal{S}$ can be defined from the feasible sets in each RAT as

$$\mathcal{S} = \{ \boldsymbol{N} : \boldsymbol{N}_t \in \mathcal{S}^t, \boldsymbol{N}_w \in \mathcal{S}^w \}. \tag{24}$$

B. CP Cases

In the following, the CP is derived for RAT WCDMA and TDMA:

1) WCDMA CP: In WCDMA, radio congestion events may arise in the UL and/or DL, thus enabling the CP in WCDMA to be formulated as

$$P_c^w(\boldsymbol{N}_w) = 1 - \left[1 - P_c^{\mathrm{UL},w}(\boldsymbol{N}_w)\right] \left[1 - P_c^{\mathrm{DL},w}(\boldsymbol{N}_w)\right] \tag{25}$$

where the assumption that the CP in the UL and DL is independent has been made.

UL case: It is well known (see, e.g., [20]) that user traffic activity, which is characterized by activity factor α, may be modeled by a binomial distribution to determine the probability of having n simultaneously transmitting users, given N admitted users, i.e.,

$$P_\alpha(n|N) = \binom{N}{n}\alpha^n(1-\alpha)^{N-n}. \quad (26)$$

Then, assuming independence between voice and data call/session generation, the probability of $n_{w,v}$ voice users and $n_{w,d}$ data users simultaneously transmitting in WCDMA when $N_{w,v}$ and $N_{w,d}$ voice and data users are admitted in the system $P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w)$ can be formulated as

$$P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w) = P_{\alpha_v^{\mathrm{UL}}}(n_{w,v}|N_{w,v}) \cdot P_{\alpha_d^{\mathrm{UL}}}(n_{w,d}|N_{w,d}) \quad (27)$$

where vector $\boldsymbol{\alpha}^{\mathrm{UL}} = (\alpha_v^{\mathrm{UL}}, \alpha_d^{\mathrm{UL}})$ denotes UL voice and data activity factors, i.e.,

$$\Pr\left(P_T^{\mathrm{DL}}(\boldsymbol{n_w}) > P_{T,c}^{\mathrm{DL}}\right) = \Pr\left\{\left(\sum_{i=1}^{n_{w,v}} \frac{L_{p,i}}{A_{v,i}^{\mathrm{DL}}+\rho} + \sum_{i=1}^{n_{w,d}} \frac{L_{p,i}}{A_{d,i}^{\mathrm{DL}}+\rho}\right) > \frac{P_{T,c}^{\mathrm{DL}}\left(1-\eta_{(\boldsymbol{n}_w)}^{\mathrm{DL}}\right)-P_p}{P_N}\right\} \triangleq \Pr\left(\gamma_{\boldsymbol{n_w}} > \gamma_{\boldsymbol{n_w}}^*\right). \quad (28)$$

In WCDMA, congestion situations may be detected in the UL by means of UL load factor η^{UL} whenever its value exceeds a given threshold η_c^{UL} during a certain percentage of frames within a period of time [7]. Thus, we may define the WCDMA UL CP $P_c^{\mathrm{UL},w}(\boldsymbol{N}_w)$ as

$$P_c^{\mathrm{UL},w}(\boldsymbol{N}_w) = \sum_{n_{w,v}=0}^{N_{w,v}} \sum_{n_{w,d}=0}^{N_{w,d}} \Pr\left(\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}} > \eta_c^{\mathrm{UL}}\right) \cdot P_{\boldsymbol{\alpha}}^{\mathrm{UL}}(\boldsymbol{n}_w|\boldsymbol{N}_w) \quad (29)$$

where we can express the UL load factor in a single-cell scenario when $n_{w,v}$ simultaneous voice users and $n_{w,d}$ simultaneous data users are in the system, i.e., $\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}}$, as [20]

$$\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}} = \sum_{i=1}^{n_{w,v}} 1/A_{v,i}^{\mathrm{UL}} + \sum_{i=1}^{n_{w,d}} 1/A_{d,i}^{\mathrm{UL}} \quad (30)$$

with $A_{v,i}^{\mathrm{UL}}$ and $A_{d,i}^{\mathrm{UL}}$ defined in (21).

Consequently and assuming users of the same service type to have equal allocated bit rates and bit-energy-to-noise-density ratio requirements, i.e., $A_{v,i}^{\mathrm{UL}} = A_v^{\mathrm{UL}}$ and $A_{d,i}^{\mathrm{UL}} = A_d^{\mathrm{UL}} \forall i$, it may be written as

$$\Pr\left(\eta_{(\boldsymbol{n}_w)}^{\mathrm{UL}} > \eta_c^{\mathrm{UL}}\right) = \begin{cases} 0, & \text{if } \boldsymbol{n}_w \cdot (\boldsymbol{A}^{\mathrm{UL}})^T \le \eta_c \\ 1, & \text{if } \boldsymbol{n}_w \cdot (\boldsymbol{A}^{\mathrm{UL}})^T > \eta_c \end{cases} \quad (31)$$

where it has been defined that $\boldsymbol{A}^{\mathrm{UL}} = (A_v^{\mathrm{UL}^{-1}}, A_d^{\mathrm{UL}^{-1}})$.

Given the WCDMA system parameters provided in Table I, the CP in the UL with respect to the number of admitted voice and data users defined in (29) is shown in Fig. 3, where congestion probabilities are represented in grayscale form, with the color bar indicating the probability values on the right of the graph.

TABLE I
WCDMA SYSTEM PARAMETERS FOR NUMERICAL EVALUATION OF UL CP

Parameter	Symbol	Value
Chip rate	W	3.84 Mcps
Load factor congestion threshold	η_c^{UL}	0.7
Maximum load factor	η_{max}^{UL}	0.9
E_b/N_0[a] for voice (data) traffic	θ_v^{UL} (θ_d^{UL})	6 (5) dB
Bit-rate for voice (data) traffic	$R_{b,v}^{\mathrm{UL}}$ ($R_{b,d}^{\mathrm{UL}}$)	12.2 (32) kbps
Voice (data) activity factor	α_v^{UL} (α_d^{UL})	0.5 (0.5)

[a] Bit energy to noise density ratio.

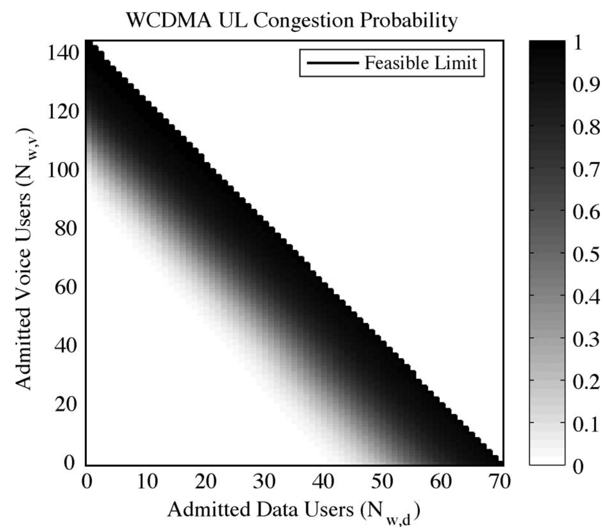

Fig. 3. WCDMA UL CP.

DL case: In WCDMA systems, the total DL transmitted power needed to satisfy all $n_{w,v}$ and $n_{w,d}$ simultaneous voice and data users in a single-cell scenario is [20]

$$P_T^{\mathrm{DL}}(\boldsymbol{n_w}) = \frac{P_p + P_N\left(\sum_{i=1}^{n_{w,v}} \frac{L_{p,i}}{A_{v,i}^{\mathrm{DL}}+\rho} + \sum_{i=1}^{n_{w,d}} \frac{L_{p,i}}{A_{d,i}^{\mathrm{DL}}+\rho}\right)}{1-\eta_{(\boldsymbol{n}_w)}^{\mathrm{DL}}} \le P_{T,\max}^{\mathrm{DL}} \quad (32)$$

with

$$A_{v,i}^{\mathrm{DL}} = \frac{W/R_{bv,i}^{\mathrm{DL}}}{\theta_{v,i}^{\mathrm{DL}}} \quad A_{d,i}^{\mathrm{DL}} = \frac{W/R_{bd,i}^{\mathrm{DL}}}{\theta_{d,i}^{\mathrm{DL}}} \quad (33)$$

along with

$$\eta_{(\boldsymbol{n}_w)}^{\mathrm{DL}} = \sum_{i=1}^{n_{w,v}} \frac{\rho}{A_{v,i}^{\mathrm{DL}}+\rho} + \sum_{i=1}^{n_{w,d}} \frac{\rho}{A_{d,i}^{\mathrm{DL}}+\rho} \quad (34)$$

where P_p and P_N are the pilot and thermal noise powers, respectively; ρ is the DL orthogonality factor; and $P_{T,\max}^{\mathrm{DL}}$ is the maximum total DL power available at the base station. W is the WCDMA chip rate, and the user requirements are in the form of requested bit rates for voice and data services $R_{bv,i}^{\mathrm{DL}}$ and $R_{bd,i}^{\mathrm{DL}}$, along with target bit-energy-to-noise-density ratios for voice and data services $\theta_{v,i}^{\mathrm{DL}}$ and $\theta_{d,i}^{\mathrm{DL}}$. The path loss experienced by each user $L_{p,i}$ may be characterized by considering macro- or microcell environments with shadowing effects modeled by means of lognormal variation [20].

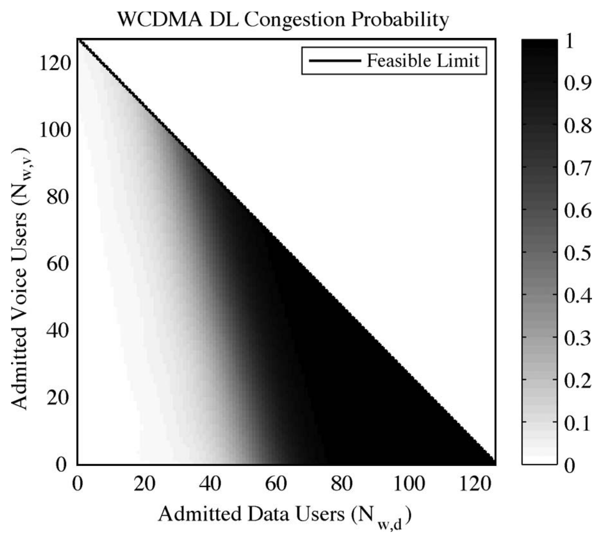

Fig. 4. WCDMA DL CP.

Consequently, for $n_{w,v}$ and $n_{w,d}$ simultaneous users, a congestion situation may be detected whenever the total DL power exceeds a given power threshold $P_{T,c}^{\mathrm{DL}} < P_{T,\max}^{\mathrm{DL}}$. Then, we may write the CP in the DL as shown in (28), which depends on the path-loss distribution and, consequently, on the user's geographical location.

For a given number of admitted voice and data users, the DL CP due to power constraints can be written as

$$P_c^{\mathrm{DL},w}(\boldsymbol{N}_w) = \sum_{n_{w,v}=0}^{N_{w,v}} \sum_{n_{w,d}=0}^{N_{w,d}} \Pr\left(\gamma_{\boldsymbol{n}_w} > \gamma_{\boldsymbol{n}_w}^*\right) \cdot P_{\boldsymbol{\alpha}}^{\mathrm{DL}}(\boldsymbol{n}_w | \boldsymbol{N}_w). \tag{35}$$

The CP in the DL as a function of the number of active voice and data users defined in (35) is shown in Fig. 4, where the terms $\Pr(\gamma_{\boldsymbol{n}_w} > \gamma_{\boldsymbol{n}_w}^*)$ and $P_{\boldsymbol{\alpha}^{\mathrm{DL}}}(\boldsymbol{n}_w|\boldsymbol{N}_w)$ are computed, given the WCDMA system parameters provided in Table II. It can be seen that the strain in the DL CP is set on the number of data users in the system due to their higher demands in terms of bit rates, which is translated into higher power demands and, consequently, higher congestion chances.

The total CP in WCDMA, considering the UL and DL, i.e., the probability of being in congestion in either the UL or the DL, can then be computed according to (25) and represented as shown in Fig. 5.

2) TDMA CP: As suggested in [7] and [8], a possible indicator of congestion in TDMA access technologies can be based on the effect of TSL sharing among data users. Accordingly, the CP in TDMA, given $n_{t,v}$ simultaneous voice users are assigned τ_v TSLs and $n_{t,d}$ simultaneous data users are assigned τ_d data TSL, can be expressed as

$$\Pr\left(\xi_{(\tau_v,\tau_d)} < \xi_c\right) = \begin{cases} 1, & \text{if } \xi_{(\tau_v,\tau_d)} < \xi_c \\ 0, & \text{if } \xi_{(\tau_v,\tau_d)} \geq \xi_c \end{cases} \tag{36}$$

where, if C is the total number of available TSLs

$$\xi_{(\tau_v,\tau_d)} = \begin{cases} 1, & \text{if } 0 \leq \tau_d \leq C - \tau_v \\ C - \frac{\tau_v}{\tau_d}, & \text{if } \tau_d > C - \tau_v \end{cases} \tag{37}$$

TABLE II
WCDMA SYSTEM PARAMETERS FOR NUMERICAL EVALUATION OF DL CP

Parameter	Symbol	Value
Cell radius	R	1000 m
User spatial distr.	-	Homogeneous
Propagation model	-	Macrocell
Log-normal shadowing std. deviation	σ	10 dB
Min. coupling loss	MCL	70 dB
Carrier frequency	f_c	1800 Mhz
Pilot power	P_p	30 dBm
Thermal noise power	P_N	-100 dBm
Orthogonality factor	ρ	0.6
Chip rate	W	3.84 Mcps
DL power congestion threshold	$P_{T,c}^{\mathrm{DL}}$	35 dBm
Max. DL transmitted power	$P_{T,max}^{\mathrm{DL}}$	43 dBm
E_b/N_0 [a] for voice (data) traffic	θ_v^{DL} (θ_d^{DL})	6 (7) dB
Bit-rate for voice (data) traffic	$R_{b,v}^{\mathrm{DL}}$ ($R_{b,d}^{\mathrm{DL}}$)	12.2 (64) kbps
Voice (data) spreading factor	SF_v^{DL} (SF_d^{DL})	128 (64)
Max. OVSF Code Condition Limit	$\mathcal{C}_{max}$	63/64 [b]
Voice (data) activity factor	α_v^{DL} (α_d^{DL})	0.5 (0.5)

[a] Bit energy to noise density ratio.
[b] For the case of a UTRAN R99 WCDMA cell with one code for the CPICH channel, another for the Primary CCPCH carrying the BCH and a couple of Secondary CCPCH channels carrying the FACH and the PCH, all of them with $SF = 256$, the fraction reserved codes for the common and shared channels would be $R_C = 4 \cdot (1/256) = 1/64$ thus yielding $\mathcal{C}_{max} = 1 - R_C = 63/64$ [20].

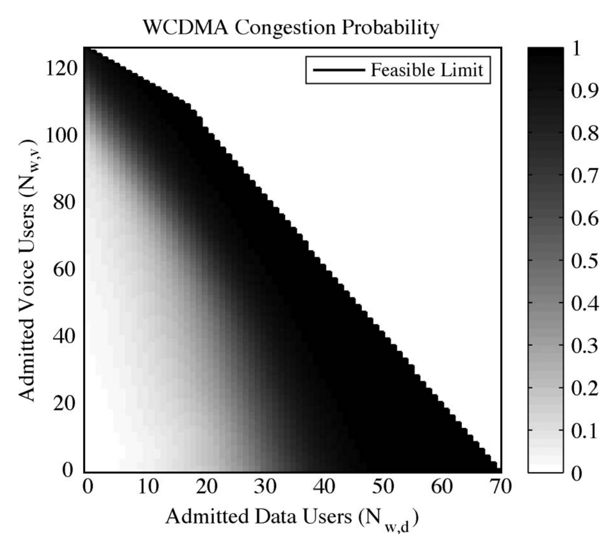

Fig. 5. WCDMA CP.

is the *reduction factor* [19] that accounts for the effect of TSL sharing among data users in a TDMA system, e.g., GSM/EDGE. It follows from (37) that ξ takes values between 0 and 1, meaning a very saturated network for ξ close to 0 (high TSL sharing) and a low loaded network for ξ close to 1 (low TSL sharing). According to (36), congestion is detected if reduction factor ξ falls below a given threshold ξ_c.

Then, given a number of $N_{t,v}$ and $N_{t,d}$ admitted users in TDMA, we may compute the probability of $n_{t,v}$ voice users and $n_{t,d}$ data users simultaneously transmitting in TDMA in the same way as (27). The resulting simultaneous users will be

TABLE III
TDMA System Parameters for Numerical Evaluation of CP

Parameter	Symbol	Value
Total number of available channels	C	23
Max. number of data users per TSL in UL (DL)	n_C^{UL} (n_C^{DL})	3 (4)
Reduction factor congestion threshold	$\xi_c^{\mathrm{UL}} = \xi_c^{\mathrm{DL}}$	0.35
Max. number of TSL per data user in UL (DL)	q^{UL} (q^{DL})	3 (8)
Voice activity factor	$\alpha_v^{\mathrm{UL}} = \alpha_v^{\mathrm{DL}}$	1
Data activity factor	$\alpha_d^{\mathrm{UL}} = \alpha_d^{\mathrm{DL}}$	0.5

assigned a number of TSLs in both the UL and DL, as given by q^{UL} and q^{DL}, respectively, allowing the following expressions for CP in each link to be defined:

$$P_c^{\mathrm{UL},t}(\boldsymbol{N}_t) = \sum_{n_{t,v}=0}^{N_{t,v}} \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{UL}}_{(n_{t,v}, n_{t,d}q^{\mathrm{UL}})} < \xi_c^{\mathrm{UL}}\right) \cdot P_{\boldsymbol{\alpha}^{\mathrm{UL}}}(\boldsymbol{n}_t|\boldsymbol{N}_t) \quad (38)$$

along with

$$P_c^{\mathrm{DL},t}(\boldsymbol{N}_t) = \sum_{n_{t,v}=0}^{N_{t,v}} \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi_{(n_{t,v}, n_{t,d}q^{\mathrm{DL}})}{}^{\mathrm{DL}} < \xi_c^{\mathrm{DL}}\right) \cdot P_{\boldsymbol{\alpha}^{\mathrm{DL}}}(\boldsymbol{n}_t|\boldsymbol{N}_t) \quad (39)$$

where it has been assumed that simultaneous voice users always occupy a single TSL, i.e., $\tau_v = n_{t,v}$, and simultaneous data users are assigned $\tau_d = n_{t,d}q$ TSLs. Finally, it may further be considered that voice users will occupy the entire TSL throughout the duration of the call, which basically means that the activity factor for voice users α_v is equal to 1, thus yielding $n_{t,v} = N_{t,v}$. In this case, (38) and (39) become

$$P_c^{\mathrm{UL},t}(\boldsymbol{N}_t) = \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{UL}}_{(N_{t,v}, n_{t,d}q^{\mathrm{UL}})} < \xi_c^{\mathrm{UL}}\right) \cdot P_{\alpha_d^{\mathrm{UL}}}(n_{t,d}|N_{t,d}) \quad (40)$$

along with

$$P_c^{\mathrm{DL},t}(\boldsymbol{N}_t) = \sum_{n_{t,d}=0}^{N_{t,d}} \Pr\left(\xi^{\mathrm{DL}}_{(N_{t,v}, n_{t,d}q^{\mathrm{DL}})} < \xi_c^{\mathrm{DL}}\right) \cdot P_{\alpha_d^{\mathrm{DL}}}(n_{t,d}|N_{t,d}). \quad (41)$$

Finally, the total CP in TDMA due to UL and DL contributions yields

$$P_c^t(\boldsymbol{N}_t) = 1 - \left[1 - P_c^{\mathrm{UL},t}(\boldsymbol{N}_t)\right]\left[1 - P_c^{\mathrm{DL},t}(\boldsymbol{N}_t)\right] \quad (42)$$

where the assumption that the congestion probabilities in UL and DL are independent is made.

For illustrative purposes, considering the TDMA system parameters provided in Table III, the total TDMA CP, as defined in (42), is given in Fig. 6. It should be observed that, with the chosen parameters in Table III, data users are granted with a higher amount of TSL in the DL than the UL (eight, as opposed to three). This asymmetry in channel allocation causes the CP to be higher in the DL case, where TSL reuse is potentially higher, than in the UL case. As a consequence, the total CP in TDMA shown in Fig. 6 will mainly be DL driven.

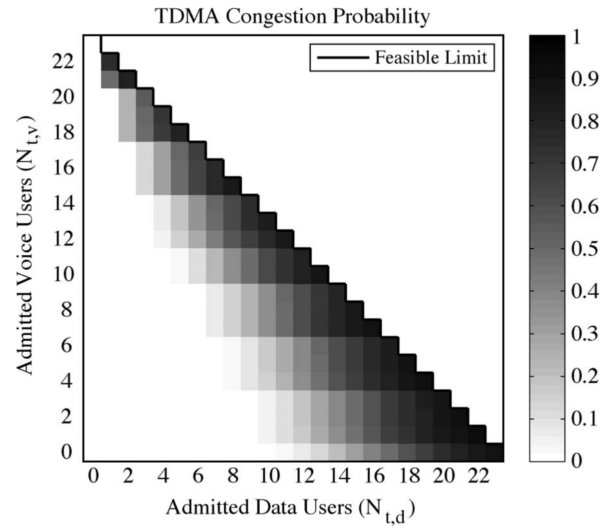

Fig. 6. TDMA CP.

VI. Model Evaluation

In this section, numerical evaluation of the proposed model is carried out. See [17] and [18] for details concerning the validation of the Markov allocation model. Numerical results for the model evaluation of congestion probabilities are computed in two steps: First, the evaluation of the allocation model (presented in Section II) determines the probability $P_{\boldsymbol{N}}$ of having a given number of users in each RAT according to a particular RAT selection policy π. Once the statistical distribution of users is known, i.e., $P_{\boldsymbol{N}}$, the congestion probabilities in each RAT given by (8) and the overall CP given by (10) can be computed. In addition, the $P_{\boldsymbol{N}}$ obtained from the allocation model will be used to compute the performance metrics, as defined later in Section VI-B.

Several RAT selection policies are evaluated in the following to assess the performance of the proposed model, considering voice and data services to be allocated in TDMA and WCDMA RATs. In particular, the LB and CA policies introduced in Sections IV-A and C, respectively, are evaluated. In addition, an illustrative example of an SB policy, which was introduced in Section IV-B, is considered, aiming to prioritize voice users to be allocated to TDMA and data users to be allocated to WCDMA. It is worth mentioning that the performance evaluation in this work is carried out for a particular illustrative case with specific parameter settings for the allocation and congestion models.

In this paper and without loss of generality, it is assumed that the load metrics used in the LB policy operation (see Section IV-A) for TDMA and WCDMA systems are taken in the UL direction as in [17]. Nevertheless, other configurations are possible in the described model.

According to the preceding discussion, in TDMA-based systems, such as, e.g., GSM/EDGE, the *TSL utilization factor* [19] can be used to measure the load in a given state $S_{\boldsymbol{N}} \in \mathcal{S}$.

It is defined as the ratio of the number of occupied TSLs to the number of available TSLs and is thus expressed as

$$L^t_{\boldsymbol{N}_t} = \min\left(C, N_{t,v} + N_{t,d}\alpha_d^{\mathrm{UL}} q^{\mathrm{UL}}\right)/C. \quad (43)$$

On the other hand, the load in a WCDMA-based system may be calculated by means of the UL load factor previously defined in (20); thus

$$L^w_{\boldsymbol{N}_w} = \eta^{\mathrm{UL}}_{\boldsymbol{N}_w}. \quad (44)$$

A. Parameter Settings

The parameter settings for TDMA/WCDMA are those given in Tables I–III. In addition, it is considered that the transmission bit rate of a single voice TSL is $\kappa_v = 12.2$ kb/s and that of a single data TSL is $\kappa_d = 29.6$ kb/s. Numerical results are obtained for a fixed total offered voice traffic of $T_v = \lambda_v/\mu_v = 10$ Erlangs (E) and a total offered data traffic $T_d = \lambda_d/\mu_d$ varying between 0 and 100 E.

B. Performance Metrics

Once the probabilities $P_{\boldsymbol{N}}$ for each state $S_{\boldsymbol{N}} \in \mathcal{S}$ are computed, a number of relevant performance metrics can be defined. In particular, as mentioned in Section III, it is of special interest to measure the CP in each RAT for a specific RAT selection policy as in (8), along with the overall CP, as in (10). Other relevant performance metrics are detailed in the succeeding sections.

1) Blocking Probability (BP): A given state $S_{\boldsymbol{N}} \in \mathcal{S}$ is said to be a *blocking state* if the addition of any service-type user into any RAT forces the system to move to a nonfeasible state $S'_{\boldsymbol{N}}$, meaning $S'_{\boldsymbol{N}} \notin \mathcal{S}$. Let the set of all blocking states be represented by $\mathcal{S}_b$. Then, the total BP yields

$$P_b = \sum_{S_{\boldsymbol{N}} \in \mathcal{S}_b} P_{\boldsymbol{N}}. \quad (45)$$

2) Throughput: Under congestion situations, users may undergo different levels of QoS degradation over a range of key performance indicators (e.g., delay, throughput, etc.). It has been shown that, in TDMA, congestion arises due to excessive TSL sharing, which, in turn, causes degradation of the perceived throughput in data users. This degradation is quantified by means of factor ξ introduced in (37), being the throughput per data user in TDMA $q\kappa_d\xi$. In WCDMA and for the sake of comparison, it is assumed that the throughput per user in WCDMA is also degraded due to congestion in such a way that the perceived throughput for voice and data users is $R_{b,v} \cdot [1 - P_c^w(\boldsymbol{N}_w)]$ and $R_{b,d} \cdot [1 - P_c^w(\boldsymbol{N}_w)]$, respectively, thus reflecting that congestion in WCDMA turns into an excess of interference so that power control cannot ensure the target requirements. Then, the overall throughput per data user in a given state $S_{\boldsymbol{N}} \in \mathcal{S}$ can be defined as

$$\Gamma_{d,u}(\boldsymbol{N}) = \frac{N_{t,d}q\kappa_d\xi_{(N_{t,v},N_{t,d}q\alpha_d)} + N_{w,d}R_{b,d}\left[1 - P_c^w(\boldsymbol{N}_w)\right]}{N_{t,d} + N_{w,d}} \quad (46)$$

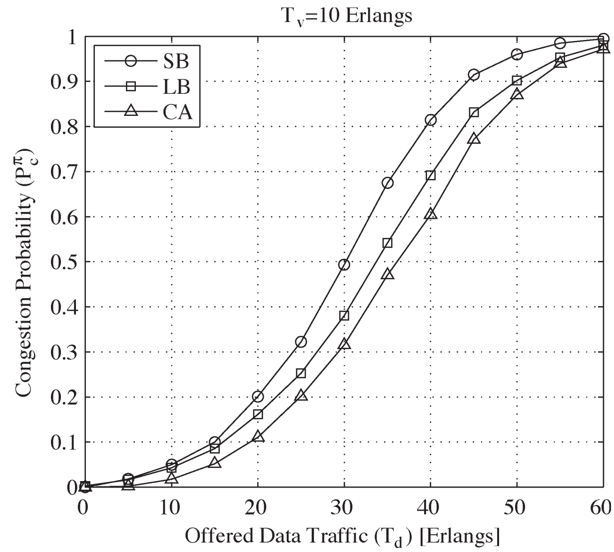

Fig. 7. Overall CP for SB, LB, and CA.

with $\xi_{(\tau_v,\tau_d)}$ defined in (37) and where q, $R_{b,d}$, and P_c^w should be particularized for the UL and DL accordingly. Then, the average throughput per data user can be obtained as

$$\Gamma_{d,u} = \sum_{S_{\boldsymbol{N}} \in \mathcal{S}} \Gamma_{d,u}(\boldsymbol{N}) \cdot P_{\boldsymbol{N}}. \quad (47)$$

Finally, it is also interesting to define the total aggregate throughput contributed by all services and RATs. For a given state $S_{\boldsymbol{N}} \in \mathcal{S}$, the aggregate throughput may be expressed as

$$\begin{aligned}\Gamma_a(\boldsymbol{N}) = &N_{t,v}\kappa_v + N_{t,d}\alpha_d q\kappa_d\xi_{(N_{t,v},N_{t,d}\alpha_d q)}\\ &+ N_{w,v}\alpha_v R_{b,v}\left[1 - P_c^w(\boldsymbol{N}_w)\right]\\ &+ N_{w,d}\alpha_d R_{b,d}\left[1 - P_c^w(\boldsymbol{N}_w)\right] \quad (48)\end{aligned}$$

where q, $R_{b,v}$, $R_{b,d}$, and P_c^w should be particularized for the UL and DL accordingly. As a result, the average aggregate throughput may be written as

$$\Gamma_a = \sum_{S_{\boldsymbol{N}} \in \mathcal{S}} \Gamma_a(\boldsymbol{N}) \cdot P_{\boldsymbol{N}}. \quad (49)$$

C. Numerical Results

Fig. 7 shows the overall CP, as a result of applying (10), when RAT-selection policies SB, LB, and CA are used. The CA policy was designed to balance the congestion probabilities in both RATs, which therefore translates into the lowest overall CP in the considered scenario. Since LB provides a similar CP in TDMA with respect to SB [see Fig. 8 (left)] and improved CP in WCDMA [see Fig. 8 (right)], the overall CP is better for the LB policy than the SB policy.

Fig. 8 shows the CP for the different considered policies in each RAT, as formulated in (8). Accordingly, since the SB policy mainly directs voice users to TDMA and congestion in TDMA is exclusively caused by data TSL reuse, hardly any congestion is detected in this RAT. Only for high offered data traffic, when WCDMA is unable to handle all the data

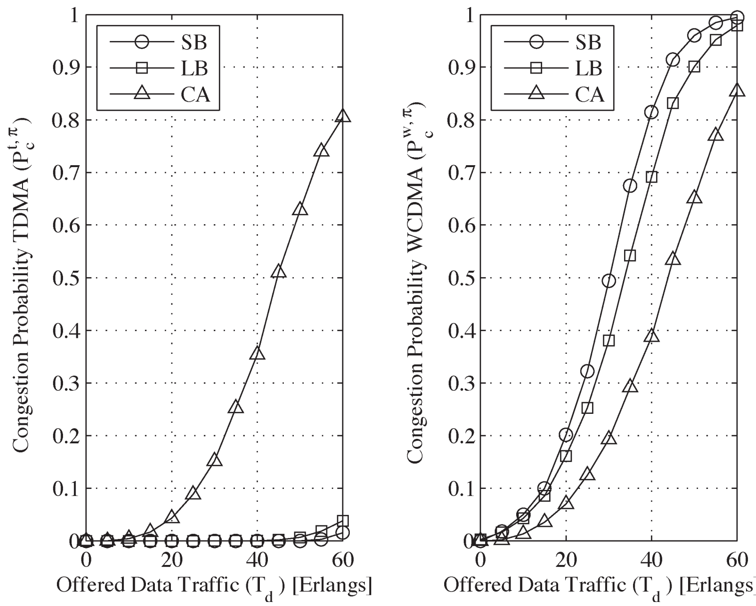

Fig. 8. CP in (left) TDMA and (right) WCDMA for SB, LB, and CA.

traffic load, does congestion in TDMA start to rise (see for $T_d = 60$ E). On the other hand, SB mainly directs data users to WCDMA, which causes congestion in this RAT to rise as the data traffic load is increased. It should be noted (see, for instance, Fig. 5) that, although excessive voice users in WCDMA may cause congestion, an excess of data users is somewhat more problematic. The LB policy, on the other hand, intends to distribute users so that the loads in each RAT [conveniently defined in (43) and (44)] are balanced. Accordingly, LB operates in such a way that data users are not forced to share TSLs in TDMA unless no capacity is left in WCDMA [17]. Then, for the case under study, LB prevents TSL reuse in TDMA, thus exhibiting a low congestion profile in this RAT. In WCDMA, LB will allocate both voice and data users; therefore, congestion will rise whenever data traffic load increases. Nevertheless, given that LB allocates voice and data resources in WCDMA, as opposed to SB, which allocates only data users, the CP in WCDMA is somewhat better for LB than SB. Finally, as expected, it can be seen how CA policy balances the CP in each RAT.

Fig. 9 shows the BP, as defined in (45), for the different policies under study. It follows from (18) that the allocation of a single voice user in TDMA implies a TSL consumption of 1 TSL over a total of C available TSLs, i.e., $1/C$. For a single data user in TDMA, this consumption is given by $\alpha_d q/n_C C$. Based on parameters given in Table III, the TSL consumption of both voice and data users in TDMA is equal to 1/23. This means that, from a resource consumption point of view, it is equally suitable to allocate voice or data users to TDMA. Moreover, the resource consumption in WCDMA may be quantified in both the UL and DL: in the UL by means of the load factor definition given in (20), with $\alpha_v^{\mathrm{UL}}/[A_v^{\mathrm{UL}}\eta_{\max}^{\mathrm{UL}}]$ and $\alpha_d^{\mathrm{UL}}/[A_d^{\mathrm{UL}}\eta_{\max}^{\mathrm{UL}}]$ being the load fractions consumed by voice and data users, respectively, and in the DL by means of the code condition provided in (22), with $1/[SF_v^{\mathrm{DL}}\mathcal{C}_{\max}]$ and $\alpha_d^{\mathrm{DL}}/[SF_d^{\mathrm{DL}}\mathcal{C}_{\max}]$ being the code consumption fractions for voice and data services, respectively.

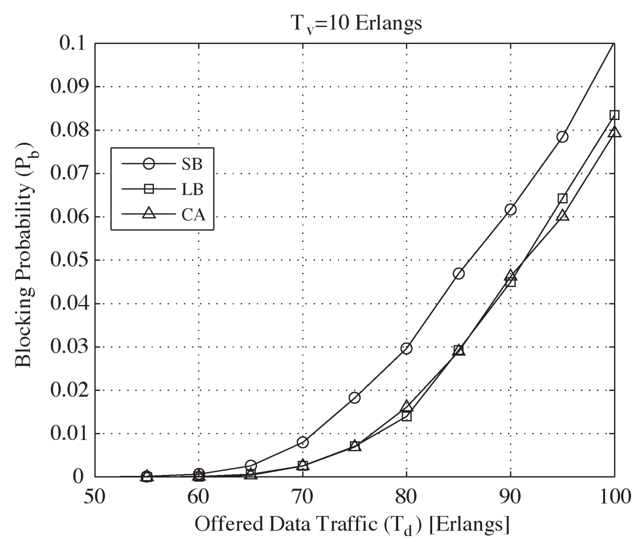

Fig. 9. Total BP for SB, LB, and CA.

With the WCDMA parameters provided in Tables I and II, it can be shown that, in WCDMA, a data user demands more resources than a voice user. In this sense, it is much more suitable, in the considered scenario, to allocate voice users in WCDMA and data users in TDMA, thus explaining the worse behavior in terms of the BP of SB, which allocates data users to WCDMA and voice users to TDMA. This is in line with conclusions raised in [11], where the allocation suitability of two services onto two RATs with linear admission limits depends on the slope of these limits, which are provided by (18), (20), and (22) and shown in Figs. 1 and 2. On the other hand, policies LB and CA are more flexible in allocating voice and data users in TDMA and WCDMA than the SB policy, thus achieving an improved BP with respect to SB.

Figs. 10 and 11 show the average DL data throughput per user and the average DL throughput, as defined in (47) and (49), respectively. For the SB policy, given that it mainly allocates

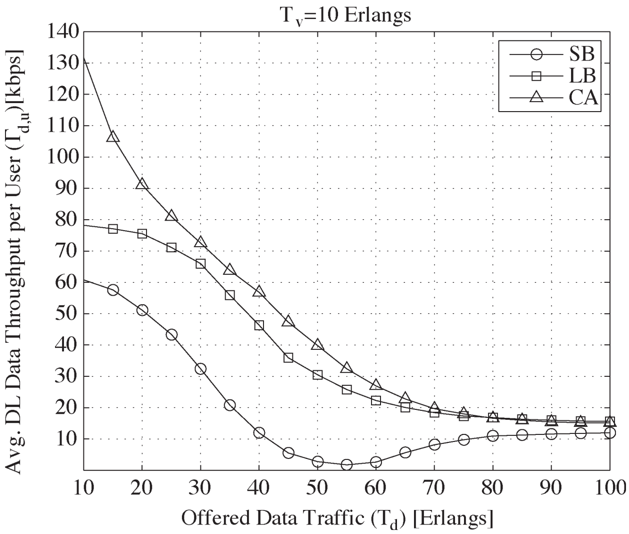

Fig. 10. DL data throughput per user for SB, LB, and CA.

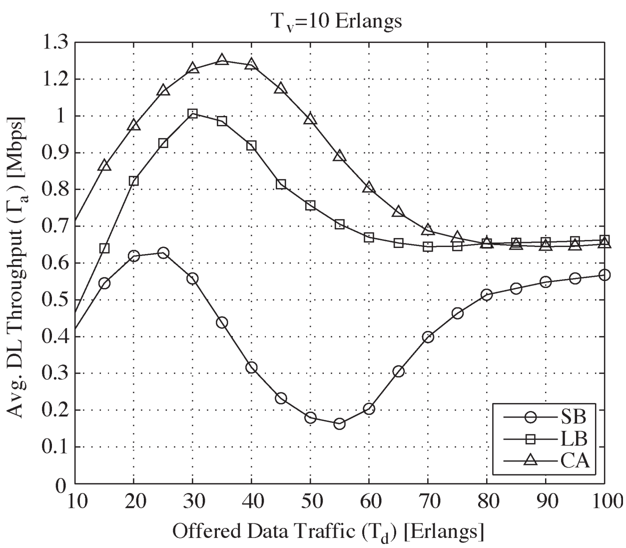

Fig. 11. DL aggregate throughput for SB, LB, and CA.

data users to WCDMA, the throughput per data user will be approximately 64 kb/s in the absence of congestion, which is the considered allocated DL bit rate for data users in WCDMA (see Table II). This happens for an offered data traffic load of $T_d = 10$ E. As the offered data traffic increases, so does the CP in WCDMA (see Fig. 8 right); thus, the throughput per data user is severely degraded up to $T_d = 55$ E. This effect is also noted in the average DL throughput in Fig. 11, where, for SB and after an initial throughput rise, the impact of congestion over data users in WCDMA causes severe throughput degradation. From $T_d = 60$ E and onward, data users start getting allocated to TDMA since WCDMA is at full capacity. Then, the throughput per user is slightly increased, given that the throughput degradation in TDMA is less harsh than that in WCDMA. This throughput increase is also noted in the average DL throughput given in Fig. 11. On the contrary, both LB and CA may eventually allocate data users to TDMA where the achieved throughput per data user depends on the reuse of TSL. If such reuse of TSL is low, with current parameter settings, the

TABLE IV
AC PARAMETERS IN TDMA AND WCDMA

AC Setup	n_C^{UL}	n_C^{DL}	η_{max}^{UL}	$\mathcal{C}_{max}$
(a)	3	4	0.9	63/64
(b)	2	3	0.8	53/64
(c)	1	2	0.7	43/64
(d)	1	1	0.6	33/64
(e)	1	1	0.5	23/64
(f)	1	1	0.4	13/64

achieved throughput per data user in TDMA can be improved with respect to WCDMA. Nevertheless, as the number of data users increases in TDMA, TSL sharing increases, and the data throughput per user decreases. It is worth noticing that the LB policy disregards the effect of throughput degradation in WCDMA due to congestion while only capturing the effect of TSL reuse in TDMA (since it prevents TSL reuse unless it is strictly necessary). This explains the improved performance of CA with respect to LB, given that CA captures both throughput degradation impacts in TDMA and WCDMA. This effect can be observed in both Figs. 10 and 11. As for the average DL throughput (see Fig. 11), to achieve high throughput values, a RAT-selection policy must achieve both low BP and high throughput per user. Then, from Figs. 9 and 10, it can be expected that policies CA and LB present an overall best performance with respect to SB and that CA outperforms LB for the case under study.

Despite having an efficient initial RAT selection policy that minimizes the occurrence of congestion events, further actions may be required in the case that such events happen. These mechanisms, which are referred to as CR mechanisms, provide the means to alleviate congestion situations by actuating over some specific network parameters. One possible mechanism for lessening the congestion status of a given RAT consists of actuating over the AC such that fewer users are allowed in the system [7], [8]. In practice, this implies a reduction in the feasible limits, as shown in Figs. 1 and 2 for TDMA and WCDMA, respectively. In WCDMA, this can be achieved by conveniently setting the maximum load factor ($\eta_{\max}^{\mathrm{UL}}$) and the maximum OVSF code condition limit ($\mathcal{C}_{\max}$) for the UL and DL, respectively. In the same way, the feasible limit in TDMA can be managed by setting the maximum number of data users per TSL n_C^{UL} and n_C^{DL} for the UL and DL, respectively.

To show the benefits of considering congestion information when performing initial RAT selection procedures, some additional results are presented. In this respect, the numerical study considers a range of values (see Table IV) for n_C^{UL} and n_C^{DL}, which determine the AC for TDMA, along with $\eta_{\max}^{\mathrm{UL}}$ and $\mathcal{C}_{\max}$, which control the AC procedure in WCDMA. A decrease in the value of these parameters indicates higher constraints in terms of user admission. Consequently, Fig. 12 shows the CP and the BP for policies SB and CA under the AC scenarios defined in Table IV. The offered voice and data traffic values are $T_v = 10$ and $T_d = 40$ E, respectively. It can be observed for both policies that, by restricting the admission, i.e., moving from case (a) to case (f), the CP is reduced. On the other hand, imposing tighter admission conditions comes at the cost of increased blocking

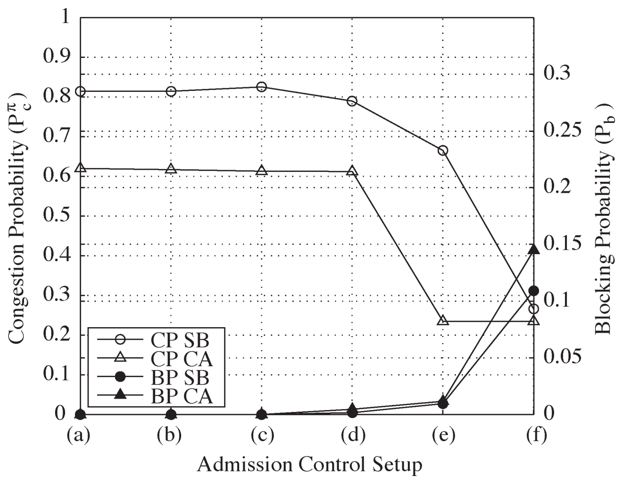

Fig. 12. BP and CP against several AC setups provided in Table IV.

probabilities, as also shown in Fig. 12, which also degrades the users' perceived QoS. Note also that, by considering the CP as an input criterion for initial RAT selection, as in policy CA, lower CP situations can be achieved at the cost of improved blocking probabilities. On the other hand, for the SB policy, to reduce the experienced CP, solely acting over the AC will severely penalize users in terms of blocking. For example, the CA policy achieves a CP of a little more than 0.2 when the AC setup is (e), which, in turn, gives a BP of below 0.02. As for the SB policy, to achieve a similar CP value, the AC needs to be more stringent than the CA policy case, which, in turn, results in a BP of more than 0.15. This suggests that, although CR mechanisms can be applied to solve congestion situations, such as actuating over the AC, an appropriate election of the initial RAT selection policy is of great importance.

D. Computational Considerations

While the values of congestion probabilities $P_c^t(\boldsymbol{N}_t)$ and $P_c^w(\boldsymbol{N}_w)$ are easily computed, determining probabilities $P_{\boldsymbol{N}}$ for all states $S_{\boldsymbol{N}} \in \mathcal{S}$ by solving the SSBEs may seem computationally complex. In general, computational complexity for solving the system of equations given by the SSBEs increases with the state dimension M and, consequently, the number of states in the state space N [28]. It was shown (see Section II-A) that the higher the number of services J and/or RATs K, the higher the dimensionality of our model M. Therefore, for large values of K and J, computational cost may dramatically increase. Nevertheless, operators may typically manage no more than three or four RATs in a given area, and although offered services are high in number, they may typically be grouped into four different QoS traffic classes according to how delay sensitive they are [29]: 1) conversational; 2) streaming; 3) interactive; and 4) background. In addition, not all RATs support all traffic classes, thus diminishing the impact on the state dimensionality. Bearing this in mind, the solution of the SSBEs may be carried out using well-known efficient numerical methods. In particular, a so-called *iterative power* procedure will be utilized for such a task [27], [28]. Other methods, such as Gauss–Seidel, have also proven to be effective (see, e.g., [30]). The iterative power operation is based on iteratively performing the product of a probability vector (of dimension $N \times 1$) with the $N \times N$ transition probability matrix $\boldsymbol{P}$. If i iterations are needed for convergence, then a total number of $i \times N^2$ multiplications are needed. 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 $i \times N_z$ multiplications are now required [28]. In addition, the involved memory storage requirements of such procedure can easily be handled by off-the-shelf computer equipment.

VII. Concluding Remarks

In this paper, a complete, detailed, and generalized framework for the evaluation of multiservice allocation in multiaccess systems by means of policies has been provided. In this sense, a generalized policy definition framework that is capable of responding to different allocation principles has been introduced. In addition, an analytical statistical characterization for radio access congestion in multi-RAT environments has also been presented. The evaluation of several RAT selection policies has been carried out in a combined TDMA/WCDMA scenario with voice and data services. In this case, specific analytical expressions for the CP have been provided for both TDMA and WCDMA.

In particular, three RAT selection policies have appropriately been defined and evaluated: LB, SB, and CA. In the case under study, results have revealed that SB prevents data-resource reuse in TDMA by allocating voice users in this RAT but at the cost of increased CP in WCDMA due to data users. Furthermore, SB provides the worst BP behavior among the considered policies. As for LB policy, it also prevents data TSL reuse in TDMA but exhibits higher flexibility in allocating voice and data users in TDMA and WCDMA. As a result, the CP and throughput per user is improved with respect to SB. However, LB disregards the impact of CP in the throughput perceived in WCDMA. Then, using congestion information, as with the CA policy, can lead to better performance in terms of both CP and throughput. In addition, the use of congestion information as a guiding principle for initial RAT selection can also prevent high blocking situations, which result from applying tighter AC mechanisms to reduce such congestion.

With the proposed framework, the impact of different RAT-selection policies on the CP can be measured, and QoS degradation is assessed. The presented framework enables extending the current results in other scenarios with other technologies and QoS requirements.

References

[1] A. Tolli, P. Hakalin, and H. Holma, "Performance evaluation of common radio resource management (CRRM)," in *Proc. IEEE ICC*, New York, 2002, vol. 5, pp. 3429–3433.

[2] A. K. Salkintzis, "Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks," *Wireless Commun.*, vol. 11, no. 3, pp. 50–61, Jun. 2004.

[3] J. Pérez-Romero, O. Sallent, R. Agusti, P. Karlsson, A. Barbaresi, L. Wang, F. Casadevall, M. Dohler, H. Gonzalez, and F. Cabral-Pinto, "Common radio resource management: Functional models and implementation requirements," in *Proc. IEEE Int. Symp. PIMRC*, Berlin, Germany, 2005, vol. 3, pp. 2067–2071.

[4] J. Zander, S.-L. Kim, M. Almgren, and O. Queseth, *Radio Resource Management for Wireless Networks*. Norwood, MA: Artech House, 2001.
[5] E. Gustafsson and A. Jonsson, "Always best connected," *Wireless Commun.*, vol. 10, no. 1, pp. 49–55, Feb. 2003.
[6] 3GPP TR 25.922 v6.0.1, *Radio Resource Management Strategies (rel. 6)*.
[7] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "Congestion control strategies in multi-access networks," in *Proc. IEEE 3rd ISWCS*, Valencia, Spain, Sep. 2006, pp. 579–583.
[8] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "On managing multiple radio access congestion events in B3G scenarios," in *Proc. IEEE 65th VTC—Spring*, Dublin, Ireland, 2007, pp. 1137–1141.
[9] 3GPP TR 25.881 v5.0.0, *Improvement of RRM Across RNS and RNS/BSS*.
[10] 3GPP TR 25.891 v0.3.0, *Improvement of RRM Across RNS and RNS/BSS (post rel-5) (release 6)*.
[11] A. Furuskar and J. Zander, "Multiservice allocation for multiaccess wireless systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 1, pp. 174–184, Jan. 2005.
[12] G. Fodor, A. Furuskar, and J. Lundsjo, "On access selection techniques in always best connected networks," in *Proc. 16th ITC Spec. Semin. Performance Eval. Wireless Mobile Syst.*, Antwerp, Belgium, 2004.
[13] R. Chakravorty, P. Vidales, K. Subramanian, I. Pratt, and J. Crowcroft, "Performance issues with vertical handovers—Experiences from GPRS cellular and WLAN hot-spots integration," in *Proc. 2nd IEEE Annu. Conf. Pervasive Comput. Commun.*, Orlando, FL, Mar. 2004, pp. 155–164.
[14] J. Pérez-Romero, O. Sallent, and R. Agustí, "A generalized framework for multi-RAT scenarios characterisation," in *Proc. IEEE 65th Veh. Technol. Conf.*, Dublin, Ireland, Apr. 2007, pp. 980–984.
[15] L. Giupponi, R. Agustí, J. Pérez-Romero, and O. Sallent, "A novel approach for joint radio resource management based on fuzzy neural methodology," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1789–1805, May 2008.
[16] F. Malavasi, M. Breveglieri, L. Vignali, P. Leaves, and J. Huschke, "Traffic control algorithms for a multi access network scenario comprising GPRS and UMTS," in *Proc. 57th IEEE Semiannual VTC*, Jeju, Korea, Apr. 2003, vol. 1, pp. 145–149.
[17] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A 4-dimensional Markov model for the evaluation of radio access technology selection strategies in multiservice scenarios," in *Proc. IEEE VTC—Fall*, Montreal, QC, Canada, Sep. 2006, pp. 1–5.
[18] X. Gelabert, J. Pérez-Romero, O. Sallent, and R. Agustí, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 10, pp. 1257–1270, Oct. 2008.
[19] T. Halonen, J. Melero, and J. Romero-Garcia, *GSM, GPRS and EDGE Performance: Evolution Toward 3G/UMTS*. New York: Wiley, 2002.
[20] J. Pérez-Romero, O. Sallent, R. Agustí, and M. Díaz-Guerra, *Radio Resource Management Strategies in UMTS*. Hoboken, NJ: Wiley, 2005.
[21] S. Nanda, "Teletraffic models for urban and suburban microcells: Cell sizes and handoff rates," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 673–682, Nov. 1993.
[22] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. VT-35, no. 3, pp. 77–92, Aug. 1986.
[23] C. Hartmann and N. Nasser, "Modeling and performance analysis of multi-service wireless CDMA cellular networks using smart antennas," *Wirel. Commun. Mob. Comput.*, vol. 9, no. 1, pp. 117–129, Jan. 2009.
[24] Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "Performance analysis for dual band PCS networks," *IEEE Trans. Comput.*, vol. 49, no. 2, pp. 148–159, Feb. 2000.
[25] J. Wang, Q.-A. Zeng, and D. P. Agrawal, "Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 1, pp. 65–75, Jan.–Mar. 2003.
[26] L. Kleinrock, *Queueing Systems*, vol. 1. Hoboken, NJ: Wiley, 1975.
[27] G. Bolch, S. Greiner, H. D. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications*. New York: Wiley-Interscience, 2006.
[28] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*. Princeton, NJ: Princeton Univ. Press, 1994.
[29] 3GPP TS 23.107 v5.7.0, *Quality of Service (QoS) Concept and Architecture (rel. 5)*.
[30] S. Lincke, "Validation of a load shared integrated network with heterogeneous services," in *Proc. 40th ANSS*, Norfolk, VA, Mar. 2007, pp. 49–59.

Xavier Gelabert (M'08) was born in Alaior, Spain, in 1978. He received the M.S. degree in electrical engineering from the Royal Institute of Technology, Stockholm, Sweden, in 2003 and the Telecommunications Engineering (B.S. plus M.S.) degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2004. He is currently working toward the Ph.D. degree with the Department of Signal Theory and Communications, UPC.

From August to December 2008, he was a Visiting Researcher with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta. Since January 2009, he has been a Visiting Researcher with the iTEAM—Mobile Communications Group, Technical University of Valencia, Valencia, Spain. He has participated in European-funded projects EVEREST, AROMA, and E^3, along with Spanish projects COSMOS and COGNOS. His current research interests are mobile radio communication systems, focusing on radio resource management strategies in multiaccess networks and opportunistic/cognitive spectrum management.

Oriol Sallent received the Telecommunications Engineer (B.S. plus M.S.) and Ph.D. degrees from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 1994 and 1997, respectively.

He is currently an Associate Professor with the Department of Signal Theory and Communications, Universitat Politècnica de Catalunya. He has participated in many research projects and consultancies funded by either public organizations or private companies. He is currently participating in E^3 project within the 7th Framework Program of the European Commission. His research interests are radio resource and spectrum management for heterogeneous cognitive wireless networks, on which he has authored more than 100 papers in IEEE journals and conference proceedings.

Jordi Pérez-Romero (M'04) received the Telecommunications Engineering (B.S. plus M.S.) and Ph.D. degrees from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1997 and 2001, respectively.

He is currently an Associate Professor with the Department of Signal Theory and Communications, UPC. He has been involved in several different European projects as well as projects for private companies. He has authored papers in international journals and conference proceedings and has coauthored one book on mobile communications. His research interests are mobile communication systems, particularly packet radio techniques, radio resource and quality-of-service management, heterogeneous wireless networks, and cognitive networks.

Ramon Agustí (M'78) received the Engineer of Telecommunications (B.S. plus M.S.) degree from the Universidad Politécnica de Madrid, Madrid, Spain, in 1973 and the Ph.D. degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1978.

He is currently with the Department of Signal Theory and Communications, UPC. In 1973, he joined the Escola Tècnica Superior d'Enginyers de Telecomunicació de Barcelona, where he became a Full Professor in 1987. After graduation, he worked on digital communications, particularly on the transmission and development aspects in fixed digital radio: both radio relay and mobile communications. He participated in the European programs COST 231 and COST 259 as a Spanish delegate. He has also participated in the RACE, ACTS, and IST European research programs, as well as in many private and publicly funded projects. He is part of the editorial board of several scientific international journals and has been conducting a postgraduate annual course on mobile communications since 1995. For the last 15 years, he has mainly been concerned with performance analysis and the development of planning tools and equipment for mobile communication systems and has authored about 200 papers in these areas. He coauthored two books on mobile communications.

Dr. Agustí was the recipient of the Catalonia Engineer of the Year Prize in 1998 and the Narcís Monturiol Medal issued by the Government of Catalonia in 2002 for his research contributions to the field of mobile communications.