Impact of the Connection Admission Process on the Direct Retry Load Balancing Algorithm in Cellular Networks

Przemyslaw Pawelczak, Member, IEEE, Shaunak Joshi, Member, IEEE, Sateesh Addepalli, John Villasenor, Senior Member, IEEE, and Danijela Čabrić, Member, IEEE

Abstract—We present an analytical framework for modeling a priority-based load balancing scheme in cellular networks based on a new algorithm called direct retry with truncated offloading channel resource pool (DR). The model, developed for a baseline case of two cell network, differs in many respects from previous works on load balancing. Foremost, it incorporates the call admission process, through random access. In specific, the proposed model implements the Physical Random Access Channel used in 3GPP network standards. Furthermore, the proposed model allows the differentiation of users based on their priorities. The quantitative results illustrate that, for example, cellular network operators can control the manner in which traffic is offloaded between neighboring cells by simply adjusting the length of the random access phase. Our analysis also allows for the quantitative determination of the blocking probability individual users will experience given a specific length of random access phase. Furthermore, we observe that the improvement in blocking probability per shared channel for load balanced users using DR is maximized at an intermediate number of shared channels, as opposed to the maximum number of these shared resources. This occurs because a balance is achieved between the number of users requesting connections and those that are already admitted to the network. We also present an extension of our analytical model to a multicell network (by means of an approximation) and an application of the proposed load balancing scheme in the context of opportunistic spectrum access.

Index Terms—Cellular networks, load balancing, connection admission, 3GPP, opportunistic spectrum access

1 INTRODUCTION

Given the rapid current and expected growth in 3G/UMTS and LTE-based networks and in the number of mobile devices that use such networks to download data intensive, multimedia-rich content, the need for QoS-enabled connection management is vital. However, the nonuniform distribution of users and consequent imbalance in usage of radio resources leads to an existence of local areas of under- and overutilization of these resources in the network. This phenomenon results in challenging network management issues. Load balancing is an important technique that attempts to solve such issues, and occurs when a centralized network controller intelligently distributes connections from highly congested cells to neighboring cells which are less occupied. This allows for an increase in network subscriber satisfaction because more subscribers meet their QoS requirements. Furthermore, it allows for an increase in overall channel utilization by leveraging the fact that users with access to multiple cells also have access to additional resources.

In this work, we aim to quantify the impact of load balancing on the overall system as well as user experience (from the separate viewpoints of users that share resources and those that use these shared resources) using a detailed analytical model of a fundamental two cell setup, later extended by means of approximations to multicell setup. Furthermore, we present an application of our model in the context of a cellular system using opportunistic spectrum access (OSA), which even further improves the teletraffic properties of the considered cellular network, to efficiently utilize system resources.

1.1 Related Work

Load balancing has been explored for many years as described, for example, in earlier works such as [2], [3]. More recent works such as [4], [5], [6], [7] have also examined the impacts of various load balancing schemes.

For simplicity, the previous studies of load balancing have assumed that the connection admission process can be neglected because nonfinite user populations are considered with a particular arrival rate of the number connections. In cellular networks, however, finite user population exists, where each new connection needs to first send a request to the serving base station (BS) through some predefined control channel. In 3GPP standards, this control channel is the Physical Random Access Channel (PRACH) [8, Section 2.4.4.4], [9, Section II]. The success of a connection request by a user is dependent on multiple factors including the number of requesting users; the pairwise channel quality between the user and the serving BS (measured, for example, in BER or outage probability);
and the actual control channel access technique itself [10], [11], [12].

More recent works, such as [13], use inter- and intracell handover techniques to alleviate issues of coverage for cell-edge users without specific focus on traffic parameters. Other recent studies propose the use of low-power cellular relays in the presence of high-power BSs with more emphasis on a heuristic to determine load balancing and mobile association [14]. It is evident that current literature on load balancing lacks an analytical model that allows for substantial flexibility in terms of finite user population consideration and the use of various traffic parameters.

1.2 Our Contribution

Until now the exact impact of random access overhead on load balancing performance is not well understood. More specifically, the quantitative relationship between the random access phase length, the user blocking probability, and system channel utilization in a load balancing-enabled cellular system is unknown. To the best of our knowledge an analytical model to quantify system wide and user experience metrics in this respect has not been previously provided. Our work:

1. Through development of an analytical model for a baseline (fundamental) two-cell network, demonstrates the benefits of load balancing, from a teletraffic point of view, using realistic traffic scenarios, various network configurations and parameters, and simplified physical layer model for channel quality.

2. Provides a detailed connection admission process to determine the effects of a finite user population on the efficiency of load balancing performance metrics.

3. Allows extension to more complex network setups. More specifically, we present an approximation to a multicell case and also provide an application of the model in the context of cellular OSA.

Our model can be used in:

1. Demonstrating that the random access phase length is an important tool that can be used to control both system wide and user experience performance metrics, for example, quantifying the tradeoffs between random access channel (RACH) collision probability and blocking probability for load-balanced users as a function of random access phase length.

2. Determining the impact of channel quality to improve the accuracy of reported load balancing efficiency, for example, quantifying the loss in accuracy due to perfect channel condition assumptions.

3. Exploring the effects of varying shared channel access on system-wide performance and user experience, for example, determining the marginal gain of adding more channels for shared access between cells.

The rest of the paper is organized as follows: The system model is introduced in Section 2, while the analytical model is introduced in Section 3. The numerical results are presented in Section 4. Lastly, the paper is concluded in Section 5.

2 System Model

In the following sections, we will describe the system model in detail. We start with a description of the channel structure in Section 2.1, followed by the description of node placement in Section 2.2. Section 2.3 describes the signal transmission model and Section 2.4 prioritization policies in load balancing. Then, in Section 2.5, introduces a random access process in the context of cellular networks, followed by the introduction of a data transfer model in Section 2.6. Finally, the whole load balancing process is introduced in Section 2.7.

2.1 Channel Structure

We consider a cellular system, where two BSs are positioned such that they create a region of overlap in coverage. Naturally, while cellular systems typically have far more than two BSs, a reduction to a two-BS system for analytical reasons enables a tractable analytical framework while still allowing exploration of a large number of microscopic parameters to use in optimizing network performance. Furthermore, two-cells provide a fundamental functional pair for the purposes of studying load balancing from the context of off-loading connections from a highly congested cell to a less congested one. We strongly emphasize that consideration of load balancing in the context of a two-BS system has been used extensively and successfully in previous treatments, e.g., [4], [7], [15], [16].

Cell 1 has $M_1$ available basic bandwidth units, as referred to in [17] or more commonly referred to as channels, and cell 2 has $M_2$ available channels. Note that channels can also represent WCDMA codes in the context of UMTS. The throughput of every channel in each cell is the same and equal to $R$ bits/second. We assume that channels are mutually orthogonal and that there is no interference in the set of channels belonging to cells 1 and 2. Each BS emits a signal using omnidirectional antennas and we assume a circular contour signal coverage model, in which full signal strength is received within a certain radius of the BS, and no signal is available beyond that radius, as used in, e.g., [2], [4], [13], [18].

2.2 Node Placement

Each mobile terminal, referred to as user equipment (UE), remains at fixed positions following an initial UE placement process, with each UE located in one of three separate regions. $N_1$ UEs are in group 1 and have access only to one BS, $N_2$ UEs are in group 2 and have access only to the second BS, and $N_3$ UEs are in group 3 and can potentially access either BS. Such nonhomogenous UE placement has been considered, for example, in [16], which allows for a tractable analysis of the considered system and includes all important groups participating in the load balancing process. The nonhomogenous case of UE distribution is the most widespread because UEs are generally distributed nonuniformly over a cellular area. UEs in group 3 are in the region of overlap in coverage between the two cells, also known as the Traffic Transferable Region (TTR) [4]. Since only one UE can occupy a channel at a time, a maximum of

1. For other, more involved channel assignment procedures including, for example, channel bonding the reader is referred to, e.g., [19], where multiple channel assignment problem is studied in an ad hoc scenario.
$M_1 + M_2$ UEs can be connected to both cells in the system at any given time. We assume that UEs from groups 1 and 3 are initially registered to cell 1 (serving as the overloaded cell), while UEs in group 2 are initially registered to cell 2.

2.3 Signal Transmission Model

Time is slotted and the minimum time unit is a frame length of $\tau$ seconds. Connections and channel conditions are assumed to remain constant for the duration of a frame, though they will in general vary from frame to frame. We assume that Adaptive Modulation and Coding (AMC) is not used in this framework because it does not aid in evaluating load balancing performance. Please note that AMC is not considered in similar previous works such as [2], [4], [18]. For simplicity, we also do not consider advanced error control methods such as Hybrid Automatic Repeat Request.\(^2\) On the other hand, we do assume that the connection and termination processes for UEs are directly dependent on the channel states experienced between each group of UEs and the BSs they are connected to. We also assume that channel states are binary and independent from slot to slot, much as occurs in [22], and that all of the UEs in each group experience the same channel quality to a given BS. Therefore, in any given time slot a UE is either experiencing a good state with probability $w_{x,y}(i)$ (denotes the particular pair-wise connection between group $x$ of UEs and associated BS $y$, and $i \in \{d, u\}$ denotes the downlink and uplink, respectively), or a bad state, with probability $1 - w_{x,y}(i)$. The value of $w_{x,y}(i)$ is dependent on the distance between UEs in group $x$ and BS $y$, which is denoted as $d_{x,y}$. In our analysis, we use the distance as an input to a combined path loss and shadowing propagation model. This serves as an average channel quality consideration for the model instead of the use of channel quality indicators on the uplink per transmitted packet.

2.4 Prioritization

Because of the strict boundaries between groups of UEs, we assign priorities on a per group basis. A single, higher priority is given to all UEs in groups 1 and 2 because there is no flexibility to reassign them to a different BS; group 3 UEs can potentially be reassigned and thus given a lower priority. Priorities for UEs are determined on a per time slot basis. A very similar priority model has been used in other treatments of load balancing. For example, in [15], [23], newly arriving connections in the non-TTR are given first priority to acquire channels from their serving BSs, while the connections from the TTR are assigned to the remaining channels. Our model allows for the implementation of a wide range of scenarios that require such traffic prioritization. One potential application is for the modeling of networks, where load balancing traffic originating from UEs in the TTR has lower priority than nonbalanced connections due to several factors including a lower average channel quality [24], [25], QoS requirements [26], nonuniform spatial distribution of traffic classes [16], or cell dwell time. Furthermore, it allows for modeling integrated hybrid cellular/WLAN/Ad Hoc networks as discussed in [6], [27], where noncellular terminals in the TTR have a lower priority than cellular UEs, and hierarchical cellular systems [28], where members of different tiers have independent priorities. Finally, it enables the modeling of femtocell traffic prioritization, where UEs in groups 1 and 3 are those in the Closed Subscriber Group (CSG) [29], while UEs in the TTR are neighboring UEs outside of the CSG.

2.5 Random Access

In the connection process, a UE first attempts to connect to the BS it is initially registered to by requesting a connection through a RACH. We assume a frequency division duplex transmission mode, where control and data traffic are transmitted and received simultaneously. In addition, time division duplexing is considered during the transmission of control packets. Specifically, each UE generates a connection request with probability $p_x$. A connection is requested randomly in one of $L_x \leq \tau$ nonoverlapping, time slotted control resources, unique to group $x \in \{1, 2, 3\}$ of UEs. In other words, each group has a unique set of sub slots within a frame during which UEs may, but are not required to, request a connection. The random access phase length is equal to slot length $\tau$. Collisions between connection requests from UEs in the same group are possible.

The random access procedure considered in this work shares features of the 3GPP-based cellular network standards, which use the PRACH, mapped on a one-to-one basis to the logical RACH. RACH uses the S-ALOHA protocol and, in relation to the priorities assumed in this paper, allows the prioritization of connection requests based on Active Service Classes (ASC) [10] that are unique to each UE, and can be adapted by the 3GPP-MAC layer once the UE is in connected mode [8, Section 2.4.2.6]. The BS advertises itself to the UEs within range through the broadcast channel using signatures (3GPP release 99, e.g., UMTS), subcarriers (3GPP release 8, e.g., LTE), or time slots, which each ASC can in-turn use for connection requests on RACH. The adaptation of the ASC is performed in the time intervals predefined by the operator. For the purpose of our paper, we assume that the BSs collectively, through the Radio Network Controller, map the received signal from every registered UE to an associated ASC.

We assume a zero-persistence protocol, i.e., a collision during a connection request implies that connections are lost, and also UEs do not retry to generate another dependent connection. Due to this assumption, a power ramping process, i.e., feedback from the UE to the BS on an unsuccessful connection request [10, Section II-B], is redundant. To isolate the impact of each group of UEs on collision rates, we assume mutually exclusive RACH resources assigned to each ASC [11, Fig. 4]. Analysis of PRACH performance in isolation can be found in [9], [11].

2.6 Data Transfer

A connection request is granted during the connection arrangement process if a good channel state occurs between the UE and its associated BS at the moment of the request, and if no collisions occur between multiple requests from different UEs. Once a connection is established, the BS randomly selects a channel and assigns the connected UE to it. The UE then begins to receive downlink data. UEs

---

\(^2\) Recent papers [20], [21] are a good source of information on the performance of Hybrid Automatic Repeat Request.
occupy a time slot with probability \( q \), where \( 1/q = r_p/(R	au) \) is the average connection transfer size and \( r_p \) is the average packet size given in bits. We assume that the transfer size is at least one time slot long. A connection terminates either when a transmission completes, or when the channel is in a bad state during transmission.

### 2.7 Load Balancing Process

In the case of a UE in group 3, if a connection request is successful and there are no resources available in cell 1, we assume that the radio network controller performs load balancing by transferring the call from cell 1 to cell 2. To avoid overloading cell 2 and protecting UEs that are already registered to it, UEs in group 3 can access a maximum of \( K \) channels from cell 2, where \( 0 \leq K \leq M_2 \). UEs in group 3 have access to an additional \( K \) shared channels; therefore, they have access to a total of \( M_1 + K \) channels.

Using the nomenclature of [4, Sections 2 and 3], this load balancing scheme belongs to a class of direct load balancing schemes. It is closest in operation to direct retry [2]. Since we allow at most \( K \) available channels to offload traffic from cell 1 to cell 2 (as in simple borrowing scheme [4, Section 2]) we denote our scheme as direct retry, while for group 2 (as in simple borrowing scheme [4, Section 2]) we denote our scheme as direct retry with truncated offloading channel resource pool (abbreviated as DRC). With \( K = M_2 \) our scheme reduces to classical direct retry, while for \( K = 0 \) it reduces to a system in which no load balancing takes place.

In our model, we do not use a take-back process, i.e., bidirectional load balancing is not considered. That is, once connections from group 3 are offloaded onto cell 2, they remain connected to cell 2 during the transmission despite whether or not resources have been freed in cell 1. In [5], the authors remark that the take-back process, although more advantageous to the network due to the high load balancing takes place.

In our model, we do not use a take-back process, i.e., bidirectional load balancing is not considered. That is, once connections from group 3 are offloaded onto cell 2, they remain connected to cell 2 during the transmission despite whether or not resources have been freed in cell 1. In [5], the authors remark that the take-back process, although more advantageous to the network due to the high load balancing takes place.

Using the nomenclature of [4, Sections 2 and 3], we define the state transition probability as

\[
Pr\left(A_t = a_t, Y_t^{(1)} = b_t, Y_t^{(2)} = c_t, C_t = d_t\right) = \rho(a, b, c, d, \alpha, \beta, \gamma, \delta, \epsilon) \Pr\left(A_{t-1} = a_{t-1}, Y_{t-1}^{(1)} = b_{t-1}, Y_{t-1}^{(2)} = c_{t-1}, C_{t-1} = d_{t-1}\right),
\]

where subscripts \( t \) and \( t-1 \) denote the current and the previous time slots, respectively. This allows for computation of the transition probability matrix required to obtain the performance metrics of the load balancing-enabled cellular system. In the subsequent sections, we describe the process of deriving the transition probability \( \rho(a, b, c, d, \alpha, \beta, \gamma, \delta, \epsilon) \). We begin by explaining the computation process for the channel quality, and then focus on the derivation of the functions that support (1).

### 3.1 Derivation of Channel Quality

In the downlink, the probability of a UE belonging to group \( x \in \{1, 2, 3\} \) and receiving a good signal from BS \( y \in \{1, 2\} \), is defined as

\[
w_{x,y}^d = 1 - \Pr\left(J_{x,y}^d \leq \gamma_q^d\right) = 1 - \int_0^{\gamma_q^d} p_{s_{x,y}}(\gamma, d_{x,y}) d\gamma,
\]

where \( \gamma_q^d \) is the signal reception threshold for the downlink, expressed as the minimum required received power of the received signal \( J_{x,y}^d \), and \( p_{s_{x,y}}(\gamma, d_{x,y}) \) is the distribution of the signal \( \gamma \) received at group \( x \), which is at a distance of \( d_{x,y} \) from BS \( y \). As an example, we consider an environment with path loss and shadowing, where \( w_{x,y}^d \) is expressed in [30, (2.52)] as

\[
w_{x,y}^d = Q\left(\frac{\gamma_q^d - P_t - 10 \log_{10} W^d(d_{x,y}) + 10 \log_{10} \frac{d_{x,y}}{d_0^d}}{\sigma_f}\right),
\]

where \( W^d(d) \) is a unit-less constant dependent on both the antenna characteristics and an average channel attenuation, and for \( W^d < 1 \) approximating empirical measurements and assuming omnidirectional antennas is given as [30, (2.41)],

\[
W^d dB = 20 \log_{10} \frac{\lambda}{4\pi d_0^d},
\]

where \( \lambda \) is the wavelength of the carrier frequency; \( P_t^d \) is the BS transmitted power (which is assumed to be the same for both BSs, \( d_{x,y} \) is the distance of the UE in group \( x \) located farthest from BS \( y \), \( d_0^d \) is the reference distance for the BS.
antenna far field; $\sigma_y$ is the log-normal shadowing variance given in dB; $\delta$ is the path loss exponent; and the $Q$ function is defined in the usual manner as

$$Q(z) \triangleq \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$  

(5)

In the uplink, the probability that a good signal is received by BS $y$ from a UE in group $x$ is $w_{x,y}^{(u)}$ and is expressed in the same manner as (2) and (3), by replacing all variables having superscript $(d)$ with $(u)$, where $P_{x}^{(u)}$ denotes the UE transmission power; $W_{x}^{(u)}$ denotes the constant for the UE antenna (which again can be calculated in the same manner as $W_x^{(d)}$ using (4)); $d_{x,y}^{(u)}$ is the reference distance for the UE antenna far-field; and $\gamma_{x}^{(u)}$ is the signal reception threshold for the uplink. The downlink and uplink channel quality information governs the success rate of the connection admission process, as well as the duration of a downlink transmission.

### 3.2 Derivation of Connection Arrangement Probability

An important feature of the model is the consideration of connection admission in the load-balancing process. This process is a function of the total number of UEs, the number of UEs receiving data from their respective serving BSs, the pairwise channel quality between the UEs and its serving BSs, and the underlying random access algorithm. The probability that $j$ new connections have successfully requested downlink data given $i_{t-1}$ in $\{a_{t-1}, b_{t-1}, c_{t-1}, d_{t-1}\}$ currently active connections from group $x$, with a RACH consisting of $L_x$ time slots is

$$S_{i_{t-1}}^{(j)} \triangleq \begin{cases} \zeta_{i_{t-1},x,y}^{(j)}, & j > 0, \\ 1 - w_{x,y}^{(u)}(i_{t-1},x,y) + \zeta_{i_{t-1},x,y}^{(1)}, & j = 0, \\ 0, & \text{otherwise}, \end{cases}$$

(6)

where

$$\zeta_{i_{t-1},x,y}^{(j)} \triangleq \frac{N_{i-1} - j}{N_{i-1}} \beta_{k}^{(j)} w_{x,y}^{(u)};$$

(7)

$p_x$ is the probability of a connection request by an individual UE in group $x$; and $\beta_{k}^{(j)}$ is the probability that among $k$ UEs requesting a connection, $j$ were successful in obtaining a resource. Note that the reference to $x, y$ in $S_{i_{t-1}}^{(j)}$ is omitted due to space constraints, keeping in mind that for $a_{t-1}, a_{t-1} = 1, y = 1$, for $b_{t-1}, b_{t-1} = 3, y = 1$, for $c_{t-1}, c_{t-1} = 3, y = 2$, and for $d_{t-1}, d_{t-1} = 2, y = 2$.

For consistency with cellular networks such as 3GPP, we consider a PRACH-like control channel, for which $\beta_{k}^{(j)}$ can be described in the manner of [31, 33]

$$\beta_{k}^{(j)} = \sum_{m=j}^{\min(k,L_x)} \frac{(-1)^{m+j}(L_x - m)^{k-m}k!}{(m-j)!(L_x - m)!(k-m)!m!}.$$  

(8)

Note that depending on the assumption of how collisions are resolved, different definitions of $\beta_{k}^{(j)}$ in (8) can be applied when calculating the connection arrangement probability according to (6).

### 3.3 Derivation of Connection Termination Probability

Once a UE successfully requests a connection from the serving BS, a downlink transmission is started provided that at least one free channel is available for the UE. The connection terminates when the BS finishes transmitting data to the UE or when the downlink signal received by the UE is in outage. The probability that $j$ connections from $i_{t-1}$ active connections at group $x$ terminate is

$$T_{i_{t-1}}^{(j)} \triangleq \binom{i_{t-1}}{j} \gamma_x^{(j-1)}(1 - \lambda_{x,y})^{i_{t-1}-j},$$

(9)

where $\lambda_{x,y} = 1 - w_{x,y}^{(d)} + w_{x,y}^{(d)}$ is the inverse of the average packet length, accounting for truncation of some packets due to a bad channel quality. Again, the indices $x, y$ have been omitted for notational simplicity in the symbol $T_{i_{t-1}}^{(j)}$, assuming that the same relationship between $x, y$ and $j, i_{t-1}$ as given in Section 3.2 holds.

### 3.4 Derivation of the State Transition Probability

Using the definitions of the arrangement and termination probabilities, expressed in (6) and (9), respectively, we can finally introduce the state transition probabilities for the complete model. The transition probability is constructed using the termination and arrangement probability definitions and the respective relationship between the variables of these two definitions (which are dependent on the start and end states of the transition). Due to the complexity of the derivation we begin with a highly simplified example.

#### 3.4.1 State Transition Probabilities for a Single UE Group

To facilitate understanding the derivation of the complete state transition probabilities, we first consider a network in which no load balancing occurs and all of the UEs are in group 1, such that $N_1 > 0$ and $N_2 = N_3 = 0$. The state of the Markov chain then simplifies to $\{A, 0, 0, 0\}$ and the transition probability becomes $r_{a_{t-1},0,0,0}$, where

$$r_{a_{t-1},0,0,0} = \begin{cases} \sum_{i=0}^{a_{t-1}} T_{a_{t-1}}^{(i+1-a_{t-1})} S_{a_{t-1}}, & a_{t-1} \geq a_t, \\ \sum_{i=0}^{a_{t-1}} T_{a_{t-1}}^{(i)} S_{a_{t-1}}, & a_{t-1} < a_t \leq a_t, \\ \sum_{i=0}^{a_{t-1}} T_{a_{t-1}}^{(i)} S_{a_{t-1}}^{(i-a_{t-1})}, & a_{t-1} < a_t, \\ \sum_{i=0}^{a_{t-1}} T_{a_{t-1}}^{(i)} S_{a_{t-1}}^{(i-a_{t-1})}, & a_{t-1} < a_t \leq a_t, \\ \sum_{i=0}^{a_{t-1}} T_{a_{t-1}}^{(i)} S_{a_{t-1}}^{(i-a_{t-1})}, & a_{t-1} < a_t \leq a_t. \end{cases}$$

(10)

In (10), the case $a_{t-1} \geq a_t, a_t < M_1$ denotes the transition from a higher to a lower channel occupancy, subject to the constraint that the number of channels occupied in the end state must be less than the total BS capacity. The number of terminating UEs is set to compensate for the UEs that generate successful connections. The case $a_{t-1} < a_t, a_t < M_1$ denotes the transition from a lower to a higher channel
occupancy (given, again, that the number of occupied channels is less than the total BS capacity). In this case, UEs from group 1 need to generate exactly as many connections as given by the end state, not forgetting to generate connections to compensate for the total number of terminations. Lastly, for the case of $a_{i-1} \leq a_i$, $a_i = M_1$ the end state equals the total channel capacity. The first term in the definition of this transition probability includes exactly the number of connections needed to reach the end state, once again compensating for terminations. The second term accounts for all successful connections generated that exceed those needed to reach the end state, which will not be admitted.

3.4.2 General Solution for the State Transition Probabilities

Due to the complexity of the general solution, the main analytical equations describing the construction of the four-dimensional transitional probability matrix are presented in the Appendix B, available in the online supplemental material.

3.5 Performance Metrics

Given the complete description of the system, we are able to derive important metrics that would describe the efficiency of the load balancing process involving connection admission. While there are many performance metrics that can be extracted given the above framework, we focus on three primary metrics: 1) the blocking probability, which describes the probability that at least one UE which requests a connection from a particular group will be denied access to a channel; 2) the channel utilization, which expresses the fraction of the available channels are being used; and 3) the collision probability on a control channel, which provides the probability that at least one requesting connection will be lost due to a collision with another UE.

3.5.1 Blocking Probability

As used here, blocking occurs when at least one UE requests a new connection, but cannot be admitted to any BS due to lack of available channels. Since each group has access to a different number of channels and can follow a different connection strategy, it is necessary to define separate blocking probability metrics for UEs in groups 1 and 2, as contrasted with UEs in group 3. For groups 1 and 2, the blocking probability is defined according to

$$
B^{(2)} = \sum_{a,b,c,d} \sum_{i=0}^{x} \sum_{j=0}^{N_d} \sum_{k=0}^{y} \pi_{a,b,c,d} T^{(i)}_{a} S^{(j)}_{b} \left( i + j + k \right),
$$

where for $x = 1$: $a = y = b = v = 1$ and for $x = 2$: $a = d$, $y = c$, $v = 2$. For group 3 UEs, the blocking probability is given as

$$
B^{(3)} = \sum_{a,b,c,d} \sum_{i=0}^{x} \sum_{j=0}^{N_d} \sum_{k=0}^{y} \sum_{l=0}^{N_l} \sum_{m=0}^{N_m} \pi_{a,b,c,d} T^{(i)}_{a} S^{(j)}_{b} S^{(l)}_{c} S^{(m)}_{d} \left( i + j + k + l + m \right),
$$

where $\phi$ is defined separately for $K < M_2$ and $K = M_2$. For $K = M_2 \phi = M_1 + M_2 - a - b - c - d - g_a - g_d + i_a + i_d + 1$, where

$$
g_a = \begin{cases} 
M_1 - a + i_a - b + i, & j_a > M_1 - a + i_a - b + i, \\
\text{otherwise}, & 
\end{cases}
$$

and $g_c$ is defined as (13) replacing $M_1$ with $M_2$, $a$ with $c$, $i_a$ with $i_d$, $b$ with $d$ and $i$ with $j$. For $K < M_2 g_a = M_1 - g_a - b - j_a + i_a + i + g_c + 1$, where $g_c$ is defined as (13) and

$$
g_c = \begin{cases} 
\max \left\{ 0, g_c \right\}, & g_c < K, \\
K - c + j, & c - j < K \text{ and } g_c \geq K, \\
0, & \text{otherwise}, 
\end{cases}
$$

where $g_c = M_2 - c - d + i_d + j - j_d$.

We briefly explain the above equations. The derivation of the blocking probability for group 3 UEs is more complicated than for those UEs in groups 1 and 2 because this group can access channels from both cells. Therefore, the number of generations for group 3 UEs that leads to blocking has to account for the terminations within the same group, and also for the possible changes in the number of connections of UEs in groups 1 and 2.

The blocking for group 3 UEs can be analyzed in two separate cases. The first case accounts for the number of generations needed to occupy all the channels in cell 1 ($K = M_2$), while the second case accounts for the number of generations needed to occupy all available channels in cell 2 ($K < M_2$). The first case is simpler to analyze because group 3 UEs need only generate as many connections as there are available resources on cell 1.

When $K < M_2$, group 3 UEs can only access a maximum of $K$ channels on cell 2. Therefore, extra conditions are added for the situation in which group 3 UEs are blocked when exceeding $K$ connections in cell 2. If there are less than $K$ available free channels after terminations of group 3 UEs connected to cell 2, the function $\max\{0, g_c\}$ represents the number of connections to cause blocking by generating the exact number of connections to occupy all available channels. The $\max\{0, g_c\}$ function is used to lower bound the necessary number of connections for blocking. This is because the number of connections on cell 2, in general, is not restricted to $K$ and can thus have more than $K$ current occupancies resulting in a possibly negative value for $g_c$. On the other hand, if there are more than $K$ available channels after terminations, exactly $K$ channels are used by group 3 UEs to cause blocking.

3.5.2 Channel Utilization and Total System Throughput

The overall channel utilization is

$$
U = \frac{1}{M_1 + M_2} \sum_{a,b,c,d} \left( a + b + c + d \right) \pi_{a,b,c,d},
$$

which refers to the fraction of the collective capacity that has been used by all UEs in all groups. The average total system throughput is obtained by multiplying (15) by $R$.

3.5.3 Collision Probability

Because the model uses a RACH for connection requests, it is necessary to compute the collision probability of the system. The collision probability for UEs in group $x$ is:

$$
D^{(x)} = \sum_{a,b,c,d} \sum_{k=0}^{N_x-a} \sum_{j=0}^{N_x-N_a} \sum_{l=0}^{N_x-N_a-N_b} \sum_{m=0}^{N_x-N_a-N_b-N_c} \pi_{a,b,c,d} \left( N_x - \eta \right) k \times p^k \left( 1 - p \right)^{N_x-\eta-k} w^{(a)}_{x,y},
$$

where $\eta$ is the number of connections needed to reach the end state.
where $I_i^j = 1$ when $i \geq j$ and $I_i^j = 0$, otherwise, $\eta = \{a, d\}$ for group $x = \{1, 2\}$, respectively, and $\eta = b + c$ for group $x = 3$.

### 3.6 A Special Case: Load Balancing with OSA

In this section, we illustrate the use of the proposed system to analyze other, more complex, cellular setups. As an example, we present a case where the connections can be offloaded to a neighboring cell belonging to a separate network (such as that of another provider when there is neither shared signaling between cells nor a central controller to coordinate the distribution of connections from one system to another). In the proposed system, users registered to one cell must detect the availability of free channels in the neighboring cell by means of passive channel observations, i.e., spectrum sensing. Such a system can be referred to as OSA-enabled load balancing [32].

We note that very little work has been done to consider the system-level performance of OSA-enabled cellular networks. A group of papers, e.g., [33], [34], analyzes sharing mechanisms of radio resources based on OSA, focusing on the physical characteristics, while abstracting higher layers. These papers provide insights on how OSA can be beneficial in a cellular system, but do not address the specific system performance and tradeoffs involved—specifically, the traffic characteristics of primary and secondary cells and the connection admission process during the random access phase. While many papers have analyzed the performance of OSA networks, e.g., [35], [36], the connection between OSA and its use in cellular load balancing has not been deeply addressed.

#### 3.6.1 Extensions to the System Model

The following changes need to be made to the system model.

**Spectrum sensing.** First, due to spectrum sensing assumption, each user in cell 1 is equipped with a spectrum sensor to detect whether any of the channels on cell 2 is occupied. It is assumed that $\tau_\sigma < \tau$ seconds are needed for spectrum sensing within each time slot. Channel availability detection is assumed to be imperfect. We denote $\mu_s$ as the probability of false alarm, i.e., that users in cell 1 will declare that a channel is occupied in cell 2 when in fact it is not. If users from group 3 attempt to access a channel detected as occupied by users from cell 2, the users in group 3 are blocked from connecting to that channel. The probability of misdetecting the presence of a vacant channel in cell 2 is denoted as $\mu_m$. If a misdetection occurs, it is assumed that the users from group 3 will occupy the channel at cell 2 and successfully be able to receive downlink data despite the presence of the licensed users in cell 2. We assume that as long as the users in cell 1 fulfills the minimum requirement on detection probability, $1 - \mu_m$, it does not violate QoS requirements of the users in cell 2. We also assume that the coding scheme of the users in cell 1 can be designed to compensate for the increased level of interference in the event of a misdetection. An existing connection of any user from cell 1 is preempted by users in cell 2 that accesses the same channel on cell 2. Once user from group 3 is preempted its connection is dropped.

**Connection termination.** Cell 2 users receive downlink data from the BS using channels assigned to cell 2 until the earlier of: 1) the termination of the connection after successful transmission of all data, or 2) preemption by users in cell 2 (if cell 2 users occupy the same channel). For the sake of brevity, we assume that perfect channel conditions are experienced in the system, i.e., $w_q^{(i)}, i \in \{a, d\}$. Due to spectrum sensing phase, the average packet length is now $1/q = r_p/(R(\tau - \tau_s))$.

**Load balancing scheme.** The same load balancing process is used, $DL_{K2}$; however, we assume $K = M_2$ throughout the analysis of OSA system, as control over access to channels used for load balancing is governed by the spectrum sensing quality, not by size of the channel pool accessible to cell 1. For the sake of brevity, we assume that there are no users in the group 1 and 2. Instead, we assume that users in cell 2 occupy any channel on cell 2, with a geometrically distributed average slot occupancy probability of $q_\sigma$. As cell 2 does not take part in load balancing process this choice is well motivated.

#### 3.6.2 Analytical Model

As in case of the general load balancing system, we construct a Markov chain to derive performance metrics. We reuse the notation from the general load balancing system. Let a state of a Markov system be given as $\{Y^{(1)}, Y^{(2)}, C\}$, where $Y^{(1)}$ denotes the number of users from group 3 connected to cell 1, $Y^{(2)}$ denotes the number of users from group 3 connected to cell 2, and $C$ denotes the number of users of cell 2 connected to cell 2. We denote the steady-state probabilities as $\pi_{b,c,d} = Pr\{Y^{(1)} = b, Y^{(2)} = c, C = d\}$. We also define the state transition probability $\pi_{b_1,c_1,d_1}^{(b,c,d)} = Pr\{Y^{(1)}_t = b_1, Y^{(2)}_t = c_1, C = d_1|Y^{(1)}_{t-1} = b_1, Y^{(2)}_{t-1} = c_1, C_{t-1} = d_1\}$, where subscripts $t$ and $t - 1$ denote the current and the previous time slots, respectively. The transition probabilities allow for the computation of the transition probability matrix, which is subsequently used to compute the steady-state distribution, which is denoted by $\pi(b, c, d)$.

The transition probabilities are governed by arrangement probability, defined as (6), termination probability, defined as (9), and preemption probability defined as [35, Section IV-C1]

$$P(x, y) = \left(\begin{array}{c} x + i \\ i \end{array}\right) \left(\begin{array}{c} M_2 - x - i \\ y - i \end{array}\right) q_\sigma^{(1 - q_\sigma)^{M_2-y}},$$

where $x$ is the current number of group 3 users connections on cell 2, $y$ is the current number of connections occupied by users of cell 2, and $i$ is the number of incoming group 3 users connection generations.

Having these three equations, we can derive the set of equations that describe the transition probability matrix for the general solution. They are presented in the Appendix C, available in the online supplemental material.

#### 3.6.3 Performance Metrics

We can define two important performance metrics for the model.

**Group 3 Users Throughput.** The Group 3 users throughput is defined as

$$Z = \frac{\tau - \tau_\sigma}{\tau} \sum_{b,c,d} (b + c)\pi(b, c, d).$$
Collision probability between Cell 1 and cell 2 users. A collision occurs when users in group 3 occupy the same channel as users in cell 2 after misdetecting their presence within a time slot. The probability that there is a misdetected channel is given by $x_q = q_d m_m$. The probability that there are $y$ collisions in a time slot is denoted as $\kappa_y$ and is described as follows: If $y \geq b$

$$\kappa_y = \sum_{a,b,c} \pi(a, b, c) \sum_{m=y}^{M_2} \left( \frac{M_2}{m} \right) x_q^m (1 - x_q)^{M_2-m}$$

$$\times \left( \sum_{a=0}^{b} \sum_{l=0}^{M_2-m} t_a^l T_b^{l+1} S_y^{b=m+l} \left( \frac{m}{y} \right) \left( \frac{M_2-m}{y+r} \right) \right)$$

$$+ \left( \frac{0}{y} \right) \sum_{a=0}^{b} \sum_{l=0}^{M_2-m} \sum_{r=M_2-b+M_1-a+1}^{M_2} T_a^r T_b^{l+1} S_y^{b=m+l} \left( \frac{m}{y} \right) \left( \frac{M_2-m}{y+r} \right) \right).$$

(19)

If the number of collisions is greater than or equal to the current number of occupancies of group 3 users on cell 2, the number of group 3 users connection generations on cell 2 are increased to meet the number of collisions. When the number of group 3 users connection generations on cell 2 exceeds the desired number of collisions, a hypergeometric term is used to calculate the exact probability of having $y$ collisions. When $y = m$ an excess term is needed for additional group 3 users connection generations that cannot be accommodated. On the other hand, if $y < b$

$$\kappa_y = \sum_{a,b,c} \pi(a, b, c) \sum_{m=y}^{M_2} \frac{M_2}{m} x_q^m (1 - x_q)^{M_2-m} \sum_{a=0}^{b} \sum_{l=0}^{M_2-m}$$

$$\times \sum_{r=\max(y+b+1)}^{M_2-b+1} t_a^r T_b^{l+1} S_y^{b=m+l} \left( \frac{m}{y} \right) \left( \frac{M_2-m}{y+r} \right) \right)$$

$$+ \left( \frac{0}{y} \right) \sum_{a=0}^{b} \sum_{l=0}^{M_2-m} \sum_{r=M_2-b+M_1-a+1}^{M_2} T_a^r T_b^{l+1} S_y^{b=m+l} \left( \frac{m}{y} \right) \left( \frac{M_2-m}{y+r} \right) \right).$$

(20)

If the number of collisions is strictly less than the number of occupancies on cell 2, then additional group 3 users connection generations are unnecessary. However, if additional group 3 users terminations occur, group 3 users generations are needed to ensure at least an equal number of collisions. In turn, the average number of collisions is defined as $\kappa_a = \sum_{y=0}^{M_2} y \kappa_y$.

4 RESULTS

Since our model incorporates a very large number of parameters, in the interest of clarity and brevity we focus our study on certain scenarios that are the most important in the context of our model. First, we present results that demonstrate the impact of a varying channel quality on the load balancing efficiency. Second, we examine the influence of the random access phase on load balancing efficiency. Third, we provide insight on the optimal channel sharing policy between BS 1 and BS 2. Fourth, we present results on load balancing in the OSA context in Section 4.4. Finally, in Section 4.5, we present insight on how to extend our analytical model to a multicell scenario.

To confirm the correctness of the analytical model, we created a simulation environment in Matlab for verifying the analytical results. The results in Sections 4.1, 4.2, 4.4, and 4.5 obtained using both the analytical and simulation approaches to confirm correctness, while those in Section 4.3 are obtained using simulation. Finally, all code used to generate numerical and simulation results in this paper is available upon request or via http://arxiv.org/abs/1004.0534.

Note, that our model is related to [2], [7], [18]. However, the exact comparison of our model with [2], [7], [18] is impossible, due to the following differences: 1) our model and those of [2], [7], [18] consider a system, where users are uniquely identified and treated as a single group, respectively; 2) our model, in contrary to [2], [7], [18] considers a more involved connection allocation process where, in addition to channel allocation (considered in [2], [7], [18] only), connection admission through a RACH is analyzed; 3) model of [7], [18] considers queuing, while our model does not (for tractability reasons).

4.1 Impact of Channel Quality on Load Balancing Process

In this simulation, we model, among others, the call admission, termination, and load balancing processes exactly as described in our system model. As an example, we consider a scenario in which two identical cells are positioned as described in our system model. As an example, we consider a scenario in which two identical cells are positioned as described in our system model. As an example, we consider a scenario in which two identical cells are positioned as described in our system model. As an example, we consider a scenario in which two identical cells are positioned as described in our system model.
Fig. 1. Impact of $K$ on channel utilization: (a) The channel utilization as a function of traffic intensity $p$. Two extremes of shared channels, i.e., $K = 0$ (no load balancing) and $K = M_2$ (all of cell 2’s channels used in load balancing) are shown. Furthermore, two network setups are considered: (a) Setup 1: $L_x = N_x = 6$, $M_1 = M_2 = 3$, and (b) Setup 2: $L_x = N_x = 12$, $M_1 = M_2 = 6$, where $x \in \{1, 2, 3\}$; the other parameters of the network are common for both cases and described in Section 4.1. The figure shows good agreement between the results from the analytical model and from simulation. (b) The channel utilization, represented as a percentage on the vertical axis, as a function of the channel quality between group 3 UEs and BS 1, $w_{3,1}$ for two extreme values of $K$. As $w_{3,1}$ improves, more group 3 UEs generate successful connections to cell 1 resulting in more UEs that connect to BS 1 and consequently are offloaded onto cell 2, resulting in an overall increase in channel utilization.

utilization as a function of increasing traffic intensity $p_1 = p_2 = p_3 = p$ for two extreme $K$ values, i.e., $K = 0$ (when no load balancing is used) and $K = M_2$ (all of cell 2’s channels may be used for load balancing) considering two network setups: setup 1 with $L_x = N_x = 6$, $M_1 = M_2 = 3$, and setup 2 with $L_x = N_x = 12$, $M_1 = M_2 = 6$. As expected, the channel utilization increases with more traffic intensity because an increase in $p$ results in more frequent connection requests from UEs in all groups leading to a higher probability of successful connections. Moreover, the larger the number of channels and users in both cells, the larger the channel utilization—for both cases considering $K = 0$ and $K = M_2$. The increase in channel utilization tails off as the system reaches saturation, i.e., close to 100 percent channel utilization. Similarly, an increase in $K$ results in a higher channel utilization as more UEs from group 3, that are blocked from cell 1, are offloaded onto cell 2, where they have access to an additional $K$ channels. We observe that there is a decreasing rate of gain in channel utilization with an increase in the number of shared channels $K$. As Fig. 1a shows a traffic intensity of $p = 0.4$ is the point at which the system begins to operate in saturation, i.e., the relative difference between channel utilizations for $K = M_2$ and $K = 0$ remains relatively constant thereafter for both network setups. With the knowledge of decreasing gains in channel utilization with increasing $K$, there may exist an intermediate value of $K$ that not only leads to an improvement in total channel utilization, but also maximizes improvement with respect to the overall UE experience. This value of $K$ is explored in Section 4.3. In the current section, we continue our investigation using $p = 0.4$ and explore the impacts of channel quality on performance.

Fig. 1b illustrates an increase in channel utilization with an increase in $w_{3,1}$ for two extreme values of $K$. Increasing $w_{3,1}$ results in group 3 UEs having more successful requests for receiving downlink transmissions because the average channel quality, in which requests are granted for group 3 UEs, improves. Therefore, ignoring the channel effects by assuming perfect channel conditions (also done by setting $w_{3,1} = 1$ in our model) in the analysis of load-balancing schemes, even for one particular group of UEs, produces a nontrivial difference in the channel utilization and leads to an exaggerated improvement in performance due to load balancing. By selecting a reasonable scheme to determine the channel quality, as presented in Section 3.1, we are able to provide a more realistic evaluation of the improvements of load balancing. Note that the average channel utilization significantly increases as more channels can be borrowed from BS 2. When $w_{3,1}$ increases, the difference in channel utilization between cases $K = 0$ and $K = 3$ becomes more profound. This proves that with low-channel quality system-wide improvement from load balancing might not be as significant as in the case of perfect channel conditions.

In Fig. 2, we examine $B^{(3)}$, $D^{(3)}$, $B^{(1)}$, and $B^{(2)}$, as a function of $w_{3,1}$ for all possible shared channels, i.e., $K = \{0, 1, 2, 3\}$. The scenario used in this result is identical to the one used previously. By comparing an increasing value of $K$ in Figs. 2a, 2b, 2c, and 2d, we observe the impact of an increasing number of shared channels on the considered performance benchmarks. The first interesting observation is that irrespective of the value of $K$ the blocking probability for group 1 UEs, $B^{(1)}$, is relatively constant. This means that the quality of service requirements for UEs not involved in load balancing will be met, even with load balancing enabled. Second, as $K$ increases, blocking probability for group 3 UEs, $B^{(3)}$, significantly decreases, which proves the effectiveness of load balancing in this context. Furthermore, the collision probability for UEs in group 3, $D^{(3)}$ also reduces because with more shared channels there are fewer unconnected UEs to request connections. Note, however, that the difference in collision probability for an increasing $K$ is not as significant as observed for the blocking probability because increasing the number of shared channels has a minimal effect on the performance of the random access phase. Finally, increasing $K$ only slightly increases blocking probability $B^{(2)}$ because these UEs have priority in connecting to any of cell 2’s free channels.

Focusing on Fig. 2d only, where load balancing is enabled, we note that as $w_{3,1}$ increases, all curves experience an increase. This can be explained as follows: With an
increase in $w_{3,1}^{(i)}$, more group 3 UEs are able to generate successful connections to BS 1 resulting in an increase in the contention for sub slots during admission control, and hence an increase in $D^{(3)}$. Also, there is an accompanied increase in $B^{(3)}$ because as more UEs generate successful connections, an increasing number of UEs contend for free channels on both cell 1 (where load balancing does not occur) and cell 2 (where load balancing occurs). Consequently, this also results in an increase in $B^{(1)}$ and $B^{(2)}$. Although these trends are obvious, the exact degradation in UE experience for each group is not. For example, in this specific scenario, Fig. 2 illustrates that $B^{(3)}$ is always the primary factor in the degradation of the group 3 UE experience as compared to $D^{(3)}$. This knowledge is significant as the network operator can determine whether an increase in $K$, or an increase $L_3$ will be more beneficial to group 3 UEs. Observe that Figs. 1b and 2 show an extremely good match between the analytical result and simulation.

4.2 Impact of Random Access Phase on Load Balancing Process

In this section, we present results on the effect of random access phase on the performance of load balancing. The results are presented in Fig. 3. All network parameters are
set identically to the network considered in Section 4.1, except for $L_x = 3$, where $x \in \{1, 2\}$.

We begin by investigating the impact of different UE distributions on the performance of load balancing. We perform three experiments and denote each experiment as a specific case. In the first case, we set the number of UEs, such that more UEs are distributed in groups 1 and 2, than in group 3, i.e., $N_1 = N_2 = 6$, $N_3 = 4$. In the second case, we set the number of UEs equal in each group, i.e., $N_1 = N_2 = N_3 = 6$. And finally, in the third case, we set the number of UEs in group 3 larger than in the other two groups, i.e., $N_1 = N_2 = 6$, $N_3 = 8$. The metric that is studied in the three cases described above is the total network-wide blocking probability, calculated as $\frac{1}{2} \sum_{i=1}^{3} B^{(i)}$, as a function of channel access probability $p_i = p$ for $i \in \{1, 2, 3\}$. This metric is used to give a simple overall indication of the blocking suffered by UEs in all groups. Results are presented in Fig. 3a.

The most interesting observation from Fig. 3a is that with an increase in the number of UEs in the TTR, the total blocking probability becomes smaller for moderate values of $p$. Surprisingly, the blocking probability starts to drop sharply as the value of $p$ continues to increase. This phenomenon occurs because as $p$ increases, so do collisions on the RACH, which in-turn limits the blocking probability because fewer UEs successfully access available channels. The result is easier to understand when one observes that the blocking probability is the probability of not finding a free data channel for a connection that has successfully connected to the BS via a control channel. It has to be kept in mind that for each case presented in Fig. 3a the length of the random access phase remains the same. What is important to note is that for moderate values of $p$, the difference between blocking probabilities for each case is small, i.e., less than 5 percent (please compare values of blocking probability for each case in the range of $p \in (0, 0.6)$).

However, as $p$ becomes very large, the curves with a higher number of group 3 UEs drop-off faster because they experience a substantial increase in the number of collisions. Therefore, a certain value of $p$ can maximize the channel utilization achieved through load balancing and also maintain the blocking probability at approximately the same level (given negligible changes in UE distribution).

We now move our focus to the impact of random access phase length $L_x$ on the performance of load balancing. The results are presented in Fig. 3b. The set of parameters remain the same as in the earlier experiment in this section, however $p_x = p = 0.4$. As an example, three network metrics are evaluated as a function of number of slots in the random access phase, $L_x$: 1) total channel utilization in both cells, $U$; 2) collision probability at group 3, $D^{(3)}$; and 3) blocking probability at group 3, $B^{(3)}$. For simplicity, the number of slots is set equal among each group of UEs.

Obviously, as the number of random access slots increases the collision probability decreases for group 3 UEs, and the overall channel utilization increases. However, as the collision probability decreases the blocking probability, within the same group of UEs, becomes larger. This is in line with the results presented in Fig. 3a. Recall, that as more UEs gain access to the BS, the probability that channels become unavailable increases. The results shown in Fig. 3b further demonstrate the fundamental tradeoff between the delay caused by random access and overall network utilization. With an increase in traffic intensity $p$, we expect that the graphs shown in Fig. 3b to shift upwards proportional to the increase in $p$ because we expect that a higher traffic intensity would result in more collisions and blocking for users in group 3. We demonstrate that the network operator has a powerful tool, i.e., random access phase length, through which network metrics can be easily regulated. It is obvious that the operator has no control over the channel access probability, $p$, of individual UEs. However, the operator is able to set a higher value of $L_x$ to the ASC of interest to maintain an expected access delay for each UE against a required blocking probability.

### 4.3 Impact of Varying Shared Channel Pool $K$ on Load Balancing Efficiency of $DR_K$

We consider a macrocell scenario in which the distribution of UEs in groups 1, 2, and 3 follow the relationship given by $N_1 = N_2 < N_3$ and $L_x = L_x^*$, where $x \in \{1, 2, 3\}$. Let $N_1 = N_2 = L_1 = L_2 = 25$ and $N_3 = L_3 = 40$. We consider a symmetric system, where each cell has 10 channels, i.e., $M_1 = M_2 = 10$, and the distances between each group of UEs and their respective serving BSs are identical, i.e., $d_{i1} = d_{i2} = d_{i3} = d_{i2} = 220$ m. Once again, we assume an average channel capacity of $R = 250$ kbps per channel with an average packet size $r_p = 1$ kB. A frame duration of $\tau = 1$ ms is used. We assume a simplified path loss model with identical parameters as in Section 4.1 except with $\delta = 3$, which is more appropriate for outdoor channel conditions.

In Fig. 4a, we explore the system-wide improvement in channel utilization (represented as a percentage on the vertical axis) as a function of the number of shared channels $K$ (represented on the horizontal axis). Note that for the remainder of our study we fix traffic intensity for all groups to $p = 0.2$ to determine the maximum gain in channel utilization for every value of $K$ at traffic intensities that are very near saturation. The line with circle markers represents the total percentage improvement experienced as a function of $K$, while the line with triangle markers represents the improvement experienced in channel utilization per shared channel. As $K$ increases, there is a decreasing improvement in channel utilization per additional channel, which indicates that the extra cost of sharing more channels for load balancing may outweigh the added benefit of serving a greater number of UEs.

Although there is an overall improvement in the system-wide channel utilization, the exact effect of the load-balancing scheme on the UE experience is unknown. Obviously, $B^{(3)}$ decreases with an increase in $K$ because group 3 has access to more channels. In contrast, $B^{(2)}$ increases because more UEs in group 3 access channels belonging to cell 2, which are of course also accessible to UEs in group 2. Although these general trends are obvious, the exact relationship between the amount of decrease in $B^{(3)}$ versus the amount of increase in $B^{(2)}$ is unknown. In Fig. 4b, we examine this relationship in more detail, where the decrease in $B^{(3)}$ (solid line with triangle markers) is plotted with the consequent increase in $B^{(2)}$ (solid line with circles) as a percentage on the vertical axis with increasing $K$ on the horizontal axis. We observe that in this particular scenario,
the total decrease in $B^{(3)}$ is always more than the total increase in $B^{(2)}$, suggesting that the overall UE experience improves with the proposed load-balancing scheme. This reaffirms the increase in channel utilization with an increase in $K$ for DB$_K$, which is previously observed in Fig. 4a.

The total improvement in overall UE blocking probability demonstrates the effectiveness of the load-balancing scheme. However, from a network operator viewpoint, knowledge of the changes in UE experience per additional shared channel is also very important. Fig. 4b examines the effect of an increase in $K$ on both the decrease in $B^{(3)}/K$ (dashed line with triangles), and the consequent increase in $B^{(2)}/K$ (dashed line with circles). We observe, that for this particular scenario, a decrease in $B^{(3)}/K$ is always more than the increase in $B^{(2)}/K$, which suggests that the UEs in group 3 experience more of an improvement in performance than the performance degradation experienced by UEs in group 2 per additional shared channel. This allows direct evaluation of the effectiveness of the load-balancing scheme on the overall UE experience per additional shared channel. In Fig. 4b, we note that $B^{(3)}/K$ reaches a maximum at an intermediate value of $K$, i.e., $K = 3$, because the system reaches a balance between the number of UEs requesting connections and those that are already connected through load balancing. Our model allows for the direct observation of this system state because of the combined modeling of a finite number of UEs together with a detailed call admission process. With the use of Fig. 4b, we are able to determine the best $K$ to improve the overall UE experience on a per shared channel basis, and then find the corresponding improvement in the overall channel utilization using Fig. 4a. For this particular scenario, the maximum difference between the increase in $B^{(3)}/K$ and decrease in $B^{(3)}/K$ occurs at $K = 6$ and corresponds to an overall improvement in channel utilization of 12.6 percent.

In summary, we can construct the following optimization function. Given fundamental descriptors of the network considered in Section 3, i.e., $N_i$, $M_i$, $L_i$, $p_i$, $u^{(i)}_{x,y}$, $r_p$, and $\tau$, the network operator should find

$$\arg\max_K U/K \text{ subject to } \forall i B^{(i)} \leq m^{(i)}, D^{(i)} \leq D^{(i)}.$$

where $m^{(i)}$ and $D^{(i)}$ are the required maximum blocking and collision probabilities, respectively, for group $i$. The developed analytical model provided in Section 3 allows solving the optimization function (21), since each metric, $U$, $B^{(i)}$ and $D^{(i)}$ is given in closed form. The optimization formula allows obtaining the value of $K$ to obtain the maximum utilization per shared channel, such that all considered quality of service metrics required by the operator, $m_i$ and $D_i$, are met. Note that finding the optimal solution to (21) is beyond the scope of this paper. On the other hand, note that because the complexity of the optimization problem (21) increases linearly with $K$, it can be solved efficiently via, e.g., exhaustive search, provided the existence of quick computation of $\pi_{a,b,c,d}$. Then, the complexity of calculating $\pi_{a,b,c,d}$ increases nearly exponentially with increasing $M_i$. Note, however, that calculating $\pi_{a,b,c,d}$ even for large values of state space (e.g., for $K = 3$, $M_1 = M_2 = 12$ and $N_e = 24$ size of $\pi_{a,b,c,d}$ is 16,744 elements), can be handled efficiently with any standard personal computer with optimized implementation of the transition probability matrix developed in this paper.

### 4.4 Load Balancing with OSA Results

We consider an energy detection technique to detect the presence of cell 2 users and an AWGN channel for which $\mu_f$ and $\mu_m$ is given as [39, Section III]. The sensing bandwidth is 200 kHz, with cell 2 users detected SNR of $-5$ dB and the detection threshold of $-109.4$ dBm is set to the noise floor.

#### 4.4.1 Impact of Varying $M_2$ on Throughput for Group 3 Users $Z$

To illustrate an application of the model, we consider the following parameters: The number of channels on cell 1 $M_1 = 7$, the number of subscribers in the region of overlap $N_3 = 40$, connection request probability $p = 0.1$, random access phase length $L = N_3/2 = 20$. Furthermore, we consider an average packet size of $r_p = 1$ kB, channel throughput $R = 250$ kbps, and a transmission time of $\tau - \tau_s = 9$ ms, where $\tau_s = 1$ ms. This results in an inverse of the packet length $q = 0.2813$, false-alarm probability $\mu_f = 0.1398$, and misdetection probability $\mu_m = 0.0861$.
The impact of $\tau_s$ on the collision probability of cell 1/cell 2 users is shown in Fig. 5b. All parameters remain the same as in Section 4.4.1, except that the sensing threshold equals $-109.7$ dBm, $r_p = 800$ kB, and $M_1 = 1$, $M_2 = 6$ for the purpose of better illustrating the impact of $\tau_s$. More collisions with an increase in $q_p$ are experienced because more misdetections of cell 2 occur. Furthermore, an increase in $\tau_s$ results in an accompanying decrease in $\mu_{m}$, thus, reducing $\kappa_m$ due to lower levels of perceived cell 2 user activity. In addition, increasing $\tau_s$ to reduce the number cell 1/cell 2 users collisions is more effective at high $q_p$. At high $q_p$, the system is highly sensitive to $\mu_{m}$, i.e., a small reduction in $\mu_{m}$ results in large improvements. Therefore, network operators need to account for relative levels of $q_p$ to determine if increasing $\tau_s$ will result in a considerable improvement for cell 2 users. At low $q_p$, network operators can determine the optimal $\tau_s$ to maximize $Z$ while measuring the exact improvement in $\kappa_m$.

### 4.5 Extension of the Model to a Multicell Scenario

Due to the complexity of deriving performance metrics based on an analytical model for a multicell network, we present a numerical approximation and simulation results to provide insight into the behavior of such a system. We consider a general, non-OSA, scenario in which a central cell overlaps with $X$ neighboring cells. UEs are divided into groups, in the same fashion as in the system model presented earlier. Specifically, there are three groups: 1) UEs in the central cell (referred to as group 1); 2) UEs in the neighboring cell (referred to as group 2); and 3) UEs in the overlap region between the central and neighboring cells (referred to as group 3). It is assumed that UEs in the central cell and in each of the overlapping regions are registered to the central cell. To avoid ambiguity, we assume that there are no overlaps between cells neighboring the center cell, so that the UEs in the overlap region observe signals only from two cells. This assumption makes evaluation of the multicell extension relatively easier.

Each of the groups considered consists of $N$ UEs, while every cell has $M$ available channels. Therefore, a particular UE belonging to any group in the overlap region has access to $2M$ possibly unoccupied channels, i.e., $M$ from the central cell and an additional $M$ from the neighboring cell, while the UEs from the remaining groups have access to only $M$ channels. Each channel is assumed to be error free (in other words channel conditions are not considered). Note that the decision on which users are granted connections to the central cell in the presence of load balancing is left solely to the radio network controller (as in the case of base two-cell case, refer to Section 2.7 for details). In the case of our analysis, we apply a simple decision strategy, where users are selected randomly from each cell when load balancing needs to occur. In practical scenarios, the radio network controller can apply user selection strategies based on, e.g., 1) user/cell priority; 2) fairness; and/or 3) channel quality. We present the results on two
essential metrics: 1) the overall channel utilization; and 2) the overall blocking probability.

We utilize a simple numerical approximation for overall channel utilization. We represent the two-cell system, as a special case of a multicell system, which has one central cell and $X$ neighboring cells. For simplicity, in this particular case, we assume that only group 3 UEs are present in the multicell system. Furthermore, cell one of the two-cell system represents the central cell of the multicell system, while cell two represents a linear combination of the $X$ neighboring cells in the multicell system. The number of UEs in the group residing in the region of overlap of the two-cell system is now $X_N$, the number of channels in each cell is $[(XM)/2]$, and the number of assigned random access slots is $XL$. In this manner, we keep the ratio of the number of UEs per channel approximately the same for both the two-cell and multicell systems. We present the results for three network scenarios: 1) a large-scale network; with $M = 15$ and $N = 30$; 2) a medium-scale network, with $M = 8$, $N = 16$; and 3) a small-scale network with $M = 3$, $N = 6$. For each scenario, we assume a channel access probability $p = 0.4$, average channel capacity of $R = 250$ kbps per channel with an average packet size $r_p = 0.833$ kB, and $\tau = 8$ ms slot length (the same parameters as in Section 4.1), which translates to a slot occupancy probability of $q = 0.3$. Since the traffic generated by users in each cell neighboring the central cell is the same, we keep $K = M$ for every cell. Note the unequal pool of shared channels would penalize some cells during the load balancing process and defeat the goal of maintaining an equal level of throughput for each cell.

In Fig. 6a, we observe that as the number of neighboring cells increases, so does the channel utilization. Interestingly, the channel utilization does not change considerably with an increase in UE population and channel pool. Observe that the channel utilization is well represented by our approximation, i.e., the difference between the exact model and the approximation does not exceed 10 percent for all network setups. Note that the approximation is looser for the small-scale network, in Fig. 6a for $X = \{3, 5\}$, i.e., when the ratio of UEs to channels is not an integer. This is simply due to the rounding down of the number of the channels by the floor function. Recall that the code used to generate the results for numerical approximation (and to test the accuracy of approximation for other parameter sets) is available upon request or via http://arxiv.org/abs/1004.0534.

In Fig. 6b, we present simulation results on the blocking probability as a function of the number of the neighboring cells $X$. We assume the same UE distribution as that of the medium-scale network, i.e., $M = 8$, $N = 16$, with $p = 0.3$ with $r_p = 0.625$ kB, $R = 250$ kB, $\tau = 8$ ms (the same values as considered earlier), which translates to $q = 0.4$. We observe the average blocking probability for each considered group in the scenario using three different values for the random access phase length $L = \{4, 8, 16\}$. The blocking probability for UEs in group 3 is more than that of the other groups with an increase in $X$. This is attributed to the increase in the number of UEs in the overlap region with an increase in the number of neighboring cells. The blocking probability for UEs in groups 1 and 2 stays relatively constant with changes in $X$ because UEs in group 1 have priority over UEs in group 3, while the UEs in group 2 do compete for cell resources with the UEs in the overlap region. Also, the relative difference between blocking for UEs in groups 1 and 2 for each value of $L$ is small because the ratio of accessible channels per UE remains the same regardless of $X$. As the number of random access slots increase, so does the blocking probability, which is fully consistent with the earlier observation expressed analytically and presented in Fig. 3b. This supports a conclusion that the analysis of the two-cell system is highly relevant to the behavior of the multicell network.

5 CONCLUSIONS

We have presented a new analytical model to assess the performance of the load balancing process in a two-cell system, later extended via approximations to a multicell system. The model differs in many respects from previous work on load balancing analysis in several important ways. First, it aids in determining the impact of channel quality to improve the accuracy of reported load balancing efficiency. Second, it demonstrates that the random access phase length can be used as an important network regulatory tool to control both system-wide and user experience performance metrics. Lastly, it facilitates the exploration of the effects of varying shared channel access on system-wide
performance and user experience. In addition, we have presented a new application of our model, which allows for the consideration of new load balancing techniques. In specific, we have considered OSA as an additional feature in the load balancing process. We have shown that OSA brings a large benefit to the cellular network, especially in the absence of central coordination. We have presented a variety of results derived from this framework, and in particular explored the tradeoffs in terms of channel utilization, blocking probability, and collision probability when traffic is transferred from a highly congested cell to a less-loaded neighboring cell.

ACKNOWLEDGMENTS

Part of this work was published in the Proceedings of the IEEE GlobeCom, 6-10 Dec. 2010, Miami, Florida [1].

REFERENCES

[1] S. Joshi, P. Pawełczak, J. Villasenor, S. Addepalli, and D. Ćibric, “Connection Admission versus Load Balancing,” Proc. IEEE GlobeCom, Dec. 2010.

[2] E. Ekland, “Channel Utilization and Blocking Probability in a Cellular Mobile Telephone System with Direct Retry,” IEEE Trans. Commun., vol. 34, no. 4, pp. 329-337, Apr. 1986.

[3] D. Everett and D. Manfield, “Performance Analysis of Cellular Mobile Communication Systems with Dynamic Channel Assignment,” IEEE J. Selected Areas in Commun., vol. 7, no. 8, pp. 1172-1180, Oct. 1989.

[4] O.K. Tonguz and E. Yannaz, “The Mathematical Theory of Dynamic Load Balancing in Cellular Networks,” IEEE Trans. Mobile Computing, vol. 7, no. 12, pp. 1504-1518, Dec. 2008.

[5] W. Song, W. Zhuang, and Y. Cheng, “Load Balancing for Cellular/WLAN Integrated Networks,” IEEE Network, vol. 22, no. 5, pp. 862-872, June 2004.

[6] W. Song, W. Zhuang, and Y. Cheng, “Load Balancing for Cellular/WLAN Integrated Networks,” IEEE Network, vol. 21, no. 1, pp. 27-33, Jan./Feb. 2007.

[7] E. Yannaz and O.K. Tonguz, “Dynamic Load Balancing and Sharing Performance of Integrated Wireless Networks,” IEEE J. Selected Areas in Commun., vol. 22, no. 5, pp. 862-872, June 2004.

[8] Radio Network Planning and Optimisation for UMTS, J. Laiho, A. Wacker and T. Novosad, eds. John Wiley & Sons, 2006.

[9] A.R. Miller, J.R. Vickerstaff and S. McLaughlin, “Performance Analysis of Slotted Random Access Channels for W-CDMA Systems in Nakagami Fading Channels,” IEEE Trans. Vehicular Technology, vol. 51, no. 3, pp. 411-424, May 2002.

[10] Y. Yang and T.-S. P. Yum, “Analysis of Power Ramping Scheme for UTRA-FDD Random Access Channel,” IEEE Trans. Wireless Commun., vol. 4, no. 6, pp. 2688-2693, Nov. 2005.

[11] C.-L. Lin, “Investigation of Third Generation Mobile Communication RACH Transmission,” Proc. IEEE 52nd Vehicular Technology Conf. (VTC-Fall), Sept. 2000.

[12] J. He, Z. Tang, D. Kaleshi, and A. Munro, “Simple Analytical Model for the Random Access Channel in WCDMA,” Electronics Letters, vol. 43, no. 19, pp. 1034-1036, Sept. 2007.

[13] K. Son, S. Chong, and G. de Veciana, “Dynamic Association for Load Balancing and Interference Avoidance in Multi-Cell Networks,” IEEE Trans. Wireless Commun., vol. 8, no. 7, pp. 3566-3576, July 2009.

[14] Y. Yu, R.Q. Hu, C. Bontu, and Z. Cai, “Mobile Association and Load Balancing in a Cooperative Relay Cellular Network,” IEEE Commun. Mag., vol. 49, no. 5, pp. 83-89, May 2011.

[15] H. Jiang and S.S. Rappaport, “Prioritized Channel Borrowing with Common Pool,” IEEE Trans. Commun., vol. 46, no. 6, pp. 163-172, Apr. 1996.

[16] W.S. Jeon and D.G. Jeong, “Comparison of Time Slot Allocation Strategies for CDMA/TDD System,” IEEE J. Selected Areas in Commun., vol. 18, no. 7, pp. 1271-1278, July 2000.

[17] D.Z. Deniz and N.O. Mohamed, “Performance of CAC Strategies for Multimedia Traffic in Wireless Networks,” IEEE Trans. Wireless Commun., vol. 21, no. 10, pp. 1557-1565, Dec. 2003.

[18] H. Wu, C. Qiao, S. De, and O. Tonguz, “Integrated Cellular and Ad Hoc Relaying Systems: iCAR,” IEEE J. Selected Areas in Commun., vol. 19, no. 10, pp. 2105-2115, Oct. 2001.

[19] S. Joshi, P. Pawełczak, D. Ćibric, and J. Villasenor, “When Channel Bonding Is Beneficial for Opportunistic Spectrum Access Networks,” IEEE Trans. Wireless Commun., vol. 11, no. 11, pp. 3942-3956, Nov. 2012.

[20] S. Tomasin, M. Levorato, and M. Zorzi, “Steady State Analysis of Coded Cooperative Networks with HARQ Protocol,” IEEE Trans. Commun., vol. 57, no. 8, pp. 2391-2401, Aug. 2009.

[21] C.-Y. Huang, W.-C. Chung, C.-J. Chang, and F.-C. Ren, “An Intelligent HARQ Scheme for HSDPA,” IEEE Trans. Vehicular Technology, vol. 60, no. 4, pp. 1602-1611, May 2011.

[22] A. Gotic, D. Konnakkos, and P. Constantinou, “Dynamic Subchannel and Slot Allocation for OFDMA Networks Supporting Mixed Traffic: Upper Bound and a Heuristic Algorithm,” IEEE Commun. Letters, vol. 13, no. 8, pp. 576-578, Aug. 2009.

[23] T.-S. P. Yum and W.-S. Wong, “Hot-Spot Traffic Relief in Cellular Systems,” IEEE J. Selected Areas in Commun., vol. 11, no. 6, pp. 934-940, Aug. 1993.

[24] J.-G. Choi, Y.-J. Choi, and S. Bahk, “Power-Based Admission Control for Multiclass Calls in QoS-Sensitive CDMA Networks,” IEEE Trans. Wireless Commun., vol. 6, no. 2, pp. 472-487, Feb. 2007.

[25] F.A. Cruz-Perez, J.L. Vazquez-Avila, G. Hernandez-Valdez, and L. Ortizgoza-Guerrero, “Link Quality-Aware Call Admission Strategy for Mobile Cellular Networks with Link Adaptation,” IEEE Trans. Wireless Commun., vol. 7, no. 9, pp. 4213-4223, Sept. 2008.

[26] W. Li, H. Chen, and D.P. Agrawal, “Performance Analysis of Handoff Schemes with Preemptive and Nonpreemptive Channel Borrowing in Integrated Wireless Cellular Networks,” IEEE Trans. Wireless Commun., vol. 4, no. 3, pp. 1222-1233, May 2005.

[27] W. Song and W. Zhuang, “Multi-Service Load Sharing for Resource Management in the Cellular/WLAN Integrated Network,” IEEE Trans. Wireless Commun., vol. 6, no. 2, pp. 725-735, Feb. 2007.

[28] S.-P. Chung and J.-C. Lee, “Performance Analysis and Overloaded Traffic Characterization in Multiservice Hierarchical Wireless Networks,” IEEE Trans. Wireless Commun., vol. 4, no. 3, pp. 904-918, May 2005.

[29] D. Choi, P. Monajemi, S. Kang, and J. Villasenor, “Dealing with Load Neighbors: The Benefits and Tradeoffs of Adaptive Femtocell Access,” Proc. IEEE GlobeCom, Nov./Dec. 2008.

[30] A. Goldsmith, Wireless Communications. Cambridge Univ., 2005.

[31] P. Zhou, H. Hu, H. Wang, and H.-H. Chen, “An Efficient Random Access Scheme for OFDMA Systems with Implicit Message Transmission,” IEEE Trans. Wireless Commun., vol. 7, no. 7, pp. 2790-2797, July 2008.

[32] M.M. Buddhikot, “Cognitive Radio, DSA and Self-X: Towards Next-Transformation in Cellular Networks,” Proc. IEEE Symp. New Frontiers in Dynamic Spectrum (DySPAN), Apr. 2010.

[33] Y. Ma, D.I. Kim, and Z. Wu, “Optimization of OFDMA-Based Cellular Cognitive Radio Networks,” IEEE Comm. Letters, vol. 58, no. 8, pp. 2265-2267, Aug. 2010.

[34] J. Xie, Y. Zhang, T. Skeie, and L. Xie, “Downlink Spectrum Sharing for Cognitive Radio Femtocell Networks,” IEEE Systems J., vol. 4, no. 4, pp. 524-534, Dec. 2010.

[35] J. Park, P. Pawełczak, and D. Ćibric, “Performance of Joint Spectrum Sensing and MAC Algorithms for Multichannel Opportunistic Spectrum Access Ad Hoc Networks,” IEEE Trans. Mobile Computing, vol. 10, no. 7, pp. 1011-1027, July 2011.

[36] H. Al-Mahdi, M.A. Kalil, F. Liers, and A. Mitschele-Thiel, “Increasing Spectrum Capacity for Ad Hoc Networks Using Cognitive Radios: An Analytical Model,” IEEE Comm. Letters, vol. 13, no. 9, pp. 676-678, Sept. 2009.

[37] WCDMA for UMTS—HSPA Evolution and LTE, H. Al-Mahdi, M.A. Kalil, F. Liers, and A. Mitschele-Thiel, “Increasing Spectrum Capacity for Ad Hoc Networks Using Cognitive Radios: An Analytical Model,” IEEE Comm. Letters, vol. 13, no. 9, pp. 676-678, Sept. 2009.

[38] WCDMA for UMTS—HSPA Evolution and LTE, H. Holma and A. Toskala, eds. John Wiley & Sons, 2010.

[39] C. Williamson, “Internet Traffic Measurement,” IEEE Internet Computing, vol. 5, no. 6, pp. 70-74, Nov./Dec. 2001.

[40] D. Digham, M.-S. Alouini, and M. Simon, “On the Energy Detection of Unknown Signals over Fading Channels,” Proc. IEEE Int'l Conf. Comm. (ICC), May 2003.
Przemyslaw Pawelczak received the MSc degree from Wroclaw University of Technology, Poland, in 2004 and the PhD degree from Delft University of Technology, The Netherlands, in 2009. During Fall 2007, he was a visiting scholar at the Connectivity Lab, University of California, Berkeley. From 2009 to 2011, he was a postdoctoral researcher at the Cognitive Reconfigurable Embedded Systems Lab, University of California, Los Angeles. Between 2011 and 2012, he has been a research staff member at Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany. Currently he is a (tenure track) assistant professor with the faculty of Electrical Engineering, Mathematics and Computer Science at the Delft University of Technology. His research interests include cross-layer analysis of opportunistic spectrum access networks. He received the NWO Veni grant in 2012. He was nominated for the 2011 UCLA Chancellor Award for Postdoctoral Research and received the annual Telecom Prize for Best PhD Student in Telecommunications in The Netherlands in 2008 awarded by the Dutch Royal Institute of Engineers. He is a member of the IEEE.

Shaunak Joshi received the BSc degree in electrical engineering from North Carolina State University in 2007 and the MSc and PhD degrees, both in electrical engineering, from the University of California, Los Angeles, in 2008 and 2012, respectively. He has served as a professional referee for many IEEE journals and conferences. He was also a board member of the Technical Entrepreneurship Club at the University of California, Los Angeles, in 2010. Since 2011, he has been with the Wireless Network Business Unit at Cisco Systems, Inc., San Jose, California, where he develops spectrum sensing capabilities for enterprise WiFi products. His research interests include next generation cellular and WiFi systems, channel bonding in wireless networks, network modeling, cognitive radio systems, and load balancing strategies. He is a member of the IEEE.

Sateesh Addepalli is currently the director of the Advanced Architecture, Research, and Innovation Group of Cisco Systems, Inc., San Jose, California. He has been responsible for launching and driving various successful initiatives related to mobile and wireless, sensors and sensor networks, cloud and ubiquitous/pervasive computing, data mining and analytics, internet of things, machine-to-machine, human-to-machine, health and wellness, and smart and connected communities. He currently serves on the advisory boards of the Center for Advanced Technology in Telecommunications, New York University, and Quality of Life Lab, University of Texas, Dallas, respectively. In addition, he serves on the editorial board for the Journal of Cyber Security and Mobility.

John Villasenor received the BSc degree from the University of Virginia, and the MSc and PhD degrees from Stanford University. He is a professor of electrical engineering at the University of California, Los Angeles, and a nonresident senior fellow at the Brookings Institution. His work addresses the technologies and policy issues associated with digital information. He is a senior member of the IEEE.

Danijela Ćabrić received the Dipl Ing degree from the University of Belgrade, Serbia, in 1998, the MSc degree in electrical engineering from the University of California, Los Angeles, in 2001, and the PhD degree in electrical engineering from University of California, Berkeley, in 2007, where she was a member of the Berkeley Wireless Research Center. In 2008, she joined the faculty of Electrical Engineering of University of California, Los Angeles, as an assistant professor. Her key contributions involve the novel radio architecture, signal processing, and networking techniques to implement spectrum sensing functionality in cognitive radios. He has written three book chapters and more than 25 major journal and conference papers in the fields of wireless communications and circuits and embedded systems. She was awarded the Samueli Fellowship in 2008, an Okawa Foundation research grant in 2009, and the National Science Foundation CAREER award in 2012. She is a member of the IEEE.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.