QoS-Aware Energy-Efficient Joint Radio Resource Management in Multi-RAT Heterogeneous Networks

Glaucio H. S. Carvalho, Isaac Woungang, Alagan Anpalagan, *Senior Member, IEEE*, and Ekram Hossain, *Fellow, IEEE*

***Abstract*—A heterogeneous wireless network (HetNet), which combines multiple cooperating radio access technologies in an overlapping structure, is a communication system that has been recognized as an efficient way to meet the increasing traffic demand in broadband wireless networks. In this paper, we exploit the network cooperation in HetNets to propose two joint radio resource management (JRRM) schemes that improve energy savings while satisfying the system quality-of-service (QoS) performance requirements. First, we present an optimal QoS-aware energy-efficient JRRM scheme, which is formulated as a semi-Markov decision process model, and provide the optimal control policy for the HetNet under analysis. Second, we present an implementation-friendly QoS-aware energy-efficient JRRM scheme that utilizes a threshold on the macrocell radio resource occupancy to trigger the switching-on/off procedure of the base transceiver station resources, as well as a load-balancing procedure to minimize the service disruptions that may occur because of radio resource shortage and to reduce power consumption during the HetNet operation. This JRRM scheme is analyzed by means of a multidimensional continuous-time Markov chain model. Third, we devise an algorithm to determine the threshold setting in the implementation-friendly JRRM scheme according to a desirable power saving level that is prespecified by the mobile network operator. Numerical results show that the proposed schemes achieve substantial energy savings while keeping satisfactory performance levels.**

***Index Terms*—Base transceiver station (BTS) switch-on/off procedure, energy efficiency, heterogeneous wireless networks (HetNets), joint radio resource management (JRRM), load balancing, Markov process, multiple overlapping radio access technologies (multi-RAT).**

I. Introduction

TO afford sustainable development, mobile network operators (MNOs) are encouraged to invest in the design of more eco-friendly network operations. Such a design would help MNOs successfully deal with the astronomical growth in the number of mobile cellular subscriptions—now estimated at almost 96% of the world population [1]. In this regard, recent works [2]–[6] have revealed that it is possible to achieve significant contributions in terms of energy savings by switching on or off the base transceiver stations (BTSs) according to the traffic load fluctuations. In doing so, an MNO can adopt two power saving strategies: a more aggressive approach, where the entire site is turned on/off, or a less aggressive approach, where only the power amplifiers (PAs) of the BTSs are turned on/off.

A heterogeneous wireless network (HetNet), which is realized by the cooperation of multiple overlapping radio access technologies (multi-RAT), consists of a hierarchical architecture that has been recognized as a promising alternative to meet the ever growing traffic demand in broadband wireless networks [7]. A typical HetNet may consist of a layer containing small cells such as femtocells and picocells on the top of a microcell and/or a macrocell. To holistically operate, each RAT in a HetNet has its radio resources included in a common pool of wireless channels. From this point onward, the radio resource allocation is executed for the common good of all RATs rather than the performance of an individual RAT. To this end, the Third-Generation Partnership Project (3GPP) [8] adopts a procedure named joint radio resource management (JRRM) that, by having a view of all available radio resources in a HetNet, enhances overall quality-of-service (QoS) provisioning and resource utilization.

Initially, the HetNet design was prevalently QoS driven. However, recently, it started targeting an eco-friendly operation [6], [9]. In such context, the cooperation among the multiple RATs is exploited by the JRRM scheme that employs a switching-on/off procedure on the BTS resources and dictates which RAT will be activated or deactivated. In this case, an appealing scenario for achieving high energy savings is shown in Fig. 1, in which a macrocell covers the entire targeted area, and the inner RATs are deployed to boost the system capacity [6], [9], [10]. The challenge in this scenario stems from the fact that the wireless networks have been deployed to provide the coverage independently of their current traffic load [4], [11]. As a result, the power consumption in every RAT is not necessarily a function of its traffic load.

Considering that scenario, the objective of our work is to design QoS-aware energy-efficient JRRM schemes for multi-RAT HetNets. In this respect, we propose an optimal JRRM scheme and an implementation-friendly JRRM scheme. The former optimizes the HetNet operation by specifying the operation as an optimal control problem with the objective of

Manuscript received November 15, 2013; revised March 29, 2015 and July 3, 2015; accepted September 2, 2015. Date of publication September 15, 2015; date of current version August 11, 2016. The work of G. H. S. Carvalho, I. Woungang, A. Anpalagan, and E. Hossain was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, by the National Science and Engineering Research Council of Canada (NSERC) under Grant RGPIN/293233-2011, by NSERC through Discovery Grants, and by the NSERC Strategic Project Grant STPGP 430285-12, respectively. The review of this paper was coordinated by Prof. R. Jäntti.

G. H. S. Carvalho and I. Woungang are with the Department of Computer Science, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: glauciohscarvalho@gmail.com; iwoungan@scs.ryerson.ca).

A. Anpalagan is with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: alagan@ee.ryerson.ca).

E. Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada (e-mail: Ekram.Hossain@umanitoba.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2015.2478852

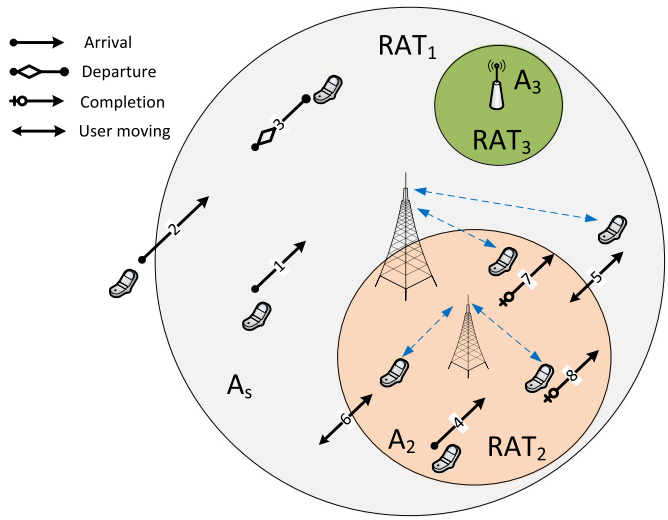

Fig. 1. HetNet with three RATs and the associated random events.

reducing the power consumption while maintaining a satisfactory system performance. The latter achieves high energy savings by efficiently triggering the switching-on/off procedure of the BTS resources based on a threshold mechanism on the macrocell radio resource occupancy. Furthermore, a load balancing procedure is proposed to minimize the service disruptions that may occur because of radio resource shortage and to reduce the power consumption during the HetNet operation. In summary, the contributions of this paper are twofold.

- We propose two QoS-aware energy-efficient JRRM schemes and present their analytical frameworks that allow us to assess the HetNet performance and quantify the power savings in every inner RAT under the coverage of the macrocell.
- Considering the implementation-friendly QoS-aware energy-efficient JRRM scheme, a systematic procedure is developed to determine the threshold setting according to a desirable power saving level that is prespecified by the MNO.

The rest of this paper is organized as follows. The related work is summarized in Section II. Section III presents the system model and assumptions. Section IV introduces the JRRM framework under analysis. Section V presents the mathematical modeling of the optimal JRRM scheme as a semi-Markov decision process (SMDP) model and the implementation-friendly JRRM scheme as a multidimensional continuous-time Markov chain (CTMC) model. In Section VI, numerical results are presented. Finally, Section VII concludes this paper.

II. Related Work and Comparison

JRRM design for HetNets has been extensively studied in the literature. In [10], Morosi *et al.* proposed two JRRM schemes, namely, the real-time traffic measurements for sleep-mode algorithm and the forecasting-based sleep-mode algorithm (FBSMA), to reduce the energy consumptions on HetNets. As far as the sleep-mode operation is concerned, these schemes achieve energy savings by turning on/off the PAs only. In [6], Ismail and Zhuang exploited the cooperation among different RATs and proposed an optimization framework that determines whether a BTS will operate in a HetNet, as well as the number of active radio channels in an operating BTS. The switching-on/off procedure was also investigated in [12] from a store-carry-and-forward relaying perspective. Unfortunately, given the mobile characteristic of the forwarding nodes, this scheme can only be applied to offload elastic services. Using the first-order analysis, Oh and Krishnamachari [13] reported that the statistics of the traffic profile and the BTS density in a flat cellular network structure are predominant factors that determine the energy savings. Because of this, the BTSs that are located in urban areas are considered the best candidates for achieving potential substantial energy savings. However, their study was not further pursued in-depth in the context of HetNets. A PA-only switching-on/off procedure in a dual-carrier Universal Mobile Telecommunications Service network was investigated in [14], and it was shown that the application of a less aggressive power saving strategy can considerably reduce the network power consumption. In [15] and [16], Bousia *et al.* investigated some strategies for achieving energy efficiency on Long-Term Evolution (LTE)-Advanced BTSs. In [15], a switching-on/off strategy is defined based solely on the distance between the BTS and its mobile users, whereas in [16], an approach that consists of sequentially testing different network layouts with on and off BTSs was proposed, assuming that mobile users are served by nearby BTSs. In [9], a Markov decision process (MDP)-based resource management scheme is proposed to study the optimal switching-on/off procedure in a HetNet with a macrocell and femtocells. A threshold-based policy was investigated for a scenario with dense femtocell deployment by means of simulation. However, contrary to our proposed schemes, the threshold mechanism was used in the femtocell radio resource occupancy rather than in the macrocell radio resource occupancy. Additionally, no load balancing procedure was used by the resource management scheme.

The application of load balancing algorithms in HetNets has attracted much attention recently. By adopting proportional fairness as a utility, Prasad *et al.*, in [17], remarkably determined that the joint optimization of load balancing and cell dormancy in an LTE-based HetNet is NP-hard. From this point onward, a set of low-complexity heuristics was proposed to efficiently handle the joint optimization problem. Cell range expansion, which is part of the 3GPP standardization efforts, is another technique used for load balancing in HetNets. By using a cell-specific offset, the work in [18] puts forward an approach to optimize the cell coverage adjustment in an LTE-based HetNet. The proposed methodology is featured by a bounding technique that approximates the solution of a system of nonlinear equations and ended up in the cell load. Numerical results illustrated that the proposed approach is able to efficiently offload the macrocell without overloading the small cells. A QoS- and load-aware user association for load balancing in HetNets was proposed in [19]. The presented method, which is formulated as a network-wide weighted utility maximization problem, consists of a nonlinear mixed-integer optimization problem. Due to the difficulties in finding the optimal solution,

particularly for a dense HetNet, a low-complexity association algorithm was designed by means of dual decomposition. Results illustrated that this method outperforms load balancing schemes that disregard QoS requirements.

Our work fundamentally differs from those in [17]–[19] in the following aspects. First, resource management approaches presented in the aforementioned papers are mostly devised to cope with the problem of maximizing the HetNet capacity and intended to run in a semistatic manner. On the other hand, our proposed JRRM schemes are designed to run in a very short time scale. In particular, the decision-making process operates at the same pace of the user arrivals and departures. Under this time scale, a load balancing algorithm could be designed to perform other functions to assist the RRM functionality. In our proposal, the load balancing scheme is designed to improve the trunking gain by redistributing the radio resources in overlapping regions under a resource exhaustion scenario. Finally, they do not pursue energy efficiency, whereas ours does.

To leverage the load balancing performance, a mechanism known as enhanced intercell interference coordination (eICIC) has been proposed by 3GPP. The principle behind eICIC is to silence the macrocell for some periods of time, which has been termed *almost blank subframes* (ABSs), over which offloaded users in small cells can transmit at reduced interference and with a higher data rate. A relevant work in this context is [20], where an algorithm is proposed to jointly address the ABS and user association problem. Due to the NP-hardness of ABS optimization, a two-step algorithm is presented to cope with the problem where, in the first step, a relaxed solution is determined, which is refined in the second step by integer rounding. Finally, Deb *et al.* inserted their proposal in the context of a self-optimized network. Results show the superiority of this method against the benchmarks. Despite the benefits of eICIC, it does not apply to our work since we assume that the macrocell is an always-on network component, and therefore, no blank subframes are transmitted.

A resource allocation framework for HetNets, considering a unified operation for admission control, handoff control, and load balancing in a self-organized HetNet, is presented in [21]. As for the system operation, when a new small cell gets self-configured, the proposed framework is invoked to shift the traffic load from the macrocell to it. After stabilizing the load between the macrocell and small cells, the process is applied to equilibrate the load among the macrocells. To perform load balancing, the load in the cells is compared, and if the difference is greater than a threshold, ongoing calls are handed off from the heaviest cell to the lightest cell. The procedure is repeated until the condition is met. No green operation is pursued in [21], and analysis is carried out by simulation.

Load balancing for machine-to-machine communications over HetNets was addressed by Osti *et al.* [22]. Taking into account knowledge of the arrival rate and the number of backlogged users in the HetNet, a performance comparison considering an optimal static policy and dynamic polices (Dyn-prob, Min-max, max-throughput, and MDP policy) is presented, and it is shown that performance gains between the static policy and the dynamic polices were only noticeable at low loads. Among the proposed schemes, the Dyn-prob policy stands out by achieving a stable and a robust performance, in addition to being able to be deployed in practice. Differences between the study in [22] and our work are described as follows. The first difference is in the type of traffic source. While we assume that traditional mobile users are originating the calls, Osti *et al.* [22] assume that machines are making the incoming requests. Second, no green operation is assumed in [22].

In [23], Sarma *et al.* investigated handoff decision making in a WiFi-WiMAX HetNet. To assist in handoff triggering, a set of RRM functionalities (bandwidth reservation, admission control, and load balancing) was engineered to minimize the cost-to-pay per bit and the power consumption while elevating QoS and quality of experience of running applications. The roles played by bandwidth reservation and admission control is to ensure the system performance while load balancing is invoked to handover mobile users from a WiFi access point to others or to a WiMAX base station. As in our work, the load balancing in [23] is applied over a very fast time scale and used to assist other RRM functionalities. However, in our work, eco-friendly operation is accomplished by turning on/off BTS resources and load balancing, whereas in [23], it is done by getting ongoing calls connected to WiFi networks only. Furthermore, we proposed analytical models based on Markov process theory, whereas the analysis in [23] is conducted entirely by simulation.

Taking the minimization of the mean delay as a criterion, the work in [24] presented a performance comparison between an optimal static and dynamic load balancing polices in HetNets. The superiority of dynamic polices, particularly the myopic dynamic policy and the modified join the shortest queue policy, over the optimal static policy became evident. The differences between our work and that in [24] lie in the following aspects. First, our works focus on polices that are QoS aware and green, whereas in [24], the emphasis is placed on the minimization of the mean delay only. Second, we focus on inelastic calls, whereas the work in [24] focuses on elastic calls. Finally, we take mobility into account, whereas the work in [24] does not.

A distributed dynamic load balancing algorithm, which is based on the K-mean++ clustering technique, is proposed in [25] to overcome congestion occurrences in an LTE-TV white space HetNet. By dynamically adjusting the cluster sizes to meet their traffic loads, the DDBL scheme ended up in a more eco-friendly operation when compared with the baseline techniques. To save energy, Aldabbagh *et al.* defined an efficient transmission mechanism in which the network overhead is optimized. In this regard, the work in [25] falls under the umbrella of *green wireless communications protocols*, whereas our work reflects the activation and deactivation BTS resources. As far as the energy efficiency is concerned, our technique leads to better results. Finally, we focus on analytical models, whereas the work in [25] focuses on simulation.

Overall, unlike previous works, which mostly rely on system-level simulations or static optimization techniques as a means for defining the system parameters and assessing the system performance measures, we propose two analytical approaches, i.e., the optimal JRRM scheme and the implementation-friendly JRRM scheme, for quantifying the system performance and energy savings. Furthermore, to the best of our knowledge, our

work is the first attempt to design a joint network selection and load balancing procedures to achieve a QoS-aware and energy-efficient wireless coverage management using the threshold mechanism and the switching-on/off process of BTS resources in a dynamic multi-RAT HetNet.

III. System Model and Assumptions

A. Network Model

We consider a HetNet composed of M RATs, where it is assumed that RAT_1 is an always-on macrocell that covers the entire targeted geographical area and overlays $M-1$ randomly distributed smaller RATs [26]. In this environment, the always-on macrocell plays a crucial role by ensuring user mobility, as well as full coverage, even after the switching-on/off process is over. This way, the green design will not jeopardize the service provision. Additionally, since the shrinking process of cell sizes becomes more and more frequent, the handoff rates will tend to grow even more. Therefore, using RAT_1 as an umbrella network is a viable solution to satisfy this fundamental design issue.

In this paper, the following assumptions are made.

A1) The ith RAT has a circular shape with radius R_i, and it covers a region A_i with area $A_{A_i} = \pi R_i^2$, whose length of perimeter is given by $L_{A_i} = 2\pi R_i$ $(1 \leq i \leq M)$. The region covered only by RAT_1, which is defined as A_s, covers an area $A_{A_s} = A_{A_1} - \sum_{i=2}^{M} A_{A_i}$. The perimeter of A_s, which is named L_{A_s}, is computed as the perimeter of the RAT_1 area added to the perimeter of all inner RATs, i.e., $L_{A_s} = L_{A_1} + \sum_{i=2}^{M} L_{A_i}$.

A2) The ith RAT $(1 \leq i \leq M)$ contributes B_i radio channels to the common pool of radio resources. This way, the total system capacity is $\sum_{i=1}^{M} B_i$ bandwidth units. The radio channels are assumed to be orthogonal so that there is no cross-RAT interference.

A3) The HetNet supports two traffic classes, namely, new calls and handoff calls. A new call is a new connection generated into the system. On the other hand, a handoff call is a connection already in progress in the system, which migrates between RATs. In HetNets, a handoff call may be categorized as horizontal handoff or vertical handoff depending on the type of RATs that are involved in the process. If the RATs are supported by the same wireless technology, then the handoff is said to be a horizontal handoff; otherwise, it is said to be a vertical handoff [27]. Moreover, mobile users, who are assumed to be uniformly distributed over the HetNet, may move freely in the network with an average speed of $E[V]$.

A4) The HetNet is a two-layer network. Thus, the JRRM scheme has only to choose between RAT_1 and an inner RAT. Moreover, the inner RATs are not deployed at the border of RAT_1. Consequently, there is no vertical handoff between an inner RAT and an external macrocell.

A5) The system is considered to be homogeneous, where all the RATs are statistically identical [26], [28]–[30], and therefore, it is sufficient to focus only on a macrocell and its inner RATs.

Given the above set of assumptions, the following random events are defined to rule the HetNet dynamics:

1) arrival of a new call in region A_s;
2) arrival of a horizontal handoff in region A_s;
3) call departure from region A_s;
4) arrival of a new call in region A_i $(2 \leq i \leq M)$;
5) user moving from region A_s to region A_i $(2 \leq i \leq M)$, and *vice versa* using RAT_1 radio resources;
6) user moving from region A_i $(2 \leq i \leq M)$ to region A_s using RAT_i radio resources;
7) call completion in region A_i $(2 \leq i \leq M)$ using the RAT_1 radio resources;
8) call completion in region A_i $(2 \leq i \leq M)$ using the RAT_i radio resources.

Fig. 1 shows a HetNet with three RATs, i.e., a macrocell RAT_1 and two smaller RATs with different coverage areas. The regions A_s and A_i, $i = 2, 3$ are also shown, along with the associated previously listed random events.

B. Model for Energy Consumption

The BTS power consumption model [6], [11] for the ith RAT $(2 \leq i \leq M)$ is given by

$$P_{ci}(h_i) = \begin{cases} \Delta_i, & \text{if } h_i = 0 \\ P_{0i} + h_i P_{rri}, & \text{if } h_i > 0. \end{cases} \tag{1}$$

In (1), h_i, $P_{ci}(h_i)$, Δ_i, P_{0i}, and P_{rri}, respectively, denote the number of ongoing mobile users using the radio resources of the ith RAT, the power consumed, the power consumed at zero load (i.e., no active radio channel), the power needed to feed the infrastructure support systems (i.e., BTS power supply and climate control, to name a few [4]), and the power component that depends on the number of active radio resources (i.e., PA, feeder loss, and transmitted power).

To achieve an eco-friendly network design, an MNO may establish two power saving strategies at the network level, namely, the deactivation of the entire BTS site or the deactivation of the PAs only. Both strategies can be implemented by appropriately setting the value of Δ_i in (1). Therefore, if the MNO adopts an aggressive power saving strategy, then Δ_i may be set to 0; otherwise, it may be set equal to P_{0i} if only the PAs are deactivated, i.e.,

$$\Delta_i = \begin{cases} 0, & \text{if the entire site is switched off} \\ P_{0i}, & \text{if only the PAs are switched off.} \end{cases} \tag{2}$$

IV. Proposed Joint Radio Resource Management Framework

A. Decision-Making Support Mechanisms

In this paper, two network mechanisms are used as a way to save energy and improve the system performance, namely, threshold and load balancing procedure. The former is exclusively used by the proposed implementation-friendly JRRM scheme, whereas the latter is used by both the optimal JRRM scheme and the implementation-friendly JRRM scheme.

1) Threshold: Let $0 \leq T \leq B_1$ be the threshold on RAT_1. The role of threshold T is to establish a simple and robust procedure for achieving energy savings. In this respect, based on the desired QoS level, threshold T may be strategically set by taking into account the offered traffic load and the expected energy savings. The proposed JRRM scheme promotes the idea that as long as the macrocell radio resource occupancy stays below the threshold value T, the BTSs or their PAs will be deactivated. Based on the time and the day of the week, two threshold settings may be of special interest. As reported in [13] and [16], the night time and weekends are the best time to save energy. Therefore, during these periods, threshold T might be set equal to B_1. In this case, all RATs under the coverage of RAT_1 will only be activated when RAT_1 gets congested. The opposite situation is also noteworthy. During the overload periods, the MNOs have to use all their resources to minimize the congestion. Under such situation, threshold T might be set 0. Thus, all RATs will be on, and the HetNet will operate with its full resources. An appealing feature of this decision-making mechanism is the fact that it only delays the start of the operation of the inner RATs; thereby, it does not degrade system performance since access to these radio resources is granted when the threshold setting is reached.

2) Load Balancing Procedure: For the implementation-friendly JRRM scheme, when the radio occupancy at RAT_1 reaches the threshold value, all RATs under its coverage are turned on. From that point on, the load balancing procedure is invoked to mitigate the service disruption situations that may occur because of radio resource shortage and to reduce the power consumption during the HetNet operation. The first step in this procedure is to compute the function f_i associated with the ith RAT $(2 \leq i \leq M)$ as follows:

$$f_i = \alpha g_t + (1-\alpha) g_e, \text{ for } 2 \leq i \leq M \tag{3}$$

where g_t is the normalized traffic load, which is obtained as

$$g_t = \frac{h_i}{B_i}, \quad 2 \leq i \leq M \tag{4}$$

and g_e is the normalized BTS total power consumption, which is obtained as

$$g_e = \frac{P_{Ti}}{\max\limits_{2 \leq j \leq M} \{P_{Tj}\}}, \quad 2 \leq i \leq M \tag{5}$$

where P_{Ti} is the total power consumption of the ith RAT $(2 \leq i \leq M)$, and α is a weighting factor. Considering (1), P_{Ti} is the maximum value of $P_{ci}(h_i)$ when the ith RAT is full.

Once f_i $(2 \leq i \leq M)$ is determined, the load balancing procedure will be invoked to perform one of the following tasks.

1) If the goal is to select a RAT to reduce its load, the JRRM scheme will search for function f^* whose value maximizes the set of $M-1$ functions, i.e.,

$$f^* = \max_{2 \leq i \leq M} \{f_i\}. \tag{6}$$

Let the ith RAT be that which maximizes (6), i.e., $f_i = f^*$. From that point onward, an ongoing call will be moved from the ith RAT to the macrocell to alleviate its traffic load and reduce its power consumption.

2) If the goal is to select a RAT to increase its load, then the RAT to be selected will be that whose function f^* minimizes the set of $M-1$ functions, i.e.,

$$f^* = \min_{2 \leq i \leq M} \{f_i\}. \tag{7}$$

Let us assume now that the ith RAT minimizes (7). In this case, an ongoing call coming from the macrocell will be moved to the ith RAT. Therefore, (7) gives the RAT whose additional call will have less impact on the overall system load and energy consumption.

Note that when multiple RATs have the same f value, then the JRRM scheme may select one among them randomly. For the optimal JRRM scheme, the application of the load balancing procedure is not associated with a particular value of the radio resource occupancy on the macrocell, but with the need to reduce the energy consumption and enhance the system performance. Thus, the load balancing procedure is executed by the optimal JRRM controller based on the incurred cost or gained reward due to its application.

B. JRRM Architecture

Fig. 2 shows the proposed JRRM framework. Each RAT in the HetNet is associated with a controller. These controllers are technology dependent. For example, they could be base station controllers for GSM Edge Radio Access Network (GERAN) systems, radio network controllers for Universal Terrestrial Radio Access Network (UTRAN) systems, multiple enhanced Node B's for LTE systems, and access service network gateways for WiMAX systems, to name a few.

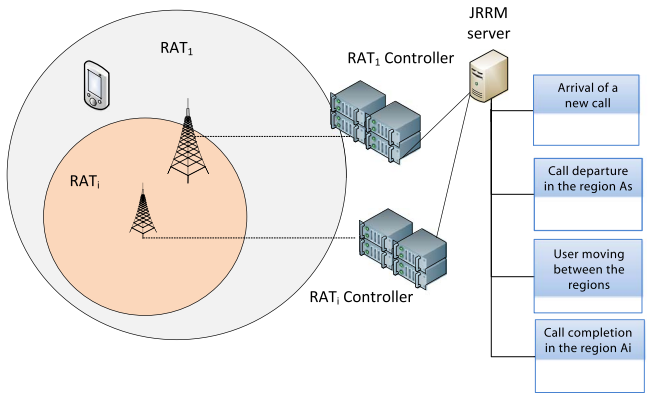

Fig. 2. JRRM framework.

From a practical perspective, the JRRM server can be implemented in any of the HetNet controllers or can alternatively be managed by a third-party entity. In this framework, the JRRM server continuously collects a set of information related to the traffic load in each RAT from the RAT controllers. These controllers always report to the JRRM server on the occurrence of any event highlighted in Fig. 1. Thus, prior to allocating any radio resource, the JRRM server combines both sets of information and invokes one of the following procedures: *arrival of a new call, call departure in region* A_s, *user moving between the regions*, and *call completion in region*

A_i. Depending on the JRRM scheme in use, the decision-making process is different. For instance, the optimal JRRM scheme applies a system control technique to decide on the best action for each of the defined random events and system load, whereas the implementation-friendly JRRM scheme employs the threshold and load balancing procedures to cope with the system dynamics.

V. Modeling

A. Traffic Model

The considered traffic model is described by means of the following parameters: dwell time, call holding time, channel holding time, and arrival patterns.

1) Dwell Time: The dwell time is defined as the amount of time the mobile user stays inside the RAT [31]. According to the work presented in [26], [28], and [29], the dwell time is assumed to be exponentially distributed, and its probability density function depends on the radius of the RAT, its area, the perimeter, and the mobile user speed. The dwell time inverse is referred to as the rate at which the mobile user leaves the RAT or makes a handoff request. Therefore, the average rate at which a mobile user moves out of region A_s is given by

$$\mu_{A_s} = \frac{E[V]L_{A_s}}{\pi A_{A_S}}. \tag{8}$$

Under the coverage of RAT_1, a mobile user in region A_s may move toward making a horizontal handoff or a vertical handoff before ending the call. Thus, that user may leave that area by crossing the boundary of RAT_1 with rate $\mu_{A_{s1}}$ or by entering the ith RAT $(2 \leq i \leq M)$ with rate $\mu_{A_{si}}$. Thus, (8) can be rewritten as

$$\mu_{A_s} = \mu_{A_{s1}} + \sum_{i=2}^{M} \mu_{A_{si}}. \tag{9}$$

To calculate the components on the right-hand side of (9), one can compute the probability of crossing the desired boundary, then multiply it by the average region A_s boundary crossing rate, i.e.,

$$\mu_{A_{s1}} = \mu_{A_s} \frac{L_{A_1}}{L_{A_s}}. \tag{10}$$

By substituting (8) in (10), we obtain

$$\mu_{A_{s1}} = \frac{E[V]L_{A_1}}{\pi A_{A_S}}. \tag{11}$$

Similarly, the average rate at which a mobile user moves from region A_s to region A_i is obtained as follows:

$$\mu_{A_{si}} = \frac{E[V]L_{A_i}}{\pi A_{A_S}}, \quad 2 \leq i \leq M. \tag{12}$$

Based on (8), the rate at which a mobile user moves out of region A_i is obtained as

$$\mu_{A_i} = \frac{E[V]L_{A_i}}{\pi A_{A_i}}. \tag{13}$$

2) Call Holding Time: The call holding time is the duration of the call. It represents the time the call takes if it experiences no forced termination; it is assumed to follow an exponential distribution with mean value $1/\mu_c$ [6], [28].

3) Channel Holding Time: The channel holding time is defined as the time elapsed from the moment a radio resource is allocated to a mobile user until the moment it is released by either ending the call or performing a handoff (vertical or horizontal). Due to the fact that the dwell time and the call duration time are both exponentially distributed random variables, the channel holding time will also follow an exponential distribution with a mean value given by the minimum of them. Thus, the channel holding time in region A_s is given by

$$\frac{1}{\mu_{A_{hs}}} = \frac{1}{\mu_{A_s} + \mu_c}. \tag{14}$$

Similarly, the channel holding time $1/\mu_{A_{hi}}$ in region A_i is given by the minimum of two events: the call completion and the vertical handoff. Therefore

$$\frac{1}{\mu_{A_{hi}}} = \frac{1}{\mu_{A_i} + \mu_c}, \quad 2 \leq i \leq M. \tag{15}$$

4) Arrival Pattern: We assumed that new calls arrive in the HetNet according to a Poisson process with mean value Λ. Furthermore, the mobile users are assumed to be uniformly spread over RAT_1. Therefore, the average arrival rates at region A_s and at the ith region are, respectively, given by

$$\lambda_{A_s} = \frac{A_{A_s}}{A_{A_1}}\Lambda \tag{16}$$

$$\lambda_{A_i} = \frac{A_{A_i}}{A_{A_1}}\Lambda, \quad 2 \leq i \leq M. \tag{17}$$

It should be noted that the Poisson process has been widely used [6], [9], [10], [32]–[34] to represent the arrival process in HetNets with RATs ranging from femtocells to macrocells. Given the system homogeneity, the horizontal-handoff call arrival rate at the macrocell level is determined by the dwell time at RAT_1. In particular, we are interested in the rate at which a mobile user moves out of RAT_1 by crossing its boundary. This rate was anticipated in (11). Thus, the rate at which l_1 mobile users on region A_s attempt a horizontal handoff is given by [35]

$$\lambda_{hh} = l_1 \mu_{A_{s1}}. \tag{18}$$

B. Optimal QoS-Aware Energy-Efficient JRRM Model

According to the SMDP framework, the definition of the optimal control problem requires the specification of the following components.

1) State Space: The state space for the SMDP-based optimal QoS-aware energy-efficient JRRM controller is defined in (19), shown at the bottom of the next page, where l_1 denotes the number of ongoing mobile users in region A_s, l_i $(2 \leq i \leq M)$ denotes the number of mobile users under the ith RAT coverage using the radio resources of RAT_1, h_i is the number of mobile users under the ith RAT coverage using its radio

resources, and $\mathbf{e}$ is a vector of size $4M$ that specifies the last occurred event. As given in (20), shown at the bottom of the page, $\mathbf{e}$ organizes the random events in six groups where the first group, i.e., $e = 0$, denotes the user moving from region A_s to region A_i $(2 \leq i \leq M)$, and *vice versa*, using the RAT_1 radio resources. The second group packs the events associated with region A_s. Thus, $e = 1$ denotes an arrival of a new call, 2 denotes an arrival of a horizontal handoff, and 3 denotes a call departure. In the third group, from 4 to $M + 2$, we have an arrival of a new call in region A_i $(2 \leq i \leq M)$. For instance, for $e = 4$, there is an arrival in region A_2, whereas for $e = M + 2$, there is an arrival in region A_M. By analogy, we can build the entire vector for the remaining events following the specifications provided in (20).

Since each RAT in the HetNet cannot support more than its capacity, the following conditions prevail: $\sum_{k=1}^{M} l_k \leq B_1$ and $h_i \leq B_i$. Furthermore, it is assumed that each state is the HetNet configuration just after an event occurrence and just before decision making. Therefore, it is infeasible to have a call completion or load balancing when there is no call in progress in a RAT. Thus, in addition to the capacity constraints, (19) presents other constraints to prevent these inconsistencies.

2) Decision Epochs and Controlling Actions: The decision epochs follow the definition of $\mathbf{e}$. They correspond to the instances in time when a new call arrives, a handoff call (vertical or horizontal) arrives, or a call releases the radio channel. Based on the decision epochs, the actions to be selected by the optimal controller can be determined. Let $A(x)$ be the set of actions available in state $x \in X$. Equation (21), shown at the bottom of the page, defines all the possible values taken by an action $a \in A(x)$. Note that there are four groups of actions. The first group stands for the acceptance in RAT_i. For instance, the action $a = 0$ means the incoming request will be connected to RAT_1, $a = 1$ means the incoming request will be connected to RAT_2, and so on. The second group refers to the handoff from RAT_i to region A_s. For instance, $a = M$ means a handoff from RAT_2 to RAT_1. The third group defines the application of the load balancing procedure in a specific region. For instance, $a = 2M - 1$ means the application of load balancing in region A_2. Finally, the action $a = 3M - 2$ means that the optimal controller will not or cannot do anything.

3) Expected Time Until the Next Decision Epoch: The expected time until the next decision epoch $\tau_x(a)$ is defined as the inverse of the sum of the rate of all constituent processes. Thus, for a state $x \in X$ and an action $a \in A(x)$, $\tau_x(a)$ is given by (22), shown at the bottom of the page.

4) Transition Probabilities: Let $p_{xy}(a)$ be the probability that in the next decision epoch, the system will be in state $y \in X$ considering that the current state is $x = (l_1, \ldots, l_i, h_i, \ldots, l_M, h_M, \mathbf{e}) \in X$ and a controlling action $a \in A(x)$ is selected. Let e^x denote the value of e in state

$$X = \Bigg\{ x = (l_1, \ldots, l_i, h_i, \ldots, l_M, h_M, \mathbf{e}) : \sum_{k=1}^{M} l_k \leq B_1; h_i \leq B_i; e = 2, 3 \text{ and } l_1 > 0; e \in \{M+3, \ldots, 2M+1\} \text{ and } h_i > 0$$
$$e \in \{2M+2, \ldots, 3M\} \text{ and } l_i > 0; e \in \{3M+1, \ldots, 4M-1\} \text{ and } h_i > 0 \Bigg\} \quad (19)$$

$$\mathbf{e} = \Bigg[\underbrace{0}_{\substack{\text{User moving from} \\ A_{s \to i} (2 \leq i \leq M) \\ \text{and vice-versa using} \\ \text{RAT}_1 \text{ radio resources}}} \Bigg| \underbrace{1\,2\,3}_{\substack{\text{Event in } A_s \\ 1\text{-Arrival of a new call} \\ 2\text{-Arrival of horiz. handoff} \\ 3\text{-Call departure}}} \Bigg| \underbrace{4, \ldots, M+2}_{\substack{\text{Arrival of new call} \\ \text{in } A_i (2 \leq i \leq M)}} \Bigg| \underbrace{M+3, \ldots, 2M+1}_{\substack{\text{User moving from} \\ A_{i \to s} (2 \leq i \leq M) \\ \text{using RAT}_i \text{ resources}}} \Bigg| \underbrace{2M+2, \ldots, 3M}_{\substack{\text{Call completion in} \\ A_i (2 \leq i \leq M) \\ \text{using RAT}_1 \text{ resources}}} \Bigg| \underbrace{3M+1, \ldots, 4M-1}_{\substack{\text{Call completion in} \\ A_i (2 \leq i \leq M) \\ \text{using RAT}_i \text{ resources}}} \Bigg]^T \quad (20)$$

$$a = \Bigg(\underbrace{0, \ldots, M-1}_{\text{Accept in RAT}_i (1 \leq i \leq M)} \Bigg| \underbrace{M, \ldots, 2M-2}_{\text{Handoff from RAT}_{i \to 1} (2 \leq i \leq M)} \Bigg| \underbrace{2M-1, \ldots, 3M-3}_{\text{Load balancing in } A_i (2 \leq i \leq M)} \Bigg| \underbrace{3M-2}_{\text{Do nothing}} \Bigg) \quad (21)$$

$$\tau_x(a) = \frac{1}{\Lambda + \lambda_{hh} + \sum_{l_1=1}^{B_1} l_1 \mu_{A_{hs}} + \sum_{i=2}^{M} \sum_{l_1=1}^{B_1} l_1 \mu_{A_{si}} + \sum_{i=2}^{M} \sum_{l_i=2}^{B_1} l_i \mu_{A_i} + \sum_{i=2}^{M} \sum_{h_i=1}^{B_i} h_i \mu_{A_i} + \sum_{i=2}^{M} \sum_{l_i=1}^{B_1} l_i \mu_c + \sum_{i=2}^{M} \sum_{h_i=1}^{B_i} h_i \mu_c} \quad (22)$$

$x \in X$. For all feasible $x, y \in X$, the following cases for $p_{xy}(a)$ can occur.

1) User moving from region A_s to A_i $(2 \leq i \leq M)$, and *vice versa*, using the RAT_1 radio resources $(e^x = 0)$. To reduce the handoff rates throughout the system, which may become critical when the number of inner RATs grows, the optimal controller will keep the mobile user linked with RAT_1, then the action $a = 3M - 2$ will always be selected. Therefore, we have the following state transitions: 1) With probability $l_1 \mu_{A_{si}} \tau_x(a)$, the system will transit to $y = (l_1 - 1, \ldots, l_i + 1, h_i, \ldots, l_M, h_M, \mathbf{e})$. 2) With probability $l_i \mu_{A_i} \tau_x(a)$, it goes to $y = (l_1 + 1, \ldots, l_i - 1, h_i, \ldots, l_M, h_M, \mathbf{e})$.
2) Arrival of a new call in region A_s $(e^x = 1)$ with probability $\lambda_{A_s} \tau_x(a)$. In this case, there are three possible actions: 1) Accept in RAT_1, where $a = 0$ is selected under the condition that $1 + \sum_{k=1}^{M} l_k \leq B_1$. In this case, $y = (l_1 + 1, \ldots, l_i, h_i, \ldots, l_M, h_M, \mathbf{e})$. 2) Perform the load balancing procedure when $1 + \sum_{k=1}^{M} l_k > B_1$, where $a \in \{2M - 1, \ldots, 3M - 3\}$ is taken considering the RATs that satisfy the conditions $l_i > 0$, $1 + h_i \leq B_i$ $(2 \leq i \leq M)$. In this case, $y = (l_1 + 1, \ldots, l_i - 1, h_i + 1, \ldots, l_M, h_M, \mathbf{e})$. 3) When no RAT meets the previous requirements, $a = 3M - 2$, and $y = x$.
3) Arrival of a horizontal handoff in region A_s $(e^x = 2)$ with probability $\lambda_{hh} \tau_x(a)$. In this case, the optimal controller defines the same rules and conditions used for $e^x = 1$.
4) Call departure in region A_s $(e^x = 3)$ with probability $l_1 \mu_{A_{hs}} \tau_x(a)$. In this case, there are two possible actions: 1) Perform the load balancing procedure in the RATs that satisfy the condition $h_i > 0$ $(2 \leq i \leq M)$. In this case, $a \in \{2M - 1, \ldots, 3M - 3\}$ and $y = (l_1 - 1, \ldots, l_i + 1, h_i - 1, \ldots, l_M, h_M, \mathbf{e})$. 2) The optimal controller may choose not to perform the load balancing and select the action $a = 3M - 2$ that leads to $y = (l_1 - 1, \ldots, l_i, h_i, \ldots, l_M, h_M, \mathbf{e})$.
5) Arrival of a new call in region A_i $(2 \leq i \leq M)$, $(e^x \in \{4, \ldots, M + 2\})$, with probability $\lambda_{A_i} \tau_x(a)$. In this case, we have the following scenarios: 1) If $h_i = 0$ and $1 + \sum_{k=1}^{M} l_k \leq B_1$, then the optimal controller may pick either $a = 0$ or $a \in \{1, \ldots, M - 1\}$. In this case, $y = (l_1, \ldots, l_i + 1, h_i, \ldots, l_M, h_M, \mathbf{e})$ or $y = (l_1, \ldots, l_i, h_i + 1, \ldots, l_M, h_M, \mathbf{e})$. Otherwise, if $1 + \sum_{k=1}^{M} l_k > B_1$, then the only available action is to accept in the ith RAT; therefore, $a \in \{1, \ldots, M - 1\}$, and $y = (l_1, \ldots, l_i, h_i + 1, \ldots, l_M, h_M, \mathbf{e})$. 2) If $h_i > 0$, $1 + \sum_{k=1}^{M} l_k \leq B_1$, and $1 + h_i \leq B_i$; hence, it is possible to select both RATs as recipients of the incoming call. However, if only RAT_1 (respectively, RAT_i) has room to admit the incoming call, then action $a = 0$ (respectively, $a \in \{1, \ldots, M - 1\}$) will be selected, which will increase by one unit the load in the corresponding RAT. 3) When $1 + \sum_{k=1}^{M} l_k > B_1$, $1 + h_i > B_i$, but $1 + h_j \leq B_j$ $(2 \leq j \leq M)$ for $j \neq i$, then the optimal controller may choose to perform the load balancing procedure considering the jth RAT by selecting an action $a \in \{2M - 1, \ldots, 3M - 3\}$. In this case, $y = (l_1, \ldots, l_i + 1, h_i, l_j - 1, h_j + 1, \ldots, l_M, h_M, \mathbf{e})$. 4) Finally, when all the previous conditions are not fulfilled, then $a = 3M - 2$, and $y = x$.
6) User moving from region A_i $(2 \leq i \leq M)$ to A_s using the RAT_i radio resources, $(e^x \in \{M + 3, \ldots, 2M + 1\})$, with probability $h_i \mu_{A_i} \tau_x(a)$. The following scenarios are supported by the optimal controller: 1) If $1 + \sum_{k=1}^{M} l_k \leq B_1$, then the optimal controller will perform the vertical handoff by selecting $a \in \{M, \ldots, 2M - 2\}$ that will result in $y = (l_1 + 1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M, \mathbf{e})$. 2) When RAT_1 is full, then the load balancing procedure will be invoked. In such a situation, if $l_i > 0$, $l_j > 0$, and $1 + h_j \leq B_j$ $(2 \leq j \leq M)$ for $j \neq i$, then the optimal controller will select an action $a \in \{2M - 1, \ldots, 3M - 3\}$. If the ith RAT is chosen, then $y = (l_1 + 1, \ldots, l_i - 1, h_i, \ldots, l_M, h_M, \mathbf{e})$. Otherwise, we will have $y = (l_1 + 1, \ldots, l_i, h_i - 1, l_j - 1, h_j + 1, \ldots, l_M, h_M, \mathbf{e})$. 3) The action $a = 3M - 2$ will be selected when no RAT meets the previous requirements and $y = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M, \mathbf{e})$, which means that the handoff will be forced to terminate.
7) Call completion in region A_i $(2 \leq i \leq M)$ using the RAT_1 radio resources, $(e^x \in \{2M + 2, \ldots, 3M\})$, with probability $l_i \mu_c \tau_x(a)$. In this case, we have the following: 1) When the call leaves RAT_1, the optimal controller may perform load balancing or do nothing. Thus, if $l_i > 0$, $h_i > 0$, and $h_j > 0$ $(2 \leq j \leq M)$ for $j \neq i$, then the optimal controller may choose the ith or the jth RAT by selecting the action $a \in \{2M - 1, \ldots, 3M - 3\}$, which will lead to $y = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M, \mathbf{e})$ or $y = (l_1, \ldots, l_i - 1, h_i, l_j + 1, h_j - 1, \ldots, l_M, h_M, \mathbf{e})$, respectively. Alternatively, the optimal controller may choose to do nothing and select the action $a = 3M - 2$ that will result in $y = (l_1, \ldots, l_i - 1, h_i, \ldots, l_M, h_M, \mathbf{e})$. It should be noticed that if there is no call in progress in the ith RAT, then it is not available to participate in the load balancing procedure. 2) When no call meets the above requirement, the only feasible action will be $a = 3M - 2$.
8) Call completion in region A_i $(2 \leq i \leq M)$, $(e^x \in \{3M + 1, \ldots, 4M - 1\})$, using the RAT_i radio resources with probability $h_i \mu_c \tau_x(a)$. For this case, we have the following: 1) Similarly to the last case, the optimal controller may choose to perform a load balancing procedure $(a \in \{2M - 1, \ldots, 3M - 3\})$ or do nothing (i.e., $a = 3M - 2$). Therefore, if $h_i > 0$, $l_i > 0$, and $h_j > 0$ $(2 \leq j \leq M)$ for $j \neq i$, then $y = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M, \mathbf{e})$ if the ith RAT is chosen or $y = (l_1, \ldots, l_i - 1, h_i, l_j + 1, h_j - 1, \ldots, l_M, h_M, \mathbf{e})$, otherwise. Optionally, the action $a = 3M - 2$ may be selected, which will lead the system to $y = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M, \mathbf{e})$. 2) If $l_i = 0$ or $h_j = 0$, then the optimal controller will select action $a = 3M - 2$.
9) For any other setting, $p_{xy}(a) = 0$.

5) Cost: Based on the action $a \in A(x)$ taken and the system state $x \in X$, a cost $C_x(a)$ is incurred. Equation (3) defines the cost structure used to regulate the behavior of the optimal

controller. Thus, whenever an action is taken to increase the load and the power consumption in an inner RAT, it leads to a cost f_i. Conversely, when it decreases the load and the power consumption in an inner RAT, a reward (i.e., negative cost) is gained, which is given by $-f_i$. The cost structure $C_x(a)$ is completely defined in (23), shown at the bottom of the page.

6) Optimization Problem and Value Iteration Algorithm: Let ζ denote a stationary policy and $\psi_x(\zeta)$ be its average cost. Let $Z(t)$ be the total cost incurrent up to time t, where $t \geq 0$. Denote $E_{x,\zeta}$ as the expectation operator when the initial state $x_0 = x \in X$ and the policy ζ is used. Then, the limit, i.e.,

$$\psi_x(\zeta) = \lim_{t\to\infty} \frac{1}{t} E_{x,\zeta}\left[Z(t)\right]$$

exists for all $x \in X$ [36]. Our optimization problem then is to minimize $\psi_x(\zeta)$ among all policies, i.e., to determine $\psi^* \leq \psi_x(\zeta)$ for all $x \in X$, which is the minimal average cost whose optimal policy is ζ^*.

In this paper, the value iteration algorithm is applied to derive the optimal policy. The principle behind this method is to approximate the minimal average cost through a sequence of value functions $V_n(x)$ for all $x \in X$. The value functions provide lower and upper bounds on the minimal average cost, which iteratively converge to the minimal average cost. The value iteration algorithm is specified as follows [36].

Step 0: Choose $V_0(x)$ such that $0 \leq V_0(x) \leq \min_a\{C_x(a)/\tau_x(a)\}$ for all $x \in X$. Choose a number τ with $0 < \tau < \min_{x,a} \tau_x(a)$. Let $n := 1$.

Step 1: Compute the recursive function $V_n(x), x \in X$, from

$$V_n(x) = \min_{a\in A(x)} \left[\frac{C_x(a)}{\tau_x(a)} + \frac{\tau}{\tau_x(a)} \sum_{y\in X} p_{xy}(a) V_{n-1}(y) + \left(1 - \frac{\tau}{\tau_x(a)}\right) V_{n-1}(x) \right].$$

Let $\zeta(n)$ be a stationary policy whose actions minimize the right-hand side of the recursive function.

Step 2: Compute the bounds

$$m_n = \min_{y\in X} \{V_n(y) - V_{n-1}(y)\} \text{ and}$$
$$M_n = \max_{y\in X} \{V_n(y) - V_{n-1}(y)\}.$$

The algorithm is stopped with policy $\zeta(n)$ when $0 \leq (M_n - m_n)/m_n \leq \varepsilon$, where ε is a prespecified accuracy number. In this paper, $\varepsilon = 10^{-10}$. Otherwise, go to Step 3.

Step 3: $n := n + 1$ and go to Step 1.

After a finite number of iterations, the algorithm terminates and outputs a policy $\zeta(n)$ whose average cost function $\psi_x(\zeta(n))$ satisfies $0 \leq (\psi_x(\zeta(n)) - \psi^*)/\psi^* \leq \varepsilon$ for all $x \in X$.

The optimal policy ζ^* is a decision rule $f : X \to A$ that dictates the action $f(x) \in A(x)$ each time the system is observed in the state $x \in X$ [36]. Under ζ^*, the underlying CTMC model is solved. To this end, its infinitesimal generator matrix $\boldsymbol{Q}$ is built following the specifications of the optimal policy. From that point on, taking into account the normalization condition $\sum_{s\in S} \pi(s) = 1$, one can compute the steady-state probability vector $\boldsymbol{\pi}$ by solving the system of linear equations $\boldsymbol{\pi Q} = 0$ using standard numerical techniques. In this paper, we have used the successive over-relaxation (SOR) method [37].

7) Performance Metrics: The blocking probability of new calls in region A_s ($P_{BP_{A_s}}$), the blocking probability of new calls in region A_i ($P_{BP_{A_i}}$), the forced termination probability of horizontal-handoff calls in region A_s ($P_{FT_{A_s}}$), and the forced termination probability of vertical-handoff calls in region A_i ($P_{FT_{A_i}}$) are computed considering that the occurrence of these events happens when action $a = 3M - 2$ is taken for the value of e^x corresponding to the event of interest (**EoI**). For instance, let $P_{\mathbf{EoI}}(a = 3M - 2)$ in (24), shown below, be a probability of the **EoI** when action $a = 3M - 2$ is chosen. Then, if $e^x = \mathbf{EoI} = 1$, we have the blocking probability of new calls in region A_s (i.e., $P_{\mathbf{EoI}}(a = 3M - 2) = P_{BP_{A_s}}$). On the other hand, if $e^x = \mathbf{EoI} \in \{M + 3, \ldots, 2M + 1\}$, then we have the forced termination probability in the ith region (i.e., $P_{\mathbf{EoI}}(a = 3M - 2) = P_{FT_{A_i}}$), and so on. Thus

$$P_{\mathbf{EoI}}(a = 3M - 2) = \sum_{x\in X : e^x = \mathbf{EoI}} \pi(x). \tag{24}$$

$$C_x(a) = \begin{cases} 0 & \forall\, e^x & a = 3M-2 \\ 0 & e^x = 1, 2, \{4, \ldots, M+2\} & a = 0 \\ f_i & e^x = 1, 2 & a \in \{2M-1, \ldots, 3M-3\} \\ -f_i & e^x = 3 & a \in \{2M-1, \ldots, 3M-3\} \\ f_i & e^x \in \{4, \ldots, M+2\} & a \in \{1, \ldots, M-1\} \\ f_i & e^x = \in \{4, \ldots, M+2\} & a \in \{2M-1, \ldots, 3M-3\} \\ 0 & e^x \in \{M+3, \ldots, 2M+1\} & a \in \{M, \ldots, 2M-2\} \\ f_i & e^x \in \{M+3, \ldots, 2M+1\} & a \in \{2M-1, \ldots, 3M-3\} \\ -f_i & e^x \in \{2M+2, \ldots, 3M\} & a \in \{2M-1, \ldots, 3M-3\} \\ -f_i & e^x \in \{3M+1, \ldots, 4M-1\} & a \in \{2M-1, \ldots, 3M-3\} \end{cases} \tag{23}$$

Based on the assumption that the mobile users are uniformly spread over the region covered by RAT_1, the average blocking probability P_{ABP} of a new call is computed as follows:

$$P_{ABP} = \frac{\lambda_{A_s} P_{bp_{A_s}} + \sum_{i=2}^{M} \lambda_{A_i} P_{bp_{A_i}}}{\Lambda}. \quad (25)$$

The bandwidth utilization is computed as the ratio of the number of busy radio resources to the total available bandwidth [35]. Let U_{RAT_1} and $U_{\mathrm{RAT}_{2\leq i\leq M}}$ denote the bandwidth utilization of RAT_1 and the ith RAT, respectively. With $\pi(s)$, these can be calculated using (26) and (27), shown at the bottom of the page, respectively.

According to the power consumption model presented in Section III-B, the average power consumption in the ith RAT $(P_{C_{\mathrm{RAT}_i}})$ is calculated following (28), shown at the bottom of the page. With $P_{C_{\mathrm{RAT}_i}}$, we can easily compute the the average power savings in the ith RAT $(2 \leq i \leq M)$, which is expressed as

$$P_{S_i} = 100 \times \left(\frac{P_{Ti} - P_{C_{\mathrm{RAT}_i}}}{P_{Ti}} \right), \text{ for all } 2 \leq i \leq M. \quad (29)$$

C. Implementation-Friendly QoS-Aware Energy-Efficient JRRM Model

In practice, the optimal policy may be calculated offline, and the state–action pair could be stored in the JRRM server as a lookup table, so that every time a system state is visited, the optimal action is selected by the JRRM controller. While this procedure is quite feasible, it may become challenging to be applied in large-scale wireless networks such as a DenseNet—the next generation of HetNets—since it will entail a system with a large number of inner RATs. In such scenario, issues related to the lack of structure of the optimal policy [38] and the size of the lookup table [39] might become critical; therefore, the design of effective suboptimal policies would be required. From that perspective, we have proposed an implementation-friendly JRRM scheme. For the random events presented in Fig. 2, this JRRM scheme performs based on the CTMC model. This model is described in the sequel.

1) System State and State Space: Let S be a finite set of system states and $s \in S$. The state of the multidimensional CTMC model representing the HetNet is defined as $s = (l_1, \ldots, l_i, h_i, \ldots, l_M, h_M)$, where l_i $(1 \leq i \leq M)$ and h_i $(2 \leq i \leq M)$ were already introduced in (19). Let $\phi \in S$ be the set of system states, where $\sum_{k=1}^{M} l_k < T$, and $h_i > 0$. These states are labeled infeasible because only RAT_1 is active when its radio resource occupancy is below the threshold value T. Thus, along with the capacity constraints presented in (19), the state space containing all feasible states is formally defined as

$$S = \left\{ s : \sum_{k=1}^{M} l_k \leq B_1; h_i \leq B_i; \sum_{k=1}^{M} l_k < T \text{ and } h_i = 0 \right\}. \quad (30)$$

2) Arrival of a New Call in Region A_s: In this region, an arrival of a new call can only be handled by RAT_1 because it lies solely under its coverage. Therefore, with rate λ_{A_s}, the system leaves the current state s toward state $s' = (l_1 + 1, \ldots, l_i, h_i, \ldots, l_M, h_M)$ if $\sum_{k=1}^{M} l_k < B_1$. On the other hand, when RAT_1 is full, and there are mobile users using its radio resources in the regions covered by noncongested RATs, the proposed JRRM scheme performs load balancing using (7) and accepts the incoming call on RAT_1. Thus, if $f_i = f^*$, the system will evolve from the current state s to state $s' = (l_1 + 1, \ldots, l_i - 1, h_i + 1, \ldots, l_M, h_M)$ with rate λ_{A_s}. In this case, if more than one RAT has the same f^* value, then the JRRM scheme may choose any of these RATs. Furthermore, if only one RAT is available and satisfies the criterion: $l_i > 0$ and $h_i < B_i$ $(2 \leq i \leq M)$, then this RAT will be chosen to perform load balancing. Let $s_{bp} \in S$ be a subset of states satisfying the following conditions:

$$s_{bp} = \left\{ s \in S | \sum_{k=1}^{M} l_k = B_1 \text{ and } l_i = 0 \text{ or } h_i = B_i \right\}. \quad (31)$$

The arrival of a new call in any state $s \in s_{bp}$ will be blocked since there is no radio resource to carry it on RAT_1, and there is no possibility to apply the load balancing.

3) Arrival of a Horizontal-Handoff Call in Region A_s: An arrival of a horizontal-handoff call in RAT_1 is processed as an arrival of a new call in region A_s. This way, assuming that the conditions described in Section V-C2 hold, if $l_1 > 0$, the system will transit from state s to state s', as previously described, but

$$U_{\mathrm{RAT}_1} = \frac{1}{B_1} \left(\sum_{l_1=1}^{B_1} \cdots \sum_{l_i=1}^{B_1} \sum_{h_i=0}^{B_i} \cdots \sum_{l_M=1}^{B_1} \sum_{h_M=0}^{B_M} \sum_{k=1}^{M} l_k \pi(l_1, \ldots, l_i, h_i, \ldots, l_M, h_M) \right) \quad (26)$$

$$U_{\mathrm{RAT}_{2\leq i\leq M}} = \frac{1}{B_i} \sum_{l_1=0}^{B_1} \cdots \sum_{l_i=0}^{B_1} \sum_{h_i=1}^{B_i} \cdots \sum_{l_M=0}^{B_1} \sum_{h_M=0}^{B_M} h_i \pi(l_1, \ldots, l_i, h_i, \ldots, l_M, h_M) \quad (27)$$

$$P_{C_{\mathrm{RAT}_i}} = \begin{cases} \sum_{l_1=0}^{B_1} \cdots \sum_{l_i=0}^{B_1} \cdots \sum_{l_M=0}^{B_1} \sum_{h_M=0}^{B_M} \Delta_i \pi(l_1, \ldots, l_i, 0, \ldots, l_M, h_M), & \text{if } h_i = 0 \\ \sum_{l_1=0}^{B_1} \cdots \sum_{l_i=0}^{B_1} \sum_{h_i=1}^{B_i} \cdots \sum_{l_M=0}^{B_1} \sum_{h_M=0}^{B_M} (P_{0i} + h_i P_{rri}) \pi(l_1, \ldots, l_i, h_i, \ldots, l_M, h_M), & \text{if } h_i > 0 \end{cases} \quad (28)$$

with rate λ_{hh}. Now, let $s_{ft} \in S$ be a subset of states satisfying the following condition:

$$s_{ft} = \{s \in S | l_1 > 0\} \bigcap s_{bp}. \quad (32)$$

Any handoff attempts during these system states will result in a premature interruption of the service due to the lack of radio resources to ensure the service provision.

4) Call Departure From Region A_s: When a mobile user performs a horizontal handoff or finishes its call in region A_s, the JRRM checks the RAT_1 status before triggering the state transition. If $l_1 > 0$ and $\sum_{k=1}^{M} l_k < T$, then the system moves from state s to state $s' = (l_1 - 1, \ldots, l_i, h_i, \ldots, l_M, h_M)$ with rate $l_1 \mu_{A_{hs}}$. Otherwise, if $\sum_{i=k}^{M} l_k = T$, then it is verified whether there are active BTSs. If there are no active BTSs, i.e., $h_i = 0$, the system will transit to the same state with the same rate. When $h_i > 0$, the load balancing will be performed using (6) to alleviate the traffic load in the selected RAT. Note that the goal here is to empty, as quickly as possible, the active BTSs, which maximizes the linear combination of the normalized traffic load and the normalized total power consumption. Hence, if $f_i = f^*$, the system will evolve from state s to state $s' = (l_1 - 1, \ldots, l_i + 1, h_i - 1, \ldots, l_M, h_M)$ with rate $l_1 \mu_{A_{hs}}$. Similarly, if more than one available RATs have the same f^*, the JRRM scheme may choose one of them randomly. If there is only one active BTS, there is no need to invoke the decision-making mechanism, and the load balancing procedure is directly performed using it. Finally, when $\sum_{i=k}^{M} l_k > T$, the system will change from state s to state $s' = (l_1 - 1, \ldots, l_i, h_i, \ldots, l_M, h_M)$ with rate $l_1 \mu_{A_{hs}}$.

5) Arrival of a New Call in Region A_i $(2 \leq i \leq M)$: When a new call arrives in region A_i, it will be handled by RAT_1 if $\sum_{k=1}^{M} l_k < T$. In this case, the system will evolve from the current state s to state $s' = (l_1, \ldots, l_i + 1, h_i, \ldots, l_M, h_M)$ with rate λ_{A_i}. Otherwise, if $\sum_{k=1}^{M} l_k \geq T$, the call will be carried by the targeted RAT. In this case, we have $s' = (l_1, \ldots, l_i, h_i + 1, \ldots, l_M, h_M)$ with rate λ_{A_i}. When the target RAT is full, the call will be carried by RAT_1, as long as it has enough room to accommodate it. In this situation, the system will evolve from state s to state $s' = (l_1, \ldots, l_i + 1, h_i, \ldots, l_M, h_M)$ with rate λ_{A_i}. However, when both the RAT_1 and the targeted RAT are full and there are other active RATs, (7) will be invoked to select a RAT to perform the load balancing. As a result, if $f_j = f^*$, $\sum_{k=1}^{M} l_k = B_1$, $h_i = B_i$, $l_j > 0$, $h_j < B_j$ $(2 \leq j \leq M)$, and $j \neq i$, then the system will move from state s to state $s' = (l_1, \ldots, l_i + 1, h_i, l_j - 1, h_j + 1, \ldots, l_M, h_M)$ with rate λ_{A_i}. If there is only one RAT available, which satisfies the requirements $l_j > 0$, $h_j < B_j$ $(2 \leq j \leq M)$, and $j \neq i$, then it will be directly selected. Let $s_{bpi} \in S$ be a subset of states satisfying the following conditions:

$$s_{bpi} = \left\{ s \in S | \sum_{k=1}^{M} l_k = B_1 \text{ and } h_i = B_i \text{ and } l_j = 0 \text{ or } h_j = B_j \text{ and } j \neq i \right\}. \quad (33)$$

The arrival of new calls in the ith RAT in states $s \in S$ that falls inside s_{bpi} will be blocked since the system will be unable to provide the radio resources to ensure the call acceptance.

6) User Moving From Region A_s to Region A_i $(2 \leq i \leq M)$, and Vice Versa, Using RAT_1 Radio Resources: As previously discussed, the JRRM scheme will keep the mobile users connected with the RAT_1. Thus, if $l_1 > 0$, the system will evolve from state s to state $s' = (l_1 - 1, \ldots, l_i + 1, h_i, \ldots, l_M, h_M)$ with rate $l_1 \mu_{A_{si}}$. Similarly, if $l_i > 0$, a mobile user will move from region A_i to region A_s with rate $l_i \mu_{A_i}$, and the system state will change from state s to state $s' = (l_1 + 1, \ldots, l_i - 1, h_i, \ldots, l_M, h_M)$.

7) User Moving From Region A_i $(2 \leq i \leq M)$ to Region A_s Using RAT_i Radio Resources: To resume its communication seamlessly, a vertical handoff has to be performed when a mobile user leaves the ith RAT toward RAT_1 before ending its call. In this case, regarding the system configuration, the CTMC model will evolve from state s to different states. Thus, if $h_i > 0$ and $T \leq \sum_{k=1}^{M} l_k < B_1$, the system will move from the current state s to state $s' = (l_1 + 1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M)$ with rate $h_i \mu_{A_i}$. In cases where RAT_1 is unable to handle this call due to lack of radio resources, i.e., $\sum_{k=1}^{M} l_k = B_1$, but $l_i > 0$ and $l_j > 0$, $h_j < B_j$ $(2 \leq j \leq M)$, and $j \neq i$; (7) will be invoked to select an available active RAT to perform the load balancing. Hence, if $f_i = f^*$, the system will transit from state s to state $s' = (l_1 + 1, \ldots, l_i - 1, h_i, l_j, h_j, \ldots, l_M, h_M)$ with rate $h_i \mu_{A_i}$. Otherwise, if $f_j = f^*$, the system will evolve from state s to state $s' = (l_1 + 1, \ldots, l_i, h_i - 1, l_j - 1, h_j + 1, \ldots, l_M, h_M)$ with the same rate. It should be noted that if only the ith RAT meets the specified requirement, then the load balancing procedure will be directly applied on it. On the other hand, if $l_i = 0$ and $l_j > 0$, $h_j < B_j$ $(2 \leq j \leq M)$, and $j \neq i$; then, (7) will be invoked again, but without involving the ith RAT. Let $s_{fti} \in S$ be a subset of states satisfying the following conditions:

$$s_{fti} = \left\{ s \in S | \sum_{k=1}^{M} l_k = B_1 \text{ and } h_i > 0 \text{ and } l_i = 0 \text{ and } l_j = 0 \text{ or } h_j = B_j \text{ and } j \neq i \right\}. \quad (34)$$

A mobile user that moves from region A_i to region A_s in a state $s \in s_{fti}$ will suffer a forced termination of its call. In this case, the system will transit from the current state s to state $s' = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M)$ with rate $h_i \mu_{A_i}$.

8) Call Completion in Region A_i $(2 \leq i \leq M)$ Using RAT_1 Radio Resources: When the mobile user ends its call before moving out of region A_i and the RAT_1 radio resource occupancy is lower than the threshold value T, the system will evolve from the current state s to state $s' = (l_1, l_2, h_2, \ldots, l_i - 1, h_i, \ldots, l_M, h_M)$ with rate $l_i \mu_c$. The same transition will occur if $\sum_{k=1}^{M} l_k = T$ and $h_i = 0$. On the other hand, provided that $h_i > 0$, (6) will be invoked to select a RAT to have its load reduced. In this case, if $f_i = f^*$, the system will leave the current state s and move to state $s' = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M)$ with rate $l_i \mu_c$. It is worth mentioning that

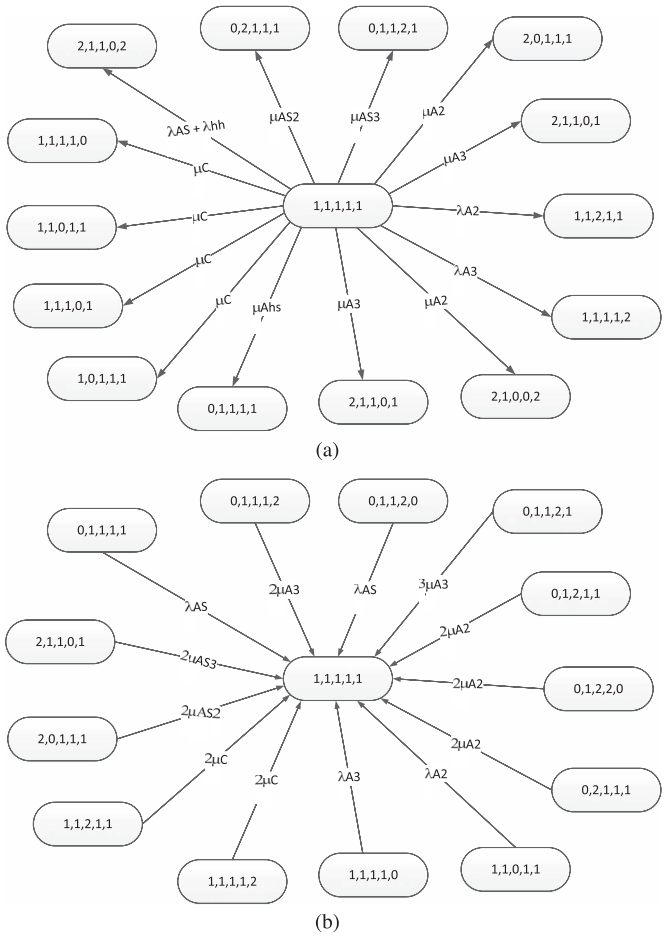

Fig. 3. State of the multidimensional CTMC for $B_1 = B_2 = B_3 = 3$ and $T = 1$.

immediately after releasing the RAT_1 radio resource, an ongoing call of the ith RAT will be transferred from the ith RAT to RAT_1 to alleviate its traffic load and reduce its power consumption. On the other hand, if $f_j = f^*$, the system will move from the current state to state $s' = (l_1, \ldots, l_i - 1, h_i, l_j + 1, h_j - 1, \ldots, l_M, h_M)$ with rate $l_i \mu_c$. This way, the released RAT_1 radio resource will now be used to reduce the traffic load of the jth RAT $(2 \leq j \leq M)$ and $j \neq i$.

It should be noted that when the ith RAT is inactive, i.e., $h_i = 0$, it is not a candidate in the decision-making procedure. Furthermore, when it is active, the system will evolve considering only that RAT in the load balancing procedure. Finally, when $\sum_{k=1}^{M} l_k > T$ and there are ongoing users finishing their calls, the system will directly transit from state s to state $s' = (l_1, \ldots, l_i - 1, h_i, \ldots, l_M, h_M)$ with rate $l_i \mu_c$ without invoking the load balancing procedure.

9) Call Completion in Region A_i $(2 \leq i \leq M)$ Using RAT_i Radio Resources: When the mobile user ends its call in the ith RAT, the system configuration is evaluated to check if the load balancing applies. This way, if $\sum_{k=1}^{M} l_k = T$, $h_i > 0$, $l_i > 0$, $h_j > 0$ for $j \neq i$; then, (6) will be invoked to decide which RAT will have its load alleviated. Therefore, if $f_i = f^*$, the system will move from the current state s to state $s' = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M)$ with rate $h_i \mu_c$. Otherwise, it will move from the current state to state $s' = (l_1, \ldots, l_i - 1, h_i, l_j + 1, h_j - 1, \ldots, l_M, h_M)$ with rate $h_i \mu_c$, as long as $f_j = f^*$. Furthermore, if $h_j = 0$ $(2 \leq j \leq M)$ for $j \neq i$ or $l_i = 0$ $(2 \leq i \leq M)$ or $\sum_{k=1}^{M} l_k > T$, then the CTMC model will directly move from the current state s to state $s' = (l_1, \ldots, l_i, h_i - 1, \ldots, l_M, h_M)$ with rate $h_i \mu_c$.

10) State Transition Diagram: Due to the complexity of the multidimensional CTMC model previously introduced, it is rather impractical to graphically represent a complete state transition diagram even for a small-scale HetNet. Thus, we provide in Fig. 3(a) and (b) an example of a particular state and all the possible transitions from/to it for a HetNet composed of three RATs $(M = 3)$, assuming that $B_1 = B_2 = B_3 = 3$, $\alpha = 0.75$, $P_{T2} = 400$ W, $P_{T3} = 300$ W, and $T = 1$. We focus on the state $s = (l_1 = 1, l_2 = 1, h_2 = 1, l_3 = 1, h_3 = 1) \in S$, which gives $f_2 = 0.5$ and $f_3 = 0.4375$.

In Fig. 3(a), it can be observed that when a new call or a horizontal-handoff call arrives into the system with rate $\lambda_{A_s} + \lambda_{hh}$, the proposed JRRM accepts the incoming service request by performing the load balancing procedure, and the system moves from state $s = (1, 1, 1, 1, 1)$ to state $s' = (2, 1, 1, 0, 2)$. It should be noted that the load balancing procedure is executed because RAT_1 is full, there are users in regions A_2 and A_3 using its radio resources, and RAT_2 and RAT_3 are not full. Thus, since $f_3 < f_2$, the JRRM scheme deallocates the radio resource used by an ongoing mobile user in region A_3 (l_3) and allocates it to the incoming request in region A_s, transferring that ongoing mobile user from RAT_1 to RAT_3. When a mobile user moves from region A_s to region A_2 and is connected to RAT_1, the JRRM scheme will keep the ongoing call associated with the same RAT. As a result, the system will transit from state $s = (1, 1, 1, 1, 1)$ to state $s' = (0, 2, 1, 1, 1)$ with rate $\mu_{A_{s2}}$. It should be noted that the next state transitions in a clockwise direction follow the same principle. An arrival of a new call in region A_2 will be accepted by the JRRM scheme into RAT_2 as long as it has space to room it, and the RAT_1 radio resource occupancy is greater than the predefined threshold. Therefore, the system will leave state $s = (1, 1, 1, 1, 1)$ and move to state $s' = (1, 1, 2, 1, 1)$ with rate λ_{A_2}. The same principle applies for an arrival of a new call in region A_3. When a mobile user connected to RAT_2 moves to region A_s, a vertical-handoff process is triggered. In this case, there is no radio resource available in the RAT_1 region; therefore, the load balancing procedure is invoked to redistribute the radio resources among the ongoing calls in the region. After transferring an ongoing call supported by RAT_1 in region A_3 to RAT_3, the JRRM scheme performs the vertical-handoff procedure, and the system will transit from state $s = (1, 1, 1, 1, 1)$ to state $s' = (2, 1, 0, 0, 2)$ with rate μ_{A_2}. In a similar way, the JRRM scheme will deal with a mobile user connected to RAT_3 that moves to region A_s. The following state transitions are the call departure in region A_s and the call completion in regions A_2 and A_3. These transitions follow the standard state-dependent service rate concept used in classical queuing theory [37]. Based on the above methodology, the state transitions to state $s = (1, 1, 1, 1, 1)$, as shown in Fig. 3(b), can be similarly obtained.

11) Performance Metrics: Consider the state space defined in (30) and the random events described in Section V-C2–C9 as the entries of the infinitesimal generator matrix $\boldsymbol{Q}$

representing the implementation-friendly CTMC model. After mounting $\boldsymbol{Q}$, we use the SOR method [37] to find the steady-state probability vector $\boldsymbol{\pi}$. With $\boldsymbol{\pi}$, we can compute the performance metrics.

Equation (31) determines the conditions under which the blocking of a new call occurs in region A_s. Thus, $P_{BP_{A_s}}$ is given by

$$P_{BP_{A_s}} = \sum_{s \in s_{bp}} \pi(s). \tag{35}$$

Similarly, (32) defines the conditions under which the forced termination probability of horizontal-handoff calls happens in region A_s. Thus, $P_{FT_{A_s}}$ is expressed as

$$P_{FT_{A_s}} = \sum_{s \in s_{ft}} \pi(s). \tag{36}$$

Considering the summation of the probabilities of all states that satisfy (33), we obtain the blocking probability of new calls in region A_i $(2 \leq i \leq M)$, i.e.,

$$P_{BP_{A_i}} = \sum_{s \in s_{bpi}} \pi(s). \tag{37}$$

Similarly, the forced termination probability of the vertical-handoff calls in region A_i $(2 \leq i \leq M)$, $P_{FT_{A_i}}$, is obtained by the summation of the probabilities of all states that satisfy (34), i.e.,

$$P_{FT_{A_i}} = \sum_{s \in s_{fti}} \pi(s). \tag{38}$$

The average blocking probability, bandwidth utilization, average power consumption in the ith RAT $(2 \leq i \leq M)$, as well as the average power savings in the ith RAT $(2 \leq i \leq M)$ are computed using (25)–(29).

D. Computational Complexity

We examine the computational aspects of the proposed JRRM schemes. To obtain the optimal policy, we initially need to build the Markov process, which means constructing a set of states $x \in X$ in which each state is defined by a $2M$-tuple of nonnegative integers. This procedure is performed in $\mathcal{O}(B_1 + 1 \times B_1 + 1 \times B_2 + 1 \times \cdots \times B_1 + 1 \times B_i + 1 \times \cdots \times B_1 + 1 \times B_M + 1 \times 4M)$. Next, the constraints that define the feasible states are verified to construct the final multilinear set. The action space A is mounted considering the number of controlling actions per state, which is done in $\mathcal{O}(A(x_1) \times A(x_2) \times \cdots \times A(x_i) \times \cdots \times A(x_{\max}))$, where $x_{\max}$ is the last state in the set. Finally, the classical value iteration algorithm is applied, which has computational complexity of $\mathcal{O}(AX^2)$ [40]. Under the optimal policy, the CTMC model is solved. In this paper, we have used the SOR method [37], which quickly converges to the steady-state distribution probability, despite the fact that its convergence has not been mathematically determined [37]. On the other hand, the computational complexity of the implementation-friendly algorithm consists of building a multilinear set where each state $s \in S$ is a $(2M-1)$-tuple of nonnegative integers. This procedure is done in $\mathcal{O}(B_1 + 1 \times B_1 + 1 \times B_2 + 1 \times \cdots \times B_1 + 1 \times B_i + 1 \times \cdots \times B_1 + 1 \times B_M + 1)$. After applying the system constraints to define the feasible states, the SOR method is again applied to find the steady-state distributions probability. To sum up, the optimal JRRM scheme takes longer time to run compared with the implementation-friendly scheme.

TABLE I
SYSTEM SETTING FOR PERFORMANCE EVALUATION

Parameter	Symbol	Value
Number of radio resources RAT_1	B_1	15
Number of radio resources RAT_2	B_2	4
Number of radio resources RAT_3	B_3	4
Radius of RAT_1	R_1	600 m
Radius of RAT_2	R_2	180 m
Radius of RAT_3	R_3	200 m
Mean call holding time	$1/\mu_c$	120 s
Total power consumption of the RAT_2	P_{T2}	3802 W
Total power consumption of the RAT_3	P_{T3}	300 W
Power consumption to support the BTS infrastructure support systems of RAT_2	P_{02}	1402 W
Power consumption to support the BTS infrastructure support systems of RAT_3	P_{03}	150 W

VI. NUMERICAL RESULTS

We consider a HetNet with three RATs: a macrocell (RAT_1) and two inner RATs (RAT_2 and RAT_3), each of which has its own BTS power consumption profile specified according to [4]. The full set of network parameters used in the numerical evaluation is outlined in Table I. The number of radio channels in the HetNet is computed in such a way that the blocking probabilities and forced termination probabilities stay around 2% for a given total offered traffic load Λ/μ_c. This way, for $\Lambda/\mu_c = 14.45$, we have $P_{BP_{A_s}} \approx 0.02$, $P_{BP_{A_2}} \approx P_{BP_{A_3}} \approx 0.01$, $P_{ABP} \approx 0.02$, $P_{FT_{A_s}} \approx 0.02$, $P_{FT_{A_2}} \approx 0.01$, and $P_{FT_{A_3}} \approx 0.01$.

A. Performance Comparison

We compare the performance of the proposed JRRM schemes against that of a chosen benchmark scheme. Although there is a variety of energy-efficient algorithms for wireless networks in the literature, it is difficult to find an algorithm that shares the same features assumed in our scenario. In addition, most of the approaches proposed in the literature rely on system-level simulations, whereas our framework consists of analytical models. In this sense, we selected the FBSMA [10] because a HetNet in which the macrocell always stays on has been considered. The main idea behind this algorithm consists in activating only the PAs needed to carry the actual offered traffic load, given the total number of radio resources computed according to the total offered traffic load Λ/μ_c and a prespecified QoS profile. In the use of this benchmark scheme, we have assumed that a new call or horizontal-handoff call will be accepted in region A_s, as long as there is an available radio resource [26]. Moreover, we have assumed that the mobile users moving from region A_s to region A_i $(i = 2, 3)$ will keep the connection with the macrocell, and *vice versa*, if they are using the RAT_1 radio resources. Moreover, a new call in region

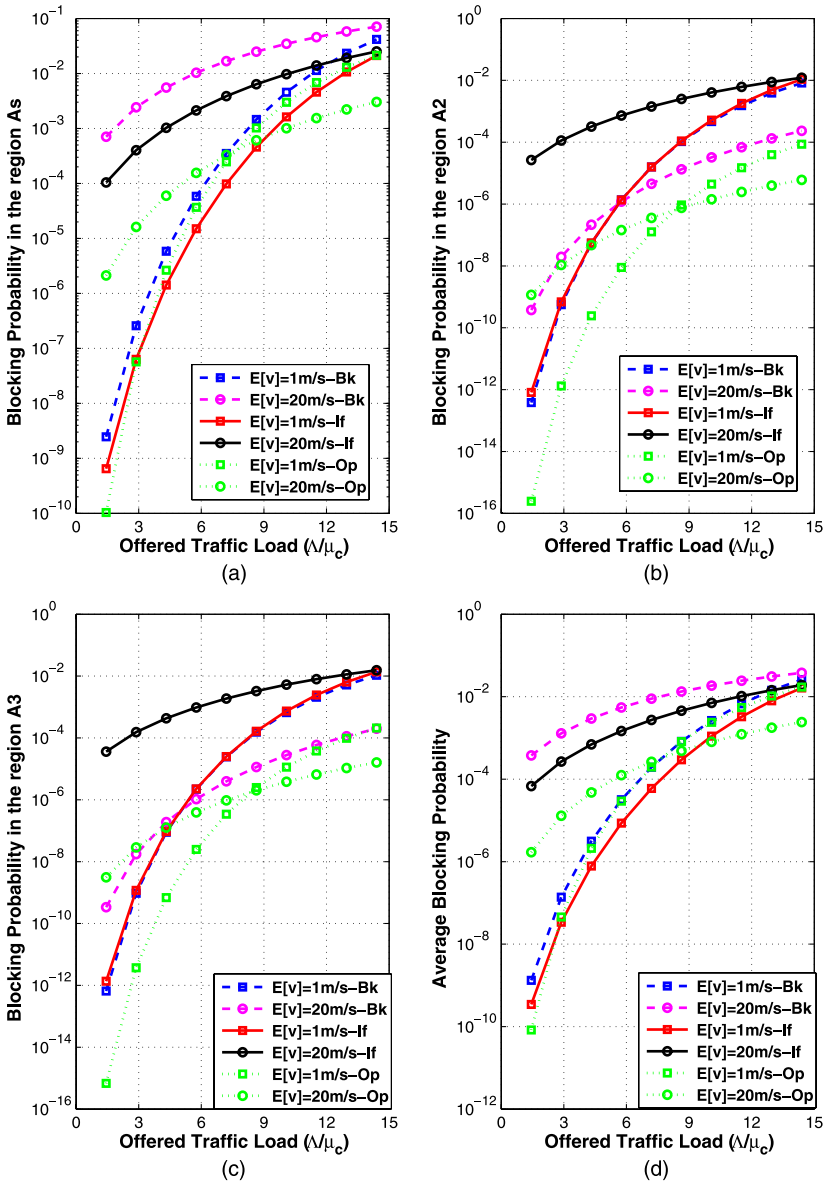

Fig. 4. Blocking probabilities of new calls for different $E[V]$ and Λ/μ_c values.

A_i will be handled by the ith RAT provided that there is a radio resource to room the call; otherwise, it will be accepted by the macrocell on the condition that it has an idle radio resource. A vertical-handoff call will be accepted by RAT_1 if it has space to accommodate the call; otherwise, it will be forced to terminate. In terms of the power saving strategy, the considered benchmark scheme does not perform the activation and deactivation of the entire BTS site, and its performance was determined by means of a CTMC model.

Figs. 4–7 show the system performance for the proposed JRRM schemes compared with that of the benchmark scheme, taking into account an increase in the offered traffic load Λ/μ_c, two mobile user average speed profiles ($E[V] = 1$ and 20 m/s), a threshold $T = 10$, and a weighting factor $\alpha = 0.5$.

Fig. 4(a) shows that all the schemes have their blocking probabilities of new calls in region A_s increased when the offered traffic load Λ/μ_c and the mobile user average speed $E[V]$ increased, but due to the load balancing procedure, the proposed JRRM schemes outperform the benchmark scheme. Note that for a low-speed mobile user, the implementation-friendly JRRM scheme performs slightly better than the optimal controller. This occurs because the optimal policy (described in the Appendix) seeks to empty the inner RATs, which makes the load in the macrocell heavier. However, as the mobile users move fast, the optimal controller achieves the lowest probability.

Considering the blocking probabilities of new calls in regions A_2 and A_3, Fig. 4(b) and (c) shows that for a low-speed mobile user, the benchmark scheme and the implementation-friendly JRRM scheme achieve a similar performance. However, for a high-speed mobile user, the benchmark scheme achieves a superior performance compared with the implementation-friendly JRRM scheme. This is attributed to the fact that when the velocity of users increases, the number of horizontal handoff (respectively, vertical handoff) requests in the entire region also increases. To successfully deal with these requests in region A_s,

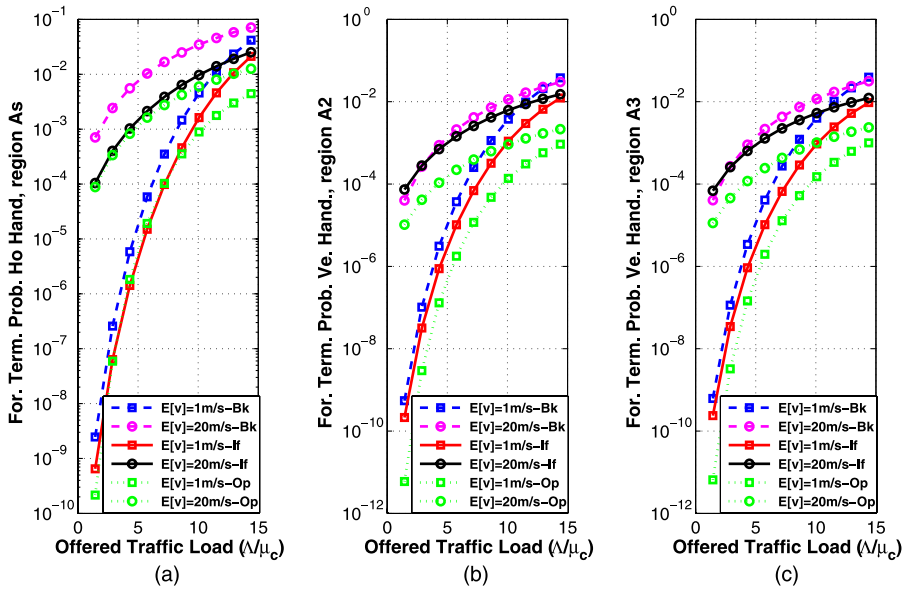

Fig. 5. Forced termination probabilities of handoff calls for different $E[V]$ and Λ/μ_c values.

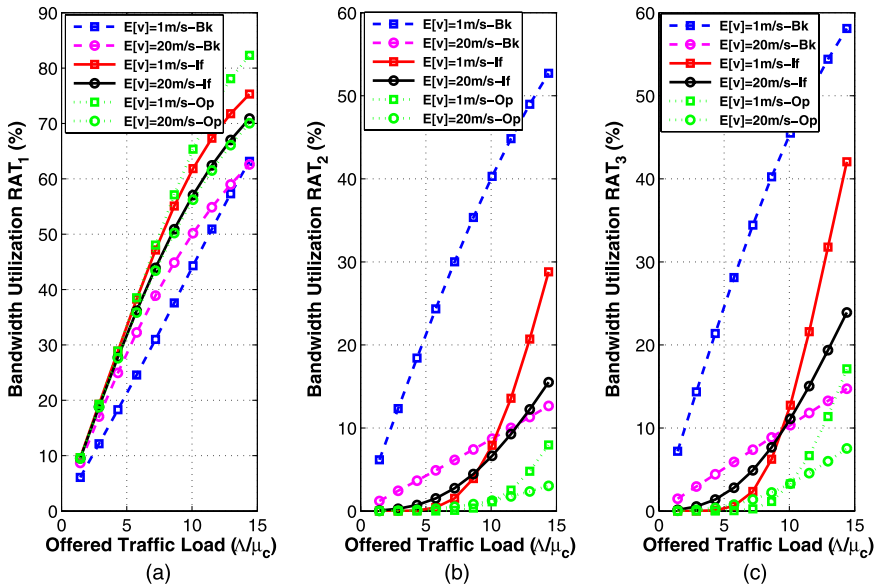

Fig. 6. Bandwidth utilization in each RAT for different mobile $E[V]$ and Λ/μ_c values.

the load balancing procedure starts to dynamically redistribute the traffic load among RAT_1, RAT_2, and RAT_3. As such, because the radio resources of RAT_2 and RAT_3 in the benchmark scheme are dedicated to cope with their own offered traffic load, they can support more new incoming calls. However, since the optimal JRRM scheme makes an optimized use of the radio resources (see Appendix), it has the lowest blocking probability in both regions. Finally, Fig. 4(d) shows that irrespective of Λ/μ_c and $E[V]$, the proposed JRRM schemes outperform the benchmark scheme. It indicates that from the aggregated HetNet performance standpoint, the proposed JRRM schemes make a more efficient use of the common pool of radio resources.

In Fig. 5, it can observed that compared with the benchmark scheme, the proposed JRRM schemes achieve the lowest forced termination probabilities of handoff calls in all the regions of the HetNet. Based on the fact that, from a user perspective, the forced termination of a handoff call is more annoying than the blocking of a new call, it can be concluded that the proposed JRRM schemes are more effective and efficient than the benchmark scheme in prioritizing the handoff calls. Notably, the optimal controller outperforms the implementation-friendly JRRM scheme. However, the difference in performance comes at the cost of an increase in the complexity of the policy.

Fig. 6 shows bandwidth utilization in every RAT comprising the HetNet. As expected, the bandwidth utilization in RAT_1 [as shown in Fig. 6(a)] is higher when it operates according to the proposed JRRM schemes. This occurs because they deactivate the inner RATs by handing off their traffic loads to the macrocell. Indeed, the optimal controller (see the Appendix)

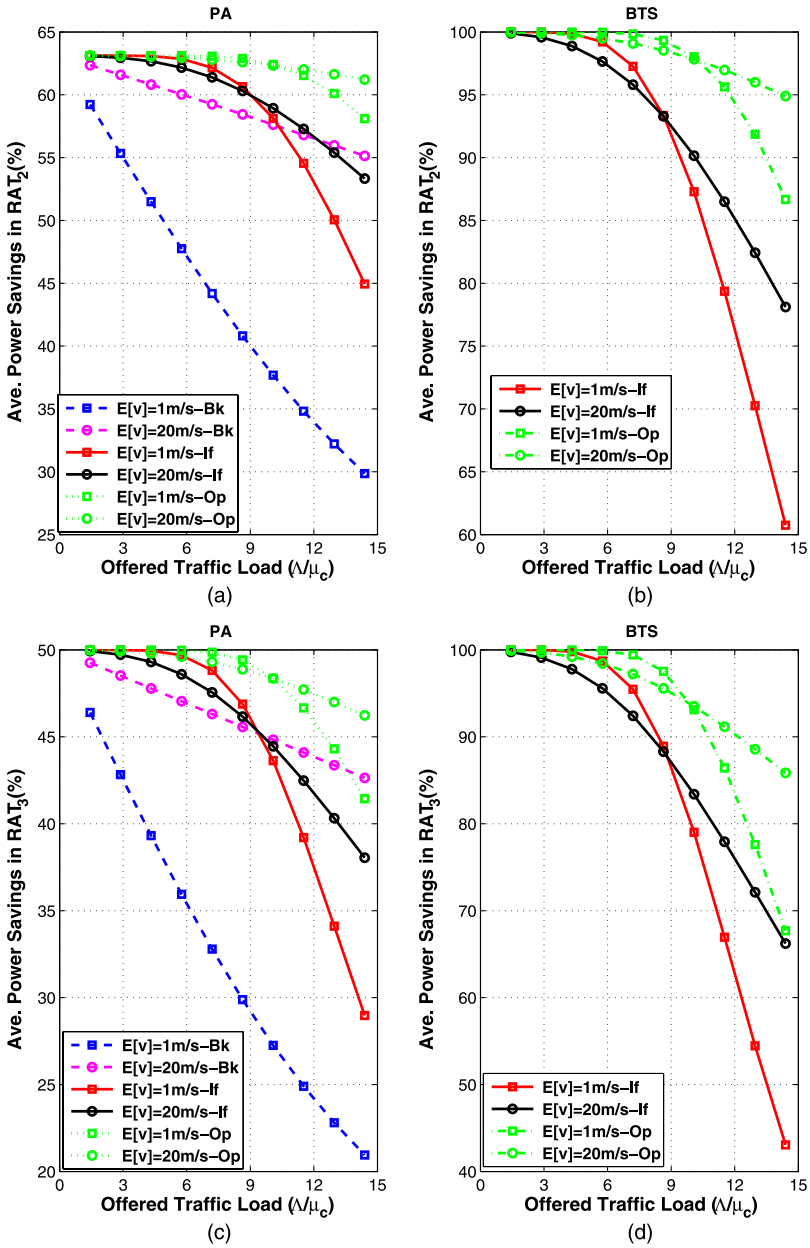

Fig. 7. Average power savings for RAT_2 and RAT_3 for different $E[V]$ and Λ/μ_c values and power saving strategies.

seeks to empty the inner RATs to save energy, and the load balancing procedure is applied to ensure QoS provisioning. Consequently, with the proposed JRRM schemes, the bandwidth utilizations in RAT_2 and RAT_3 are the lowest possible, particularly when the optimal controller is operating under a low offered traffic load, leading to the highest average power savings in both RATs, as shown in Fig. 7. Moreover, in Fig. 7, it can be observed that the benchmark scheme also presents the lowest average power savings in RAT_2 and RAT_3, particularly for low-speed mobile users. This phenomenon occurs because these users tend to stay longer in those RATs using their radio resources.

Fig. 7 displays a performance comparison between both types of power saving strategies. Focusing on the implementation-friendly JRRM scheme, it can be observed that while the PA-only switching-on/off procedure achieves a substantial level of power saving (up to almost 63%), the switching-on/off procedure of the entire BTS site achieves remarkable levels of power savings (up to almost 100%), particularly for low traffic load where the energy efficiency has been envisioned to lead to better results. Regarding the optimal JRRM scheme, it can be seen that it reaches a similar level of power savings. However, as the traffic load increases, its power saving levels decay at a slower rate than those of the implementation-friendly JRRM scheme.

In summary, it can be concluded that the optimal controller presents the best performance, particularly in terms of the achieved power savings. However, the optimal policy may be complex (see the Appendix). Given this fact, the adoption of the implementation-friendly JRRM scheme may be useful for MNOs. Therefore, the following sections are meant to unveil the implementation-friendly properties.

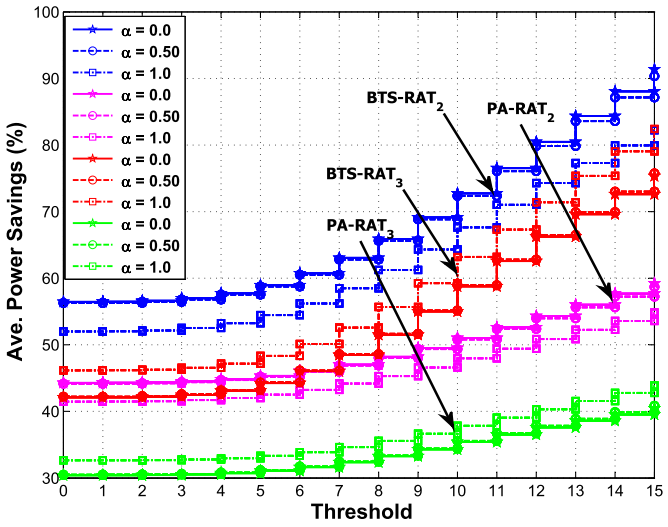

Fig. 8. Average power savings in RAT_2 and RAT_3 considering both power saving strategies: activation and deactivation of the entire BTS site and PA only.

B. Analysis of Threshold and Weighting Factor

We study how the implementation-friendly JRRM scheme can be fine tuned to save more power by properly setting threshold T and weighting factor α. We consider that $\Lambda/\mu_c = 14.45$, $E[V] = 10$ m/s, whereas T and α are varying parameters.

As shown in Fig. 8, by increasing threshold T, the average power savings in RAT_2 and RAT_3 also increase, irrespective of the adopted power saving strategy. It can also be observed that there is a considerable upward shift on the average power saving value when the entire BTS site switching-on/off procedure is adopted. For instance, when considering RAT_2, and $T = 15$, and $\alpha = 0.0$, it can be observed that the average power savings can reach up to 60% when the PA-only switching-on/off procedure is selected (respectively, more than 90% when the switching-on/off procedure of the entire BTS site is chosen). The same trend is observed for RAT_3. When $\alpha = 1$, a significant difference in the average power savings is observed. This can be attributed to the fact that for this setting, the load balancing procedure takes into account only the normalized traffic load as the criterion. Therefore, RAT_2 is more often selected than RAT_3. Indeed, RAT_2 is the smallest RAT, and according to the assumption that mobile users are uniformly distributed over the HetNet, it has the lowest traffic load. Therefore, the average power saving is higher for RAT_3 compared with that obtained for RAT_2. For the other settings, when the energy factor is taken into account in the load balancing procedure, the situation is the opposite as long as the total power consumption of RAT_2 is much higher than that of RAT_3. In this case, the JRRM scheme will select RAT_3 as often as possible.

Fig. 9(a) shows the bandwidth utilization in RAT_1, RAT_2, and RAT_3, whereas Fig. 9(b) portrays the contributions of the mobile users to the bandwidth utilization of RAT_1 in regions A_s, A_2, and A_3. It can be observed that regardless of α, the bandwidth utilization of RAT_1 quickly rises as threshold T increases. This is due to the fact that the higher the threshold T is, the longer is the duration that RAT_2 and RAT_3 will remain deactivated. As a consequence, their bandwidth utilizations drastically drop due to the increase in threshold T. In Fig. 9(b), it can be observed that this drop is outweighed by an increase in the bandwidth utilization of RAT_1 due to the presence of mobile users in regions A_2 and A_3, which are mostly served by RAT_1 when threshold T increases. It can also be observed that mobile users in region A_s have a fixed contribution to the bandwidth utilization of RAT_1.

C. Finding the Desirable Threshold Setting

The results described previously have shown that the implementation-friendly JRRM scheme may be considered as a viable solution to achieve an eco-friendly network operation by properly setting threshold T on the macrocell radio resources. In this respect, we propose a systematic approach (shown in Fig. 10) to find the desirable threshold T^* setting, given a prespecified power saving level D_{ps} and the employed power saving strategy. In a nutshell, the algorithm iteratively increases the threshold T value between 0 and B_1 and compares the achieved average power saving P_{S_i}, $(2 \leq i \leq M)$ against the prespecified power saving level for all inner RATs. After finding the prespecified power saving level for all inner RATs or ending its search in the whole set of threshold values, the algorithm returns the desirable threshold T^* setting or informs that the prespecified power saving level D_{ps} cannot be achieved for the current network configuration if at least one RAT does not satisfy this specification. For this algorithm, we consider that $\alpha < 1$ since an energy-efficient network design is assumed.

Table II presents an analysis of how the algorithm presented in Fig. 10 performs for different prespecified power saving levels D_{ps}. It is assumed that $\alpha = 0.0$, and the same scenario described in the previous section is considered. Therefore, Fig. 8 can be used for comparison purposes. Table II shows that, independently of the chosen power saving strategy, when D_{ps} increases, the desirable threshold T^* setting also increases to reach the prespecified energy savings for all inner RATs. Initially, for $D_{ps} = 38\%$, both power saving strategies meet the desired specification. It is worth mentioning that when the option of switching on/off the entire BTS site is adopted, the algorithm meets D_{ps} with $T^* = 0$. On the other hand, when the option of switching on/off the PAs only is considered, the obtained threshold is $T^* = 13$. The next specification, i.e., $D_{ps} = 45\%$, is only satisfied when the option of switching on/off the entire BTS site is adopted. In this case, it can be observed that for the option of switching on/off the PAs only, the average power savings of RAT_3, i.e., P_{S_3}, is 40.44%, which is below the prespecified power saving level. For the last test, the algorithm is unable to meet the prespecified power saving level $D_{ps} = 90\%$ for both power saving strategies.

VII. Conclusion

We have proposed two QoS-aware energy-efficient JRRM schemes for HetNets, which ensure energy savings while preserving the system performance. The optimal JRRM scheme achieved remarkable levels of energy savings and system performance. However, its optimal policy may be intricate, which makes it challenging to be deployed in practice. Although outperformed by the optimal JRRM scheme, the

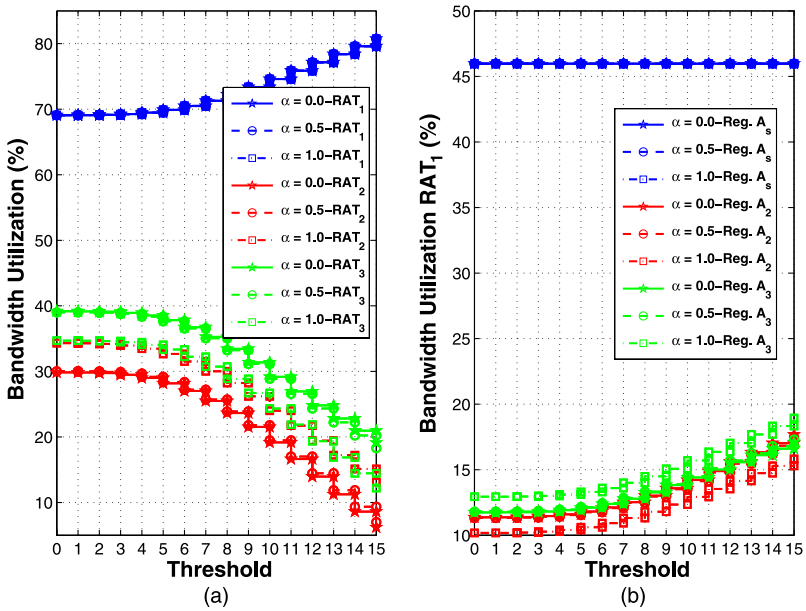

Fig. 9. Bandwidth utilization: (a) RAT_1, RAT_2, and RAT_3. (b) Contributions due to regions A_s, A_2, and A_3 in RAT_1.

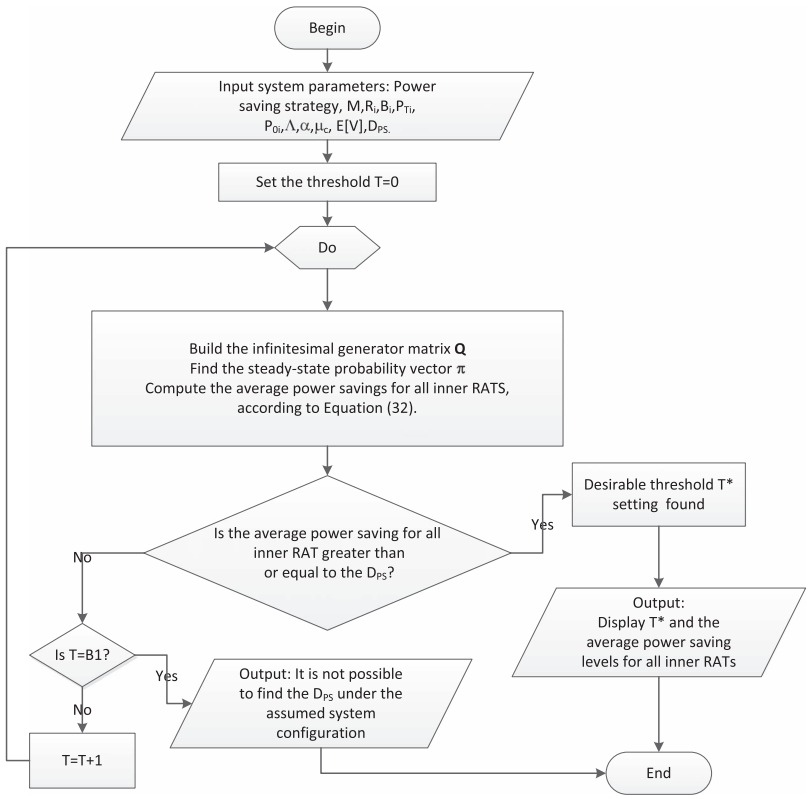

Fig. 10. Flowchart of the desirable threshold setting procedure.

implementation-friendly JRRM scheme also considerably reduces the power consumption of the inner RATs while keeping the system performance. Thus, it can be used in practice as a feasible eco-friendly design solution for MNOs in dense urban areas.

In this paper, to provide a fair comparison with the implementation-friendly JRRM scheme, we have used (3) and (23) as the decision-making criteria to govern the optimal JRRM scheme's dynamic. Nevertheless, it is worthy to mention that other criteria might be used for the system design. An example is the weighted sum of the total normalized traffic load and the BTS power consumption for all inner RATs. For this criterion, rather than considering the inner RATs individually, we take all of them into account.

TABLE II
ANALYSIS OF THE DESIRABLE THRESHOLD T SETTING

Pre-specified power saving level D_{ps} %	Power saving strategy	Desirable Thre. T^*	Av. Power. Sav. P_{S_2} %	Av. Power. Sav. P_{S_3} %
38	Switching on/off the entire BTS site	$T^* = 0$	56.45	42.10
38	Switching on/off the PAs only	$T^* = 13$	56.02	38.59
45	Switching on/off the entire BTS site	$T^* = 6$	60.73	46.03
45	Switching on/off the PAs only	Not possible	59.20	40.44
70	Switching on/off the entire BTS site	$T^* = 14$	88.06	72.63
70	Switching on/off the PAs only	Not possible	59.20	40.44
90	Switching on/off the entire BTS site	Not possible	91.36	75.34
90	Switching on/off the PAs only	Not possible	59.20	40.44

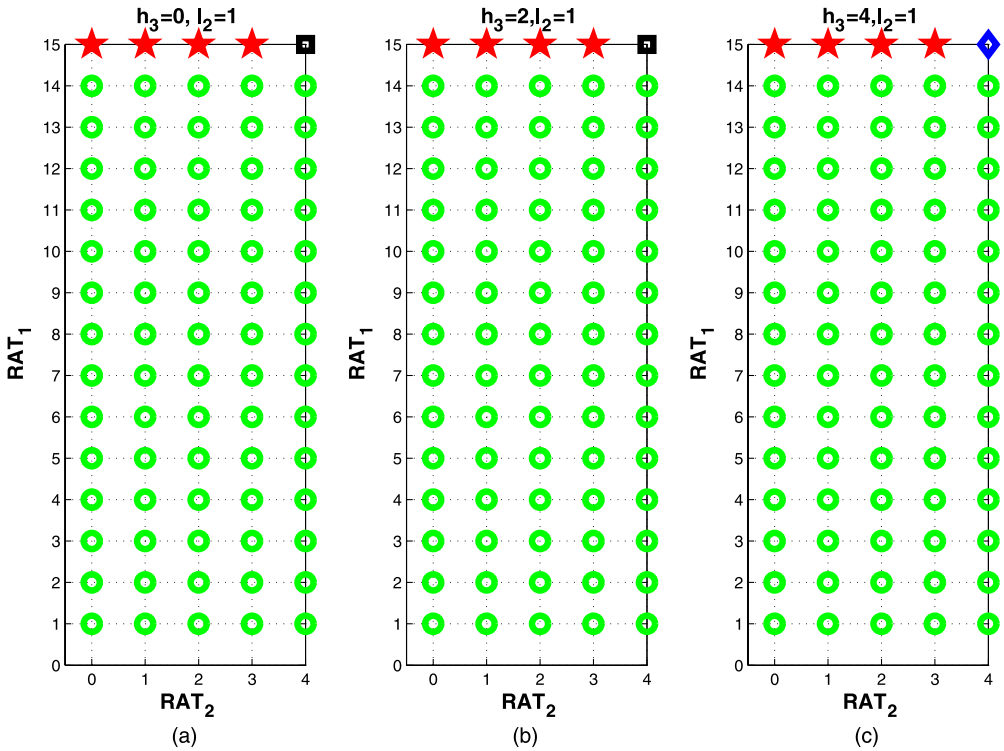

Fig. 11. Arrival of a new call in region A_2.

To obtain further energy efficiency gains, the proposed JRRM schemes could be refined considering the following directions.

- We consider that the activation of inner RATs is triggered when the load in macrocell reaches the threshold value. This procedure could be enhanced through the application of the sequential activation process, which turns on only the RATs covering regions supporting ongoing connections while keeping the remainder off saving energy.
- We consider a general multi-RAT HetNet and design JRRM schemes that are applicable to it. However, when the HetNet is formed by a massive number of low-power BTSs, as will be expected in an LTE HetNet, the power consumption could be optimized if the load balancing shifts the macrocell load to the small cells after their activation. At the limit, the macrocell could be turned off, saving even more energy.

APPENDIX
STRUCTURE OF THE OPTIMAL POLICY

Here, we provide some examples of the optimal policy structure in region A_2. As follows, we adopt the following convention: a) ○ denotes the acceptance in RAT_1; b) ☆ denotes the acceptance in RAT_2; c) △ denotes the load balancing in region A_2; d) □ denotes the load balancing in region A_3; e) + denotes the vertical handoff from RAT_2 to RAT_1; f) ◇ denotes the action "doing nothing." Given the complexity of the optimal policy, only the settings considering $\Lambda/\mu_c = 14.45$, $E[V] = 20$ m/s, $\alpha = 0.5$, $l_2 = 1$ for $h_3 = 0, 2, 4$, and the BTS activation/deactivation are outlined.

Fig. 11 shows that new calls in region A_2 are often served by RAT_1, whereas RAT_2 and RAT_3 are only demanded when RAT_1 is full. In this sense, the incoming request is only blocked when it is not possible to accept it in RAT_2 or perform the load balancing. Fig. 12 shows that when the vertical handoff cannot be directly performed, the optimal controller calls the

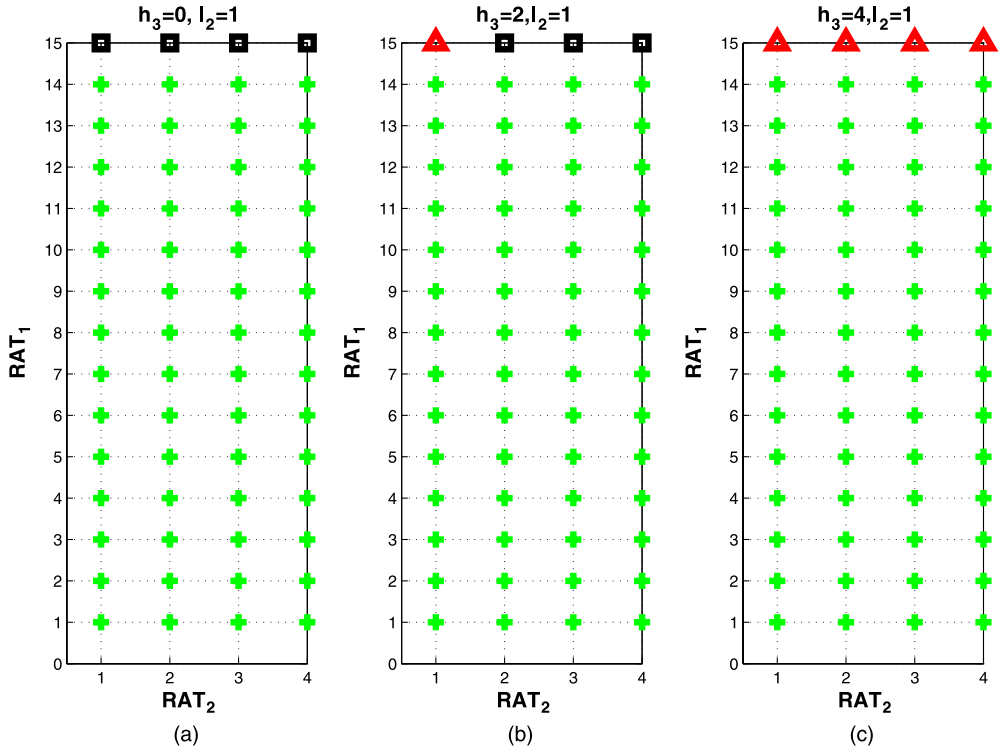

Fig. 12. User moving from region A_2 to region A_1 using RAT_2 radio resources.

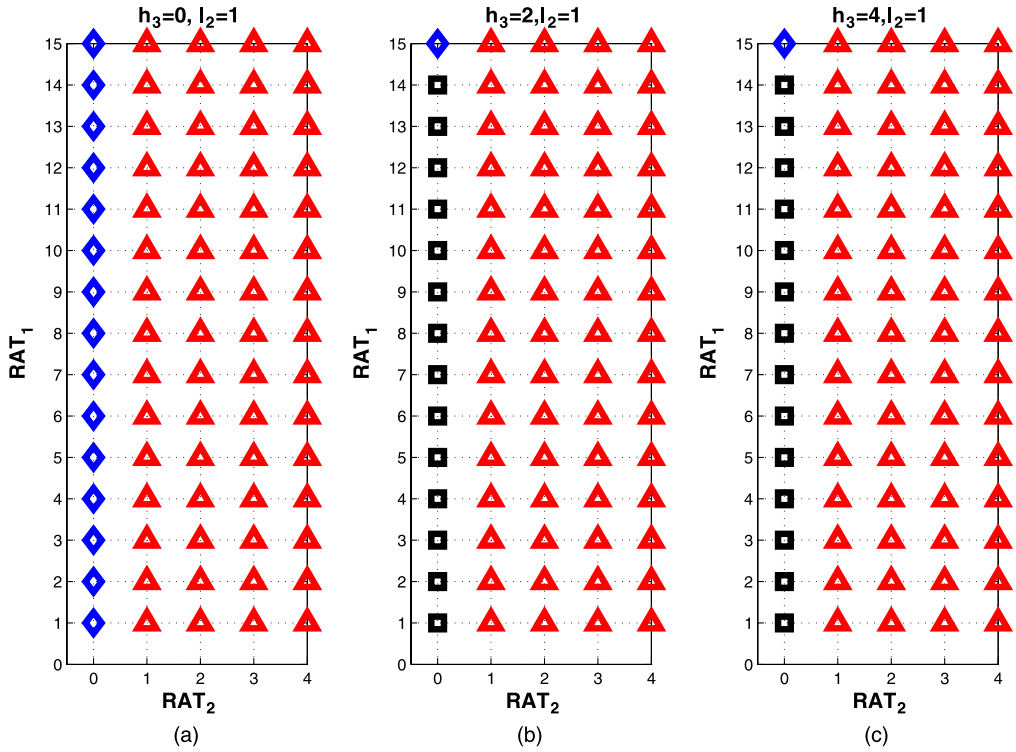

Fig. 13. Call completion in region A_2 using RAT_1 radio resources.

load balancing procedure. In this case, RAT_3 is the preferable option. However, as the its load becomes heavier, it is no longer chosen.

Fig. 13 shows that when a call leaves RAT_1, the optimal controller seeks to empty RAT_2 and RAT_3. Given the higher power consumption in RAT_2, it is selected more frequently. In this regard, RAT_3 is only chosen when RAT_2 is idle. Fig. 14 follows the same reasoning. It should be noted that when there is no call in progress in RAT_3, the optimal controller does nothing [see Fig. 14(a)]. When RAT_3 is 50% loaded [see Fig. 14(b)],

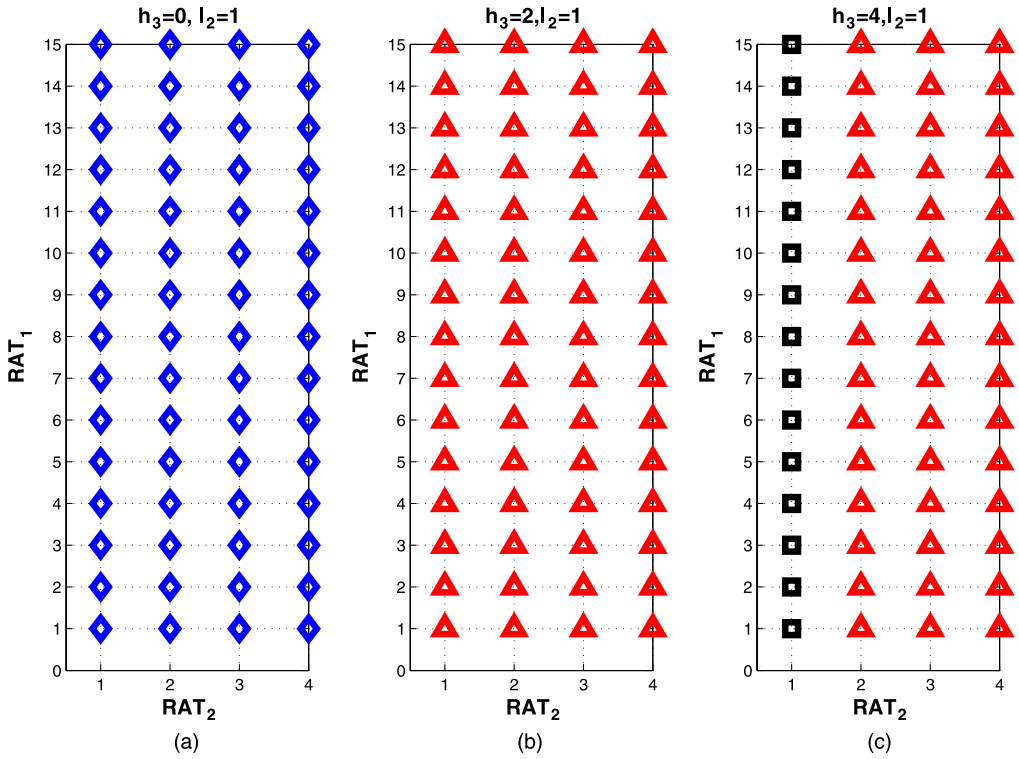

Fig. 14. Call completion in region A_2 using RAT_2 radio resources.

the decision is to empty RAT_2. However, when it is full, and RAT_2 is experiencing a light traffic load, then RAT_3 is selected by the optimal policy. However, as the load in RAT_2 grows, it becomes the best option again.

REFERENCES

[1] Statistical Highlights, 2012. [Online]. Available: http://www.itu.int/ITU-D/ict/statistics/

[2] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, Jun. 2011.

[3] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, 4th Quart. 2011.

[4] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 30–38, Oct. 2011.

[5] J. Lorincz, A. Capone, and D. Begui, "Optimized network management for energy savings of wireless access networks," *Comput. Netw.*, vol. 55, no. 3, pp. 514–540, Feb. 2011.

[6] M. Ismail and W. Zhuang, "Network cooperation for energy saving in green radio communications," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 76–81, Oct. 2011.

[7] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 34–44, Jun. 2013.

[8] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Improvement of RRM Across RNS and RNS/BBS, Release 5*, 3GPP TR25.881 V5.0.0.(2001-12), 2002.

[9] L. Saker, S. E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 664–672, Apr. 2012.

[10] S. Morosi, E. D. Re, and P. Piunti, "Traffic based energy saving strategies for green cellular networks," in *Proc. 18th Eur. Wireless Conf.*, Apr. 2012, pp. 1–6.

[11] L. M. Correia *et al.*, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66–72, Nov. 2010.

[12] P. Kolios, V. Friderikos, and K. Papadaki, "Switching off low utilization base stations via store carry and forward relaying," in *Proc. IEEE 21st Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2010, pp. 312–316.

[13] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–5.

[14] B. Song, S. Das, F. Akashi, C. Chevallier, and S. Soliman, "Network scaling for achieving energy efficient cellular networks—A quantitative analysis," in *Proc. IEEE VTC Fall*, Sep. 2011, pp. 1–5.

[15] A. Bousia, A. Antonopoulos, L. Alonso, and C. Verikoukis, "'Green' distance-aware base station sleeping algorithm in LTE-advanced," in *Proc. IEEE ICC*, Jun. 2012, pp. 1347–1351.

[16] A. Bousia, E. Kartsakli, L. Alonso, and C. Verikoukis, "Energy efficient base station maximization switch off scheme for LTE-advanced," in *Proc. IEEE 17th Int. Workshop CAMAD Commun. Links Netw.*, Sep. 2012, pp. 256–260.

[17] N. Prasad, M. Arslan, and S. Rangarajan, "Exploiting cell dormancy and load balancing in LTE HetNets: Optimizing the proportional fairness utility," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3706–3722, Oct. 2014.

[18] I. Siomina and D. Yuan, "Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization," in *Proc. IEEE ICC*, Jun. 2012, pp. 1357–1361.

[19] T. Zhou, Y. Huang, L. Fan, and L. Yang, "Load-aware user association with quality of service support in heterogeneous cellular networks," *IET Commun.*, vol. 9, no. 4, pp. 494–500, Mar. 2015.

[20] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Trans. Netw.*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[21] M. Boujelben, S. B. Rejeb, and S. Tabbane, "A novel self-organizing scheme for 4G advanced networks and beyond," in *Proc. Int. Symp. Netw., Comput. Commun.*, Jun. 2014, pp. 1–5.

[22] P. Osti, S. Aalto, and P. Lassila, "Load balancing for M2M random access in LTE HetNets," in *Proc. IEEE 22nd Int. Symp. MASCOTS*, Sep. 2014, pp. 132–141.

[23] A. Sarma, S. Chakraborty, and S. Nandi, "Deciding handover points based on context aware load balancing in a WiFi-WiMAX heterogeneous network environment," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 348–357, Jan. 2016.

[24] A. Khalid, P. Lassila, and S. Aalto, "Load balancing of elastic data traffic in heterogeneous wireless networks," in *Proc. 25th ITC*, Sep. 2013, pp. 1–9.

[25] G. Aldabbagh *et al.*, "Distributed dynamic load balancing in a heterogeneous network using LTE and TV white spaces," *Wireless Netw.*, vol. 21, no. 7, pp. 2413–2424, Oct. 2015.
[26] W. Shen and Q.-A. Zeng, "Cost-function-based network selection strategy in integrated wireless and mobile networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3778–3788, Nov. 2008.
[27] A. Sgora and D. D. Vergados, "Handoff prioritization and decision schemes in wireless cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 4, pp. 57–77, 4th Quart. 2009.
[28] Q.-A. Zeng and D. P. Agrawal, "Modeling and efficient handling of handoffs in integrated wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, pp. 1469–1478, Nov. 2002.
[29] 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.
[30] S. Tang and W. Li, "An adaptive bandwidth allocation scheme with preemptive priority for integrated voice/data mobile networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 10, pp. 2874–2886, Oct. 2006.
[31] R. Fantacci, "Performance evaluation of prioritized handoff schemes in mobile cellular networks," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 485–493, Mar. 2000.
[32] L. B. Le, D. Niyato, E. Hossain, D. I. Kim, and D. T. Hoang, "QoS-aware and energy-efficient resource management in OFDMA femtocells," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 180–194, Jan. 2013.
[33] X. Gelabert, O. Sallent, J. P. Romero, and R. Agusti, "Performance evaluation of radio access selection strategies in constrained multi-access/multi-service wireless networks," *Comput. Netw.*, vol. 55, no. 1, pp. 173–192, Jan. 2011.
[34] S. S. Rappaport and L.-R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis," *Proc. IEEE*, vol. 82, no. 9, pp. 1383–1397, Sep. 1994.
[35] T.-L. Sheun and W.-F. Wei, "A channel preemption model for vertical handoff in a WLAN-embedded cellular network," *Wireless Netw.*, vol. 16, no. 4, pp. 929–941, May 2010.
[36] H. C. Tijms, *A First Course in Stochastic Models*. Hoboken, NJ, USA: Wiley, 2003.
[37] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Hoboken, NJ, USA: Wiley, 2006.
[38] W.-B. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 4, pp. 654–664, May 1994.
[39] F. R. Yu, V. W. S. Wong, and V. C. M. Leung, "A new QoS provisioning method for adaptive multimedia in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1899–1909, May 2008.
[40] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Cross-layer design for video transmissions in metro passenger information systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 1171–1181, Mar. 2011.

Glaucio H. S. Carvalho received the Ph.D. degree in electrical engineering from the Federal University of Pará (UFPA), Pará, Brazil, in 2005.

He is currently a Professor with the Faculty of Computing, Federal University of Pará. From 2013 to 2014, he was a Postdoctoral Fellow and an Instructor with the Department of Computer Science, Ryerson University, Toronto, ON, Canada. His research interests include wireless and mobile networks, green protocols, cloud and mobile cloud computing, software-defined networks, queueing theory, Markov decision processes, and operational research.

Dr. Carvalho served as an Associate Editor for Elsevier's *Computers and Electrical Engineering* (CAEE) from 2010 to 2015, where he was a Top Associate Editor in 2011. He was a Guest Editor for the CAEE Special Issue on the *Design and Analysis of Wireless Systems: New Inspirations*.

Isaac Woungang received the M.Sc. degree in mathematics from the Université de la Méditerranée Aix-Marseille II, Marseille, France, in 1990; the Ph.D. degree in mathematics from the Université du Sud, Toulon, France, in 1994; and the M.Sc. degree from the National Instutute of Scientific Research—Energy, Materials and Telecommunications (INRS-EMT), University of Quebec, Montreal, QC, Canada, in 1999.

From 1999 to 2002, he was a Senior Software Engineer with Nortel Networks, Ottawa, ON, Canada. Since 2002, he has been with Ryerson University, Toronto, ON, where he is currently a Professor of computer science and the Director of the Distributed Applications and Broadband (DABNEL) Laboratory. He has published eight books and over 80 refereed technical articles in scholarly international journals and proceedings of international conferences. His current research interests include radio resource management in next-generation wireless networks, network security, and cloud computing.

Dr. Woungang has served as an Associate Editor for *Computers and Electrical Engineering* (Elsevier) and the *International Journal of Communication Systems* (Wiley). He has Guest Edited several Special Issues of various reputed journals, such as *IET Information Security*, *Mathematical and Computer Modeling* (Elsevier), *Computer Communications* (Elsevier), *Computers and Electrical Engineering* (Elsevier), and *Telecommunication Systems* (Springer). Since January 2012, he has been the Chair of the Computer Chapter, IEEE Toronto Section.

Alagan Anpalagan (SM'04) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto, Toronto, ON, Canada.

In 2001, he joined the Department of Electrical and Computer Engineering, Ryerson University, Toronto, where he was promoted to Full Professor in 2010. He served the department as the Graduate Program Director (2004–2009) and the Interim Electrical Engineering Program Director (2009–2010). He directs a research group working on radio resource management and radio access and networking areas within the WINCORE Lab. During his sabbatical (2010–2011), he was a Visiting Professor with the Asian Institute of Technology, Bangkok, Thailand, and a Visiting Researcher with Kyoto University, Kyoto, Japan. His industrial experience includes working at Bell Mobility, Nortel Networks, and IBM Canada. He has coauthored three edited books, namely, *Design and Deployment of Small Cell Networks* (Cambridge University Press, 2015), *Routing in Opportunistic Networks* (Springer, 2013), and *Handbook on Green Information and Communication Systems* (Academic, 2012). His current research interests include cognitive radio resource allocation and management, wireless cross-layer design and optimization, cooperative communication, machine-to-machine communication, small-cell networks, and green communications technologies.

Dr. Anpalagan has served as an Associate Editor for IEEE COMMUNICATIONS SURVEYS AND TUTORIALS since 2012 and Springer *Wireless Personal Communications* since 2009. He was an Associate Editor for IEEE COMMUNICATIONS LETTERS during 2010–2013 and an Editor of the EURASIP *Journal of Wireless Communications and Networking* during 2004–2009. He also served as a Guest Editor for two EURASIP Special Issues on Radio Resource Management in 3G+ Systems in 2006 and Fairness in Radio Resource Management for Wireless Networks in 2008, as well as MONET Special Issues on Green Cognitive and Cooperative Communication and Networking in 2012. He served as the Technical Program Committee Cochair for the 2012 IEEE International Symposium on Wireless Personal Multimedia Communications Wireless Networks; the 2011 IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications Cognitive Radio and Spectrum Management; the 2011 IEEE International Wireless Communications and Mobile Computing Conference Workshop on Cooperative and Cognitive Networks; the 2004/2008 IEEE Canadian Conference on Electrical and Computer Engineering; and the Wireless Com 2005 Symposium on Radio Resource Management. He served as the IEEE Canada Central Area Chair during 2013–2014, the IEEE Toronto Section Chair during 2006–2007, the ComSoc Toronto Chapter Chair during 2004–2005, and the IEEE Canada Professional Activities Committee Chair during 2009–2011. He received the Deans Teaching Award in 2011; the Faculty Scholastic, Research, and Creativity Award in 2010 and 2013; and the Faculty Service Award in 2010 from Ryerson University. He has completed a course on Project Management for Scientists and Engineers from the University of Oxford Continuing Professional Development Center. He is a Registered Professional Engineer in the province of Ontario and Fellow of the Institution of Engineering and Technology.

Ekram Hossain (F'15) received the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2001.

He is a Professor with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada. He has authored/edited several books in his areas of interest. His current research interests include the design, analysis, and optimization of wireless/mobile communications networks; cognitive radio systems; and network economics.

Dr. Hossain was elected IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks." He currently serves as the Editor-in-Chief for the *IEEE Communications Surveys and Tutorials* and as an Editor for the *IEEE Wireless Communications*. He is a member of the IEEE Press Editorial Board. Previously, he served as the Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in the area of "Resource Management and Multiple Access" from 2009 to 2011, as an Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING from 2007 to 2012, and an Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS—COGNITIVE RADIO SERIES from 2011 to 2014. He has received several research awards, including the University of Manitoba Merit Award in 2010 and 2014 (for Research and Scholarly Activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the 2012 IEEE Wireless Communications and Networking Conference Best Paper Award. He is a Distinguished Lecturer of the IEEE Communications Society (2012–2015). He is a registered Professional Engineer in the province of Manitoba.