Decentralized Radio Resource Allocation for Single-network and Multi-homing Services in Cooperative Heterogeneous Wireless Access Medium

Muhammad Ismail, *Student Member, IEEE,* and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—This paper studies radio resource allocation for mobile terminals (MTs) in a heterogeneous wireless access medium. Unlike the existing solutions in literature, we consider the simultaneous presence of both single-network and multi-homing services in the networking environment. In single-network services, an MT is assigned to the best wireless access network available at its location. On the other hand, in multi-homing services, an MT utilizes all available wireless access networks simultaneously. The objective of the radio resource allocation is of twofold: to determine the optimal assignment of MTs with single-network service to the available wireless access networks, and to find the corresponding optimal bandwidth allocation to the MTs with single-network and multi-homing services. We develop a sub-optimal decentralized implementation of the radio resource allocation, which relies on network cooperation to perform the allocation in a dynamic environment in an efficient manner. The MT plays an active role in the resource allocation operation, whether by selecting the best available wireless network for single-network services or by determining the required bandwidth share from each available network for multi-homing services. Simulation results are presented to demonstrate the performance of the proposed algorithm.**

***Index Terms*—Heterogeneous wireless access networks, resource allocation, single-network service, multi-homing service.**

I. Introduction

Currently, the wireless communication network is a heterogeneous environment with different wireless networks that offer a variety of access options. These wireless access networks include the cellular networks, the IEEE 802.16 wireless metropolitan area networks (WMANs), and the IEEE 802.11 wireless local area networks (WLANs). It is expected that these networks will continue to coexist due to their complementary service capabilities in terms of bandwidth, coverage area, and cost [2]. Hence, in such a networking environment with overlapped coverage from different networks, network integration will lead to better service quality to mobile users and enhanced performance for the networks [2], [3]. As a result, it is essential to develop new radio resource management mechanisms for bandwidth allocation and call admission control in order to satisfy the required quality-of-service (QoS) by the mobile users via different available wireless access networks and to make efficient utilization of the available resources from these networks.

In literature, several works have studied radio resource allocation in a heterogeneous wireless access medium. Two types of services can be distinguished for the resource allocation in this networking environment. The first service type, referred to as single-network service, includes the solutions where an MT is assigned to the best wireless access network available at its location and obtains its required bandwidth from that network. The second service type, referred to as multi-homing service, includes the solutions where an MT obtains its required bandwidth for a certain application from all wireless access networks available at its location using its multi-homing capability. However, these two service types are treated separately in literature. It is envisioned that these two service types will coexist. As a result, it is necessary to develop a radio resource allocation algorithm with such a consideration.

In this paper, the radio resource allocation problem for MTs in a heterogeneous wireless access medium is studied. The contributions of this paper are summarized in the following.

- We develop a centralized optimal resource allocation (CORA) algorithm which takes account of both single-network and multi-homing services. The objective of the CORA algorithm is to find the optimal network assignment for MTs with single-network services and to determine the corresponding optimal bandwidth allocation for MTs with single-network and multi-homing services;
- We develop a decentralized sub-optimal resource allocation (DSRA) algorithm, which is desirable when different networks are operated by different service providers. The MTs play an active role in the resource allocation operation, whether by selecting the best available wireless network for single-network services or by determining the required bandwidth share from each available network for multi-homing services;
- We evaluate and compare the performance of the DSRA and CORA algorithms and study performance trade-offs, based on computer simulations.

The rest of the paper is organized as follows: Section II reviews the related work. In Section III, the system model is presented. In Section IV, the radio resource allocation problem

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada, e-mail:{m6ismail, wzhuang}@uwaterloo.ca.

This work is presented in part at IEEE VTC 2012 [1].

TABLE I
SUMMARY OF IMPORTANT SYMBOLS

Symbol	Definition
A	Network assignment vector for MTs with single-network service
$B_m^{\min}/B_m^{\max}$	Minimum/maximum required bandwidth of MT m
b_{nms}	Allocated bandwidth from network n to MT m through BS/AP s
C_n	Transmission capacity of network n BSs/APs
C_{lvk}	Maximum number of calls of service type v and service class l which can be supported in service area k for a given network subscribers
f_{lks}^{j+1}	Maximum number of single-network calls with service class l in service area k which can be supported by BS/AP s for a given network subscribers during T_{j+1} for a given B^{j+1}
fb_{lks}	A flag bit to indicate whether or not an incoming call of a given network subscribers with single-network service and service class l in service area k can be admitted by BS/AP s
$\mathcal{K}$	Set of service areas in the geographical region
$\mathcal{L}_v$	Set of service classes for service type v
$\mathcal{M}$	Set of MTs in the geographical region
$\mathcal{M}_{ns}$	Set of MTs in the coverage area of network n BS/AP s
$\mathcal{M}_{vk}$	Set of MTs with service type v in service area k
$\vec{\mathcal{M}}_{lvk}^{j+1}$	Vector of predicted number of calls of service type v and service class l in service area k during period T_{j+1} for a given network subscribers
$M_{lvk}(t_a^j)$	Number of existing calls of service type v and service class l in service area k at time instant t_a^j for a given network subscribers
$\widetilde{M}_{lvk}(T_{j+1})$	The maximum predicted number of calls of service type v and service class l in service area k during period T_{j+1} for a given network subscribers
$\mathcal{N}$	Set of available wireless access networks in the geographical region
$\mathcal{N}_k$	Set of wireless access networks available in service area k
p_{nms}	A priority parameter assigned by network n to MT m on its resources in BS/AP s
$\mathcal{S}_n$	Set of BSs/APs of network n in the geographical region
$\mathcal{S}_{nk}$	Set of BSs/APs of network n covering service area k
$\mathcal{S}_k$	Set of BSs/APs from all networks covering service area k
$\vec{\mathcal{T}}_{lvk}^{j}$	Time vector of arrival events for calls of service type v and service class l in service area k for a given network subscribers during period T_j
T_c^{lv}	Time duration of a video call that belongs to service type v and service class l
T_r^k	User residence time in service area k
T_h^{lvk}	Channel holding time for a video call of service type v and service class l in service area k
x_{nms}	A binary assignment variable of MT m to BS/AP s of network n
υ_{lvk}	Arrival rate of both new and handoff video calls of service type v and service class l in service area k
λ_{ns}	Link access price of network n BS/AP s
$\nu_m^{(1),(2)}$	Largangian multipliers to guarantee that the required QoS of MT m with single-network service is satisfied
$\mu_m^{(1),(2)}$	Largangian multipliers to guarantee that the required QoS of MT m with multi-homing service is satisfied
α_j	Fixed step size, $j \in \{1, 2, 3, 4, 5\}$
ϵ_{lvk}	Upper bound on call blocking probability for service type v and service class l in service area k for a given network subscribers
τ	Prediction duration

is formulated and the CORA algorithm is introduced. Section V presents a sub-optimal decentralized implementation of the radio resource allocation problem and the DSRA algorithm. In Section VII, simulation results and discussions are given. Finally, conclusions are drawn in Section VIII. Table I summarizes the important symbols used in this paper.

II. RELATED WORK

In literature, several works have studied the problem of radio resource allocation in a heterogeneous wireless access medium. Two service types can be distinguished in these works, namely single-network and multi-homing services.

In the single-network service type, an MT is allocated its required bandwidth from the best available wireless access network. The selection of the best wireless access network available at the MT location is based on a predefined criterion, which can be for example the received signal strength (RSS) [4] or the available bandwidth [5]. Further, different criteria such as RSS, available bandwidth, and monetary cost can be combined in a utility function, and the network assignment for the MT is based on the results of this function for the candidate networks' BSs/APs [6]. One limitation of single-network service type is that an incoming call is blocked if no network at its location can individually satisfy the required bandwidth by the call. As a result, the available radio resources from different networks are not efficiently utilized.

An MT can maintain multiple simultaneous associations with different networks using its multi-homing capability. As a result, in a multi-homing service type, the MT obtains its required bandwidth from all networks available at its location using its multi-homing capability. This has the advantage that the available resources from different networks can be aggregated to support applications with a high required data rate using multiple threads at the application layer, thus reducing the call blocking rate. In addition, it allows for mobility support, since at least one of the radio interfaces will remain active during the call duration. Multi-homing radio resource allocation has been studied in [7] using the concept of utility fairness, in [8] using convex optimization formulation for constant bit rate (CBR) service, and in [9] for both CBR and variable bit rate (VBR) services.

In literature, the existing solutions for radio resource allocation in a heterogeneous wireless access medium focus on either a single-network service or a multi-homing service. However, it is very likely that both single-network and multi-homing services will coexist. It is expected that not all MTs will be equipped with multi-homing capabilities, hence, some MTs can only utilize a single-network service. In addition, even for an MT with multi-homing capabilities, the utilization of the multi-homing service should depend on the residual energy at the MT. When there is no sufficient energy available at the MT, the MT should switch to the single-network service where the radio interface of the best available network is utilized while all other interfaces are turned off to save energy. As a result, it is required to develop a radio resource allocation algorithm that can support both single-network and multi-homing services.

A radio resource allocation algorithm which requires a central controller over heterogeneous wireless networks is not practical in a case that these networks are operated by different service providers. A central resource manager that controls the operation of different networks in this case raises some issues [9] related to: Firstly, the question of which network will be in charge of the operation and maintenance of the central resource manager; Secondly, the changes required in different network structures in order to account for the central manager; Finally, the fact that the central resource manager is a single point of failure. Hence, in such a networking environment, it is desirable to have a decentralized solution that enables each BS/AP to perform its own resource allocation and admission control while at the same time to cooperate with available BSs/APs of other networks, to support MTs with single-network and multi-homing services. While [9] presents a decentralized resource allocation algorithm in a heterogeneous wireless access medium, it mainly targets multi-homing resource allocation and studies a static system model without new call arrivals or departures of existing calls in different service areas.

In this paper, a decentralized sub-optimal algorithm is proposed for resource allocation in a heterogeneous wireless access medium supporting both single-network and multi-homing services. Towards this end, we first develop an optimal centralized radio resource allocation algorithm (CORA algorithm). Based on the centralized algorithm, call traffic load prediction [10], and network cooperation, we then propose the decentralized sub-optimal algorithm (DSRA algorithm). The DSRA algorithm accounts for the system dynamics, in terms of call arrivals and departures, in order to achieve an acceptable call blocking probability and provide a sufficient amount of allocated resources to each call.

III. System Model

A. Wireless Access Networks

Consider a geographical region with a set, $\mathcal{N} = \{1, 2, \ldots, N\}$, of different available wireless access networks. Each network, $n \in \mathcal{N}$, is operated by a unique service provider and has a set, $\mathcal{S}_n = \{1, 2, \ldots, S_n\}$, of BSs/APs in the geographical region. The BSs/APs of each network, $n \in \mathcal{N}$, have different coverage areas from those of other networks. The BSs/APs of different networks have overlapped coverage in some areas. As a result, the geographical region is partitioned to a set, $\mathcal{K} = \{1, 2, \ldots, K\}$, of service areas. Each service area, $k \in \mathcal{K}$, is covered by a unique subset of BSs/APs from all networks, as shown in Figure 1. The subset of available networks at service area k is denoted by $\mathcal{N}_k$, and the subset of BSs/APs from network n covering service area k is denoted by $\mathcal{S}_{nk}$. The BSs/APs from all networks covering service area k are given in the subset $\mathcal{S}_k$ with cardinality $|\mathcal{S}_k|$. Each BS/AP, $s \in \mathcal{S}_n$, has a downlink transmission capacity C_n Mbps. An identification (ID) beacon is broadcasted by each BS/AP, which is used in the MT attachment procedure [11]. It is assumed that different networks are already connected through a backbone to exchange their roaming signalling information. We rely on the roaming signalling backbone in order to exchange the signalling information required by our proposed DSRA algorithm.

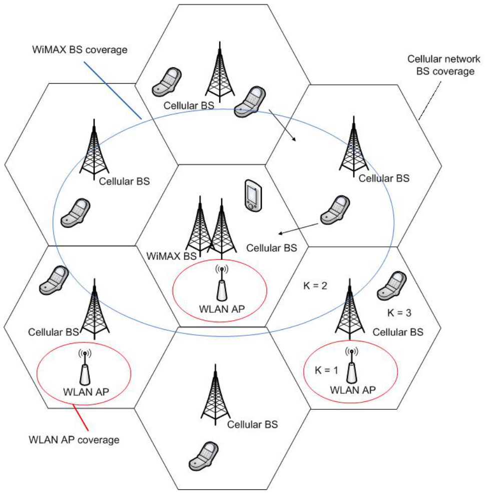

Fig. 1. The network coverage areas.

B. Service Types

The set of MTs in the geographical region is denoted by $\mathcal{M}$. The subset of MTs in a given service area, k, is given by $\mathcal{M}_k$. Each MT, $m \in \mathcal{M}$, has its own home network, but can also get service from other networks available at its location. An MT, m, using its own home network, n, is referred to as a network subscriber, while an MT using a network other than its home network is referred to as a network user. A priority parameter p_{nms} is used to represent service priority of network n in allocating its resources to MT m via BS/AP s, where $p_{nms} = 1$ for high-priority network subscribers and $p_{nms} \in [0, 1)$ for low-priority network users [9]. Two service types are considered, namely single-network and multi-homing services. The subset of MTs with same service type in a given service area k is denoted by $\mathcal{M}_{vk}$, where $v = 1$ for single-network service and $v = 2$ for multi-homing service. An MT, $m \in \mathcal{M}_{1k}$ in service area k, is assigned to a single network n BS/AP $s \in \mathcal{S}_{nk}$. The network assignment criterion is based on the available bandwidth for the MT. The network assignment vector, A, in the geographical region for MTs with single-network service is given by $A = [a_1, \ldots, a_m, \ldots, a_{|\mathcal{M}_{1k}|}]$, where $a_m = ns$ is the assignment of MT $m \in \mathcal{M}_{1k}$ to network n BS/AP s. For instance, $a_1 = 12$ is the assignment of MT 1 to network 1 BS/AP 2. On the other hand, an MT $m \in \mathcal{M}_{2k}$

in a given service area k, receives its required bandwidth from all BSs/APs, $s \in \mathcal{S}_k$, using its multi-homing capability. The set of MTs assigned to network n BS/AP s, including both multi-homing and single-network MTs, is denoted by $\mathcal{M}_{ns}$.

C. Service Traffic Models

Consider a downlink scenario, with video service applications such as on-demand video streaming. A video call to MT m is considered to be a VBR service that is allocated a total bandwidth of B_m in the range $[B_m^{\min}, B_m^{\max}]$, where $B_m^{\min}$ is the total minimum required bandwidth by MT m which guarantees a minimum QoS requirement for the video call, and $B_m^{\max}$ is the total maximum required bandwidth by MT m which is enforced to incorporate the MTs technical limitations. The more allocated bandwidth to a video call, the higher the perceived video quality experienced on the MT.

For each service type, v, there exists a set, $\mathcal{L}_v = \{1, 2, \ldots, L_v\}$, of service classes. Each service class, l_v, has unique $B_{lv}^{\min}$ and $B_{lv}^{\max}$ values. In general, class $\mathcal{L}_2$ for an MT with multi-homing service type requires larger bandwidth than class $\mathcal{L}_1$ for an MT with single-network service. The total allocated bandwidth to MT m with a VBR call of service type v and service class l is B_{lv}. The allocated bandwidth from network n to MT m via BS/AP s is denoted by b_{nms}. Let $B = [b_{nms}]$ be a matrix of allocated bandwidth from network $n \in \mathcal{N}$ to MT $m \in \mathcal{M}$ through BS/AP $s \in \mathcal{S}_n$, where $b_{nms} = 0$ if MT $m \notin \mathcal{M}_{ns}$ and for single-network MT if $a_m \neq ns$. For a given network subscribers, the number of existing calls of service type v and service class l in service area k is denoted by M_{lvk}. The maximum number of calls of each service type v and service class l which can be supported in each service area k for a given network subscribers, given the transmission capacities of available BSs/APs, is denoted by C_{lvk}. This maximum number of calls can be determined using a capacity analysis similar to that in [12]. A call admission control procedure is in place, which guarantees that $M_{lvk} \leq C_{lvk}$, such that feasible resource allocation solutions exist. It is worth mentioning that, although the proposed algorithm studies radio resource allocation on the downlink, it can be applied for radio resource allocation on the uplink.

The arrivals of video calls are modeled as a Poisson process, which is a widely adopted assumption [12]. In particular, the arrival process of both new and handoff calls of service type v and class l to service area k is modeled by a Poisson process with parameter v_{lvk}. According to statistics of on-demand video streaming [13], [14], the video call duration is very likely to be heavy-tailed. For analysis tractability, it is proposed to fit a large class of heavy-tailed distributions using hyper-exponential distributions [15], such as a two-stage hyper-exponential distribution [12]. For a video call of service type v and class l, the probability density function (PDF) of the call duration, T_c^{lv}, with mean $\bar{T}_c^{lv}$, is given by [12]

$$f_{T_c^{lv}}(t) = \frac{a_{lv}}{a_{lv}+1} \cdot \frac{a_{lv}}{\bar{T}_c^{lv}} \cdot e^{-\frac{a_{lv}}{\bar{T}_c^{lv}}t} + \frac{1}{a_{lv}+1} \cdot \frac{1}{a_{lv}\bar{T}_c^{lv}} \cdot e^{-\frac{1}{a_{lv}\bar{T}_c^{lv}}t}, \qquad a_{lv} \geq 1, t \geq 0. \quad (1)$$

D. Mobility Models and Channel Holding Time

User mobility within a given service area $k \in \mathcal{K}$ is characterized by the user residence time, denoted by T_r^k, which is assumed to follow an exponential distribution with mean $\bar{T}_r^k$. The channel holding time for a given service type v with service class l in service area k is given by $T_h^{lvk} = \min(T_c^{lv}, T_r^k)$, where T_c^{lv} and T_r^k are independent of each other. It can be easily derived that the PDF of T_h^{lvk} is

$$f_{T_h^{lvk}}(t) = \frac{a_{lv}}{a_{lv}+1} \cdot (\frac{1}{\bar{T}_r^k} + \frac{a_{lv}}{\bar{T}_c^{lv}}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{a_{lv}}{\bar{T}_c^{lv}})t} + \frac{1}{a_{lv}+1} \cdot (\frac{1}{\bar{T}_r^k} + \frac{1}{a_{lv}\bar{T}_c^{lv}}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{1}{a_{lv}\bar{T}_c^{lv}})t}, \ t \geq 0. \ (2)$$

IV. Centralized Optimal Resource Allocation (CORA)

In this section, the problem formulation of radio resource allocation (call-level bandwidth reservation) for MTs with single-network and multi-homing services in the heterogeneous wireless access medium is presented. A centralized optimal resource allocation (CORA) algorithm is proposed based on the problem formulation.

The utility of network n allocating bandwidth b_{nms} to MT m via BS/AP s is given by [9]

$$u_{nms}(b_{nms}) = \log(1 + \eta_1 b_{nms}) - \eta_2(1 - p_{nms})b_{nms} \quad (3)$$

where η_1 and η_2 are used for scalability of b_{nms} [16]. For a network subscriber, with $p_{nms} = 1$, the utility function of (3) accounts only for the attained network utility by that subscriber, which is represented by the first term in the right hand side (RHS) of (3) [16]. On the other hand, a network user with $p_{nms} \in [0, 1)$ suffers from a trade-off between the attained network utility and the cost that the network sets on its resources. The cost is represented by the second term in the RHS of (3). As a result, each network gives higher priority in allocating its resources to its subscribers as compared to other users [9].

For a given network assignment vector A, the overall resource allocation objective of all networks in the geographical region is to determine the optimal bandwidth allocation b_{nms}, $\forall n \in \mathcal{N}, m \in \mathcal{M}_{ns}, s \in \mathcal{S}_n$ which maximizes the total utility in the region, given by

$$U(b_{nms}) = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \sum_{m \in \mathcal{M}_{ns}} u_{nms}(b_{nms}). \quad (4)$$

The allocated resources from network n BS/AP s should satisfy the capacity constraint of the BS/AP, that is

$$\sum_{m \in \mathcal{M}_{ns}} b_{nms} \leq C_n, \qquad \forall s \in \mathcal{S}_n, n \in \mathcal{N}. \quad (5)$$

For MTs with single-network service, given a network assignment vector A, the allocated resources from the assigned network n BS/AP $s \in \mathcal{S}_{nk}$ to MT $m \in \mathcal{M}_{1k}$ in service area k should satisfy the application required bandwidth, given by

$$B_m^{\min} \leq b_{nms} \leq B_m^{\max}, \qquad \forall m \in \mathcal{M}_{1k}, k \in \mathcal{K}. \quad (6)$$

On the other hand, for MTs with multi-homing service, the total allocated resources from all available BSs/APs in $\mathcal{S}_k$ to

MT $m \in \mathcal{M}_{2k}$ in service area k should satisfy the application required bandwidth, given by

$$B_m^{\min} \le \sum_{n\in\mathcal{N}_k} \sum_{s\in\mathcal{S}_{nk}} b_{nms} \le B_m^{\max}, \qquad \forall m \in \mathcal{M}_{2k}, k \in \mathcal{K}. \tag{7}$$

In order to determine the optimal network assignment vector A and the corresponding optimal bandwidth allocation matrix B for MTs with single-network and multi-homing services, the radio resource allocation problem is expressed by the following optimization problem

$$\begin{aligned} \max_{A}\{\max_{B} \quad & U(b_{nms}) \\ s.t. \quad & (5) - (7)\}. \end{aligned} \tag{8}$$

While the radio resource allocation problem for a given network assignment vector is a convex optimization problem and therefore can be solved efficiently using polynomial time algorithms [17], finding the optimal vector A incurs high computational complexity. In a given service area k with a total of $|\mathcal{S}_k|$ BSs/APs available from different networks and $|\mathcal{M}_{1k}|$ MTs with single-network service, there exist $|\mathcal{S}_k|^{|\mathcal{M}_{1k}|}$ distinct assignment vectors. As a result, in the whole geographical region, the total number of distinct assignment vectors is $\prod_k |\mathcal{S}_k|^{|\mathcal{M}_{1k}|}$. For instance, consider one service area with 3 BSs/APs having overlapped coverage and a total of 50 MTs with single-network service. In this case, there are a total of $3^{50} = 7 * 10^{23}$ distinct network assignments in this service area. Hence, for the whole geographical region, it is expected that the inner maximization problem of (8) needs to be solved for a huge number of times in order to determine the optimal radio resource allocation. As a result, it is desirable to develop a less complex formulation of problem (8). In order to do so, a binary assignment variable x_{nms} is introduced, which is determined from the network assignment vector A for MT $m \in \mathcal{M}_{1k}$ by

$$x_{nms} = \begin{cases} 1, & \text{if } a_m = ns \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

On the other hand, $x_{nms} = 1$ for MTs with multi-homing service in service area k for all $s \in \mathcal{S}_k$. Using the binary assignment variable, the problem of (8) can be reformulated as in (10) where the fourth constraint ensures that an MT with single-network service is assigned to one and only one BS/AP available at its location and the last constraint allows an MT with multi-homing service to obtain its required bandwidth from all wireless networks available at its location. The problem of (10) is a non-convex mixed integer non-linear programming (MINLP) problem. In general, MINLP problems are difficult to solve, since they combine the difficulty of optimizing over integer variables with the handling of non-linear functions. This is especially true when the objective and/or constraint functions are non-convex, which is the case in (10). Recently, several new methods are developed to solve MINLP problems [18], this includes deterministic algorithms [18], [19] and stochastic ones [20]. The different methods of solving MINLP problems are available through many solvers [21]. The BARON solver [22], which is available through GAMS, is a global deterministic solver which can address non-convexities in MINLP problems and provide global optima under fairly general assumptions [22]. Since the BARON solver has proven to be the most robust one among the currently available global solvers [23], we use it to solve the radio resource allocation problem of (10).

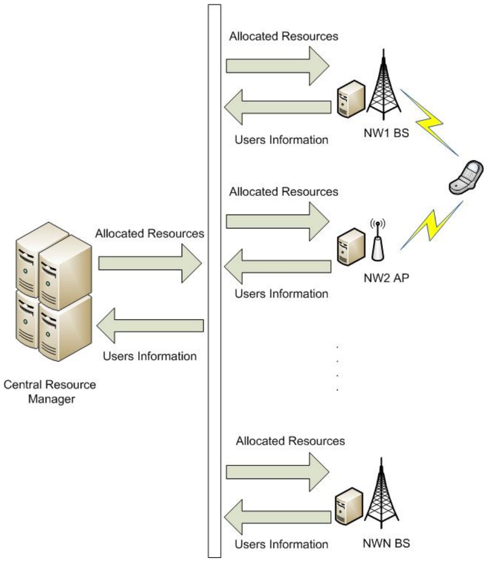

Fig. 2. Centralized implementation of the CORA algorithm.

$$\begin{aligned} \max_{x_{nms}, b_{nms}} \quad & \sum_{n=1}^{N} \sum_{s=1}^{S_n} \sum_{m\in\mathcal{M}_{ns}} \{\log(1 + \eta_1 x_{nms} b_{nms}) \\ & - \eta_2 (1 - p_{nms}) x_{nms} b_{nms}\} \\ s.t. \quad & \sum_{m\in\mathcal{M}_{ns}} x_{nms} b_{nms} \le C_n, \quad \forall s \in \mathcal{S}_n, n \in \mathcal{N} \\ & B_m^{\min} \le \sum_{n=1}^{N} \sum_{s\in\mathcal{S}_k} x_{nms} b_{nms} \le B_m^{\max}, \\ & \qquad \forall m \in \mathcal{M}_k, k \in \mathcal{K} \\ & x_{nms} \in \{0, 1\}, \quad \forall m \in \mathcal{M}_{1k}, k \in \mathcal{K} \\ & \sum_{n=1}^{N} \sum_{s\in\mathcal{S}_k} x_{nms} = 1, \quad \forall m \in \mathcal{M}_{1k}, k \in \mathcal{K} \\ & x_{nms} = 1, \quad \forall m \in \mathcal{M}_{2k}, k \in \mathcal{K}. \end{aligned} \tag{10}$$

A centralized implementation of the radio resource allocation problem (CORA algorithm) based on the formulation of (10) is illustrated in Figure 2. In this implementation, each MT reports to all BSs/APs available at its location its service type, service class, and home network using its multiple radio interfaces. This information is made available to the central resource manager via different BSs/APs. As a result, the central resource manager has the information of the service area k for each MT, MT minimum and maximum required bandwidth, and MT priority parameter. Given the transmission capacities of all the BSs/APs, the central resource manager solves (10) in order to determine network assignment and bandwidth allocations for new incoming MTs with single-network and multi-homing services, updates bandwidth allocations and initiates vertical handovers for existing MTs if necessary.

V. Decentralized Sub-optimal Resource Allocation (DSRA)

In this section, a decentralized sub-optimal resource allocation (DSRA) algorithm is proposed for the radio resource allocation problem which is desirable when different networks are operated by different service providers.

In the CORA algorithm, signalling information is exchanged over the backbone between the central resource manager and the BSs/APs with every call arrival to or departure from any service area k in order to allocate/update resources to incoming/existing MTs. In the DSRA algorithm, the time is partitioned into a set of periods $\mathcal{T} = \{T_1, T_2, \ldots, T_j, \ldots\}$ of constant duration τ. The call traffic load at each BS/AP at current period, T_j, is used to predict the call traffic load during the next period, T_{j+1}. Cooperative BSs/APs, by exchanging their predicted call traffic load information for the next period, can determine the distribution of the total call traffic load in the geographical region for the next period, T_{j+1}. Every BS/AP based on the predicted call traffic load broadcasts a parameter, referred to as a predicted link access price, which enables incoming and existing MTs to perform network selection and bandwidth request without the need for a central resource manager. The reason for employing a discrete time system is that in a dynamic environment, with call arrivals and departures in different service areas, the link access price value would be time varying. In a continuous time system, an incoming MT using the broadcasted link access price values would make a false resource request as the instantaneous link access price value (with the new call) is different from the broadcasted one. Hence, we propose to 1) discretize time, 2) at the beginning of each time slot make a prediction of the number of calls that will be in service during this time slot, 3) based on this number calculate the link access price values, fix them for the time slot and broadcast them. Incoming calls can use the broadcasted link access price values to make their service requests as the values already account for the service requirement of potential new calls during the given time slot. The traffic load prediction is a probabilistic one which ensures that the prediction error is lower than a target value ϵ, and ϵ is chosen based on the target call blocking probability of the system. The DSRA algorithm can be carried out in the following 8 steps.

Step 1: For clarity of presentation, we focus our discussion in steps 1 - 3 on one network subscribers. The same steps hold for other networks subscribers. Consider calls of service type v and class l in service area k. Let $\vec{\mathcal{T}}_{lvk}^{j}$ denote a time vector of call arrival events for calls of service type v and service class l in service area k during period T_j. With a call arrival at $t_a^j \in \vec{\mathcal{T}}_{lvk}^{j}$, $a = \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lvk}^{j}\right|\}$, in period T_j, the number of calls at the time instant, $M_{lvk}(t_a^j)$, is used by the BSs/APs in the service area to probabilistically predict the number of calls at time instant $t_a^j + \tau$ in the next time period T_{j+1}. As a result, we refer to τ as the prediction duration. The predicted number is denoted by $\widetilde{M}_{lvk}(t_a^j + \tau)$. With call arrivals and departures, the number of calls at t, $M_{lvk}(t)$, is a random variable. Using the probability distribution of $M_{lvk}(t_a^j + \tau)$ given $M_{lvk}(t_a^j)$, alternatively we can represent $\widetilde{M}_{lvk}(t_a^j + \tau)$ by a design parameter ϵ_{lvk}, such that

$$Pr[M_{lvk}(t_a^j + \tau) > \widetilde{M}_{lvk}(t_a^j + \tau) | M_{lvk}(t_a^j)] \leq \epsilon_{lvk}, \quad \forall v, l \in \mathcal{L}, k \in \mathcal{K}. \quad (11)$$

The design parameter $\epsilon_{lvk} \in [0, 1]$ denotes the probability that $M_{lvk}(t_a^j + \tau)$ exceeds the predicted number $\widetilde{M}_{lvk}(t_a^j + \tau)$. In order to find $\widetilde{M}_{lvk}(t_a^j + \tau)$, we need to calculate the conditional probability mass function (PMF) of $M_{lvk}(t_a^j + \tau)$ given $M_{lvk}(t_a^j)$, $P_{M_{lvk}(t_a^j+\tau)|M_{lvk}(t_a^j)}(i)$. Since call arrivals follow a Poisson process, the channel holding time follows a general distribution, and all calls are served simultaneously without queuing, the transient distribution of the $M/G/\infty$ model [24] can be used to calculate $P_{M_{lvk}(t_a^j+\tau)|M_{lvk}(t_a^j)}(i)$. First, we make the following definitions assuming a stationary call arrival and departure process:

- p_τ^{lvk} - The probability that a call which is in service area k at time t_a^j is still present in the same service area at time $t_a^j + \tau$;
- q_τ^{lvk} - The probability that a call that arrives in service area k during $(t_a^j, t_a^j + \tau]$ is still present at the same service area at time $t_a^j + \tau$;
- $X_B(\kappa, \alpha)$ - A binomial random variable with parameters κ and α;
- $X_P(\alpha)$ - A Poisson random variable with mean α.

Given the number of calls at time instant t_a^j, $M_{lvk}(t_a^j)$, we have [24]

$$M_{lvk}(t_a^j + \tau) =_d X_B(M_{lvk}(t_a^j), p_\tau^{lvk}) + X_P(v_{lvk}^n \tau q_\tau^{lvk}) \quad (12)$$

where $=_d$ denotes equality in distribution and v_{lvk}^n is the arrival rate of new and handoff calls to network n in service area k. In oder to determine v_{lvk}^n for BS/AP of network n, a BS/AP can count the number of its new call arrivals to service area k (excluding vertical handoff calls as these are not arrivals to service area k) and divide it by the total elapsed time. In (12), the probabilities p_τ^{lvk} and q_τ^{lvk} are given by [24]

$$\begin{aligned} p_\tau^{lvk} &= \frac{1}{E[T_h^{lvk}]} \int_\tau^\infty Pr(T_h^{lvk} > s) ds \\ &= \frac{1}{E[T_h^{lvk}]} \int_\tau^\infty (1 - F_{T_h^{lvk}}(s)) ds \quad (13) \\ q_\tau^{lvk} &= \int_0^\tau \frac{1}{\tau} Pr(T_h^{lvk} > s) ds \\ &= \int_0^\tau \frac{1}{\tau} (1 - F_{T_h^{lvk}}(s)) ds \\ &= \frac{E[T_h^{lvk}]}{\tau} (1 - p_\tau^{lvk}) \quad (14) \end{aligned}$$

where $E[T_h^{lvk}]$ is the average channel holding time and can be calculated from (2) as

$$E[T_h^{lvk}] = \frac{a_{lv}}{a_{lv} + 1} \cdot \frac{1}{\frac{1}{\overline{T}_r^k} + \frac{a_{lv}}{\overline{T}_c^{lv}}} + \frac{1}{a_{lv} + 1} \cdot \frac{1}{\frac{1}{\overline{T}_r^k} + \frac{1}{a_{lv} \overline{T}_c^{lv}}}, \quad \forall v, l \in \mathcal{L}, k \in \mathcal{K} \quad (15)$$

and $F_{T_h^{lvk}}(s) = \int_0^s f_{T_h^{lvk}}(t)dt$ is the CDF of T_h^{lvk}. Using (12) - (14), $P_{M_{lvk}(t_a^j+\tau)|M_{lvk}(t_a^j)}(i)$ can be found, from which $\widetilde{M}_{lvk}(t_a^j + \tau)$ can be calculated using (11) as the minimum integer which satisfies

$$\sum_{i=0}^{\widetilde{M}_{lvk}(t_a^j+\tau)} P_{M_{lvk}(t_a^j+\tau)|M_{lvk}(t_a^j)}(i) \geq (1 - \epsilon_{lvk}), \qquad \forall v, l \in \mathcal{L}, k \in \mathcal{K}. \quad (16)$$

Step 2: The predicted values of $\widetilde{M}_{lvk}(t_a^j+\tau), \forall v, l \in \mathcal{L}, k \in \mathcal{K}$ and $a = \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lvk}^j\right|\}$, are recorded at each BS/AP in service area k in a vector $\vec{\mathcal{M}}_{lvk}^{j+1}$.

Step 3: At the beginning of period T_{j+1}, the maximum predicted number of calls of each service type v and service class l in each service area k during T_{j+1}, $\widetilde{M}_{lvk}(T_{j+1})$, can be found from $\vec{\mathcal{M}}_{lvk}^{j+1}$. That is, $\widetilde{M}_{lvk}(T_{j+1}) = \max(\vec{\mathcal{M}}_{lvk}^{j+1})$ if it is less than or equal to C_{lvk}, otherwise we let $\widetilde{M}_{lvk}(T_{j+1}) = C_{lvk}$. This ensures that for $\widetilde{M}_{lvk}(T_{j+1}) \leq C_{lvk}$

$$Pr[M_{lvk}(t_a^{j+1}) > \widetilde{M}_{lvk}(T_{j+1})] \leq \epsilon_{lvk}, \qquad \forall v, l \in \mathcal{L}, k \in \mathcal{K}, a \in \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lvk}^{j+1}\right|\}. \quad (17)$$

Step 4: The cooperating BSs/APs in the geographical region exchange their information of $\widetilde{M}_{lvk}(T_{j+1})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$. As a result, $\mathcal{M}_{lvk}$, for every l, v, k, and networks subscribers can be determined and problem (10) can be solved at each BS/AP in order to determine the binary assignment variable x_{nms}^{j+1} for all MTs with single-network service in the geographical region during T_{j+1} and the corresponding bandwidth allocation B^{j+1}. Therefore, the network assignment vector A^{j+1} for MTs with single-network service during T_{j+1} can be determined. Based on the network assignment vector A^{j+1}, each BS/AP s can determine the maximum number of single-network calls with service class l in service area k which can be supported by this BS/AP during T_{j+1}, $f_{lks}^{j+1}, \forall l \in \mathcal{L}, k \in \mathcal{K}$, given B^{j+1}.

Step 5: Given the network assignment vector A^{j+1} calculated in step 4, problem (8) is reduced to

$$\begin{aligned} \max_{B} \quad & U(b_{nms}) \\ s.t. \quad & (5) - (7). \end{aligned} \qquad (18)$$

From the utility function definitions in (3) and (4), the objective function of (18) is concave and the problem has linear constraints. Hence, problem (18) is a convex optimization problem, which makes a local maximum a global maximum as well and strong duality holds [17]. Full dual decomposition [25], [26] of (18) is applied, which helps in the decentralized resource allocation as described in the next step. In order to apply full dual decomposition, we first find the Lagrangian function, $L(B, \lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)})$, of (18) [25], where $\lambda = (\lambda_{ns} : n \in \mathcal{N}, s \in \mathcal{S}_n)$ is a matrix of Lagrangian multipliers corresponding to the capacity constraint of (5) with $\lambda_{ns} \geq 0$, $\nu^{(1)} = (\nu_m^{(1)} : m \in \mathcal{M}_{1k}, \forall k \in \mathcal{K})$ and $\nu^{(2)} = (\nu_m^{(2)} : m \in \mathcal{M}_{1k}, \forall k \in \mathcal{K})$ are vectors of Lagrangian multipliers corresponding to the maximum and minimum required bandwidth constraints of (6) for MTs with single-network service with $\nu_m^{(1)}, \nu_m^{(2)} \geq 0$, and $\mu^{(1)} = (\mu_m^{(1)} : m \in \mathcal{M}_{2k}, \forall k \in \mathcal{K})$ and $\mu^{(2)} = (\mu_m^{(2)} : m \in \mathcal{M}_{2k}, \forall k \in \mathcal{K})$ are vectors of Lagrangian multipliers corresponding to the maximum and minimum required bandwidth constraints of (7) for MTs with multi-homing service with $\mu_m^{(1)}, \mu_m^{(2)} \geq 0$. The dual function is given by

$$h(\lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}) = \max_{B} L(B, \lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}) \qquad (19)$$

and the dual problem corresponding to the primal problem of (18) is given by

$$\min_{(\lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}) \geq 0} h(\lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}). \qquad (20)$$

The maximization problem of (19) gives the bandwidth allocation B for fixed value of the Lagrangian multipliers, which can be solved using the Karush-Kuhn-Tucker (KKT) conditions [17], and we have

$$b_{nms} = [(\frac{\eta_1}{\lambda_{ns} + (\nu_m^{(1)} - \nu_m^{(2)}) + \eta_2(1 - p_{nms})} - 1)/\eta_1]^+, \qquad \forall m \in \cup_k \mathcal{M}_{1k} \qquad (21)$$

$$b_{nms} = [(\frac{\eta_1}{\lambda_{ns} + (\mu_m^{(1)} - \mu_m^{(2)}) + \eta_2(1 - p_{nms})} - 1)/\eta_1]^+, \qquad \forall m \in \cup_k \mathcal{M}_{2k} \qquad (22)$$

where $[\cdot]^+$ is a projection on the positive orthant to account for the fact that $b_{nms} \geq 0$. The optimum values of the Lagrangian multipliers which result in the optimum bandwidth allocation can be determined by solving the dual problem of (20). For a differentiable dual function, a gradient descent method can be applied in order to determine the optimum values for the Lagrangian multipliers [17], given by

$$\lambda_{ns}(i+1) = [\lambda_{ns}(i) - \alpha_1(C_n - \sum_{m \in \mathcal{M}_{ns}} b_{nms}(i))]^+ \qquad (23)$$

$$\nu_m^{(1)}(i+1) = [\nu_m^{(1)}(i) - \alpha_2(B_m^{\max} - \sum_{n=1}^{N} \sum_{s=1}^{\mathcal{S}_n} b_{nms}(i))]^+ \qquad (24)$$

$$\nu_m^{(2)}(i+1) = [\nu_m^{(2)}(i) - \alpha_3(\sum_{n=1}^{N} \sum_{s=1}^{\mathcal{S}_n} b_{nms}(i) - B_m^{\min})]^+ \qquad (25)$$

$$\mu_m^{(1)}(i+1) = [\mu_m^{(1)}(i) - \alpha_4(B_m^{\max} - \sum_{n=1}^{N} \sum_{s=1}^{\mathcal{S}_n} b_{nms}(i))]^+ \qquad (26)$$

$$\mu_m^{(2)}(i+1) = [\mu_m^{(2)}(i) - \alpha_5(\sum_{n=1}^{N} \sum_{s=1}^{\mathcal{S}_n} b_{nms}(i) - B_m^{\min})]^+ \qquad (27)$$

where i is the iteration index and α_j with $j = \{1, 2, 3, 4, 5\}$ is a fixed sufficiently small step size. Convergence towards the optimum solution is guaranteed as the gradient of (20) satisfies the Lipchitz continuity condition [17].

The Lagrangian multiplier λ_{ns} serves as an indication of the capacity limitation experienced by network n BS/AP s. When the total call traffic load on network n BS/AP s ($\sum_{m \in \mathcal{M}_{ns}} b_{nms}$) reaches the capacity limitation (C_n), λ_{ns}

increases to denote that it is expensive to use that link and the bandwidth allocation is reduced towards the minimum required bandwidth. Hence, we refer to λ_{ns} as network n BS/AP s link access price. On the other hand, $\nu_m^{(1)}$, $\nu_m^{(2)}$, $\mu_m^{(1)}$, and $\mu_m^{(2)}$ are used to guarantee that the total allocated bandwidth to MT m satisfies its minimum and maximum required bandwidth.

Given the predicted maximum number of calls during T_{j+1}, $\widetilde{M}_{lvk}(T_{j+1})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$, each BS/AP can determine its predicted link access price value $\widetilde{\lambda}_{ns}^{j+1}$ using the BARON solver while solving (10) at the beginning of T_{j+1} using $\widetilde{M}_{lvk}(T_{j+1})$.

Step 6: Each BS/AP updates its link access price value with $\widetilde{\lambda}_{ns}^{j+1}$ at the beginning of T_{j+1} and this value is fixed over T_{j+1}, independent of call arrivals to and departures from different service areas, and is broadcasted on the BS/AP ID beacon. Moreover, a flag bit, fb_{lks}, is set to 1 if $M_{lvk} < f_{lks}$ and is broadcasted by each BS/AP s on its ID beacon to denote that a new incoming call with single-network service and service class l in service area k can be admitted by the BS/AP. Otherwise, $fb_{lks} = 0$.

The fixed link access price values, $\widetilde{\lambda}_{ns}^{j+1}$ $\forall n \in \mathcal{N}, s \in \mathcal{S}_n$ which are broadcasted during T_{j+1}, distribute the radio resources of all networks exactly over the maximum predicted number of calls $\widetilde{M}_{lvk}(T_{j+1})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$. Hence, during T_{j+1}, when $M_{lvk} = \widetilde{M}_{lvk}(T_{j+1})$, any incoming call of service type v with service class l in service area k will be blocked. As a result, from (11) and (17), ϵ_{lvk} is the upper bound of the call blocking probability, given that $\widetilde{M}_{lvk}(T_{j+1}) \leq C_{lvk}$. Otherwise, $\widetilde{M}_{lvk}(T_{j+1}) = C_{lvk}$, and both the CORA and DSRA algorithms achieve the same call blocking probability.

Step 7: An incoming MT to service area k during T_{j+1} obtains the link access price values $\widetilde{\lambda}_{ns}^{j+1}$ $\forall n \in \mathcal{N}, s \in \mathcal{S}_n$ via its multiple radio interfaces. The MT then performs the following based on its service type.

First, consider MTs with single-network service. An MT uses the link access price values to solve for the allocated bandwidth from each BS/AP available at its location with $fb_{lks} = 1$. This can be done at MT, m, of a call of service class l in service area k, using Algorithm V.1, where I is the number of iterations required for the algorithm to converge to the required bandwidth allocation. The MT orders the available BSs/APs based on the calculated bandwidth allocation from maximum to minimum. The MT then asks the BS/AP with the maximum calculated bandwidth allocation for the b_{nms} resource allocation. The BS/AP provides the required bandwidth if it has sufficient resources. Otherwise, the incoming call is blocked. For MTs which are already in service, the $\widetilde{\lambda}_{ns}^{j+1}$ values for $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$ with $fb_{lks} = 1$, are used at the beginning of T_{j+1} in a similar way as described before in order to perform a vertical handover if necessary.

Algorithm V.1 Calculation of Bandwidth Allocation from Each Available Network BS/AP at MT m with Single-network Service

```
Input: $\widetilde{\lambda}_{ns}^{j+1}$ $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, $B_m$;
Initialization: $i \longleftarrow 0$; $\nu_m^{(1)}(0) \geq 0$; $\nu_m^{(2)}(0) \geq 0$;
for $n \in \mathcal{N}_k$ do
    for $s \in \mathcal{S}_{nk}$ do
        for $i = 1 : I$ do
            $b_{nms}(i) = [(\frac{\eta_1}{\widetilde{\lambda}_{ns}^{j+1} + (\nu_m^{(1)}(i) - \nu_m^{(2)}(i)) + \eta_2(1 - p_{nms})} - 1)/\eta_1]^+$;
            $\nu_m^{(1)}(i+1) = [\nu_m^{(1)}(i) - \alpha_2(B_m^{\max} - b_{nms}(i))]^+$;
            $\nu_m^{(2)}(i+1) = [\nu_m^{(2)}(i) - \alpha_3(b_{nms}(i) - B_m^{\min})]^+$;
        end for
    end for
end for
Output: $b_{nms}$ $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$.
```

Next, consider MTs with multi-homing services. During T_{j+1}, each MT in the geographical region, including both incoming and already existing ones, uses the broadcasted link access price values received at its location to determine the required bandwidth share from each BS/AP, such that the total amount of resources allocated from all the BSs/APs satisfies its required bandwidth. This is performed at MT, m, with service class l in service area k using Algorithm V.2. The MT asks for the required bandwidth share b_{nms} from BS/AP s of network n $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, which allocates the required bandwidth if it has sufficient resources. The incoming call is blocked if the total allocated resources do not satisfy its required bandwidth.

Algorithm V.2 Calculation of Bandwdith Share from Each Available Network BS/AP at MT m with Multi-homing Service

```
Input: $\widetilde{\lambda}_{ns}^{j+1}$ $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, $B_m$;
Initialization: $i \longleftarrow 0$; $\mu_m^{(1)}(0) \geq 0$; $\mu_m^{(2)}(0) \geq 0$;
for $i = 1 : I$ do
    for $n \in \mathcal{N}_k$ do
        for $s \in \mathcal{S}_{nk}$ do
            $b_{nms}(i) = [(\frac{\eta_1}{\widetilde{\lambda}_{ns}^{j+1} + (\mu_m^{(1)}(i) - \mu_m^{(2)}(i)) + \eta_2(1 - p_{nms})} - 1)/\eta_1]^+$;
        end for
    end for
    $\mu_m^{(1)}(i+1) = [\mu_m^{(1)}(i) - \alpha_4(B_m^{\max} - \sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(i))]^+$;
    $\mu_m^{(2)}(i+1) = [\mu_m^{(2)}(i) - \alpha_5(\sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(i) - B_m^{\min})]^+$;
end for
Output: $b_{nms}$ $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$.
```

Step 8: Each MT reports to its serving BSs/APs its service type, service class, home network, and a list of the BS/AP IDs that the MT can receive signal from. This information is used by BSs/APs to predict $\widetilde{M}_{lvk}(T_{j+2})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$, during the next period T_{j+2} in order to update their link access price values at the beginning of T_{j+2}.

The link access price value for different networks' BSs/APs are updated every τ. As a result, the choice of the τ duration should reflect some change in the call traffic load in the geographical region. Let δ_{lvk} be the minimum of durations to the arrival of a new call and to the departure of an existing call for service class l with service type v in service area k. Define δ= $\min(\delta_{lvk})$ $\forall l, v, k$. As a guideline, the time duration τ is chosen such that the probability$Pr[\delta < \tau]$ is less than a small threshold γ.

VI. Simulation Results and Discussion

This section presents simulation results for the resource allocation in a heterogeneous wireless access medium for MTs with single-network and multi-homing services. Consider a geographical region which is entirely covered by an IEEE 802.16e WMAN BS and partially covered by a 4G cellular network BS and IEEE 802.11b WLAN AP. Therefore, $\mathcal{N} = \{1, 2, 3\}$, with the WMAN, cellular network, and WLAN indexed as 1, 2, and 3 respectively. Each network has one BS/AP in the geographical region, i.e. $s = 1$ for all the networks. Three service areas can be distinguished. One service area ($k = 1$) is covered only by the WMAN BS, another ($k = 2$) is covered by both the WMAN and cellular network BSs, and the last one ($k = 3$) is covered by all three networks. We consider a single service class $l = 1$ for each service type (single-network and multi-homing) and study the performance of the proposed algorithms in the service area that is covered by the WMAN and cellular network BSs ($k = 2$) in terms of the allocated resources per call and the call blocking probability. Due to space limitation, we only show the results of resource allocation for the cellular network subscribers. For simplicity, we consider a complete partitioning strategy for each network BS transmission capacity [27], where the total bandwidth of each BS is divided into two separate parts, dedicating to single-network and multi-homing services respectively[1]. The allocated capacity from the network n BS/AP to the service area under consideration for cellular network subscribers with service type v, C_{nv}, is given by $C_{11} = 1.344$ Mbps, $C_{12} = 2.864$ Mbps, $C_{21} = 0.576$ Mbps, and $C_{22} = 2$ Mbps. The C_{nv} values can support a total of 30 VBR calls with required bandwidth allocation $B_m \in [0.064, 0.128]$ Mbps for MTs with single-network service, i.e. $C_{112} = 30$, and 19 VBR calls with required bandwidth allocation $B_m \in [0.256, 0.512]$ Mbps for MTs with multi-homing services, i.e. $C_{122} = 19$. The arrival process of new and handoff video calls to the service area under consideration is modeled as a Poisson process with parameter υ_{112} (call/minute) for single-network service and υ_{122} (call/minute) for multi-homing service. The video call duration is modeled by a hyper-exponential distribution with the PDF given in (1) and $a_{1v} = 1$. The average call duration for single-network service $\bar{T}_c^{11} = 15$ minutes and for multi-homing service $\bar{T}_c^{12} = 10$ minutes. The user residence time in the service area under consideration follows an exponential distribution with an average time $\bar{T}_r = 20$ minutes [12]. The parameters η_1 and η_2 are set to 1 [16]. The WMAN and cellular networks set different costs on their resources using the priority parameter $p_{1m1} = 0.8$, $p_{2m1} = 0.6$ for network users, while $p_{nms} = 1$ for network subscribers [9]. The GDXMRW utilities [28] are used to create an interface between GAMS and MATLAB in order to make use of the BARON solver of GAMS in solving the optimization problem of (10) while using the MATLAB simulation and visualization tools.

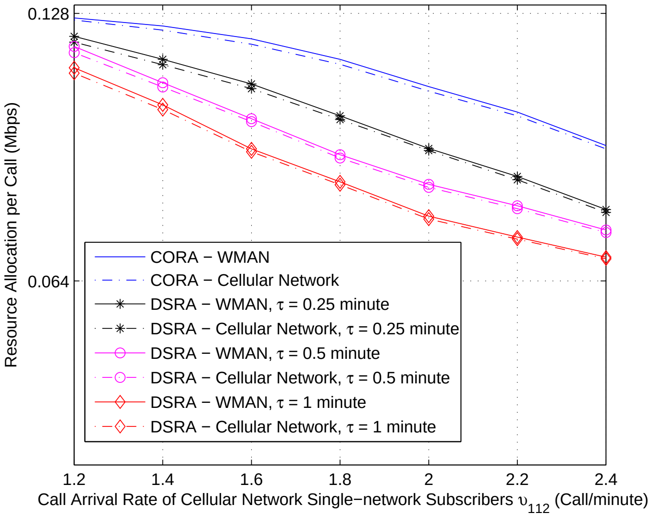

Fig. 3. Resource allocation per call versus υ_{112}.

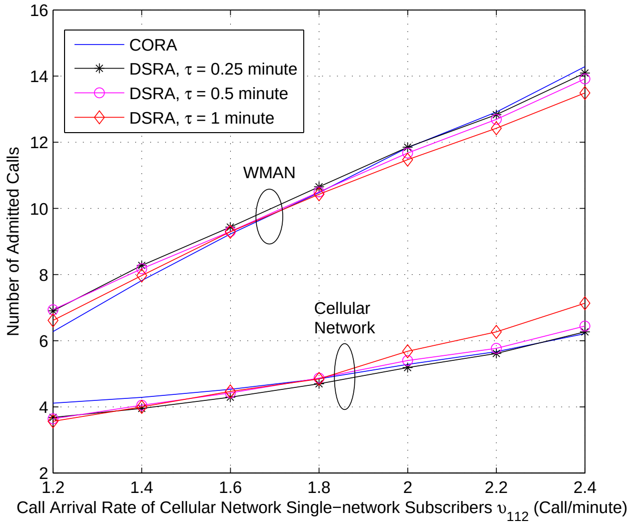

Fig. 4. Number of admitted calls versus υ_{112}.

A. Performance Comparison

In the following, the performance of the DSRA algorithm is compared to the CORA algorithm. Although it is not appropriate for practical implementation when different networks are operated by different service providers, the CORA algorithm is used as a performance bound for the allocated resources per call and the call blocking probability. In the simulation, we set the upper bounds on call blocking probability ϵ_{112}, ϵ_{122} to 1% and the prediction duration τ to 0.25, 0.5, and 1 minute. Due to space limitation, we only show the results for single-network service and same observations hold for multi-homing service.

Figures 3 - 5 show performance comparison between the DSRA and CORA algorithms for MTs with single-network service versus the call arrival rate υ_{112}. Figure 3 shows the allocated bandwidth per call for MTs assigned to the WMAN

[1] Some numerical results are presented in [1] for a complete sharing strategy for each BS/AP transmission capacity [27] where both service types can occupy up to the total capacity of each BS/AP.

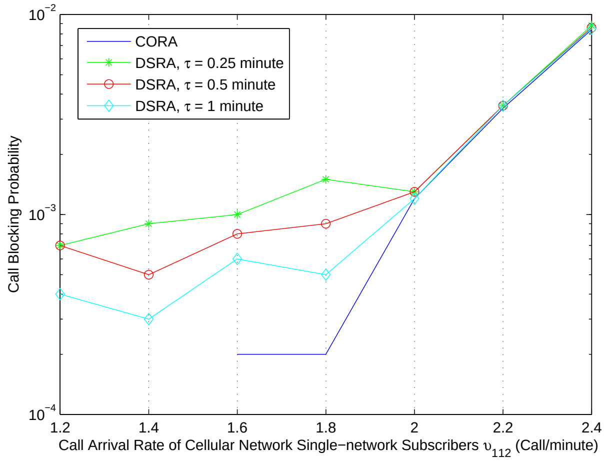

Fig. 5. Call blocking probability versus υ_{112}.

and MTs assigned to the cellular network. At a low call arrival rate, the predicted number of simultaneously present calls is low, hence the allocated bandwidth per call using the DSRA algorithm for different τ values is high. At a high call arrival rate, the predicted number of simultaneously present users is high, hence less bandwidth is allocated to each call. Moreover, less bandwidth per call is allocated for larger values of τ as explained in the next sub-section. Figure 4 shows that more MTs with single-network service are assigned to the WMAN as compared to the cellular network due to the WMAN larger capacity C_{11}. In Figure 5, using the CORA algorithm, there is no call blocking probability for $\upsilon_{112} < 1.6$ call/minute. For call arrival rate $\upsilon_{112} < 2.2$ call/minute, the DSRA algorithm does not exceed the upper bound on call blocking probability of 1%. For call arrival rate $\upsilon_{112} \geq 2.2$ call/minute, the predicted number of calls simultaneously present in the service area is larger than C_{112}. Hence, according to the DSRA algorithm, the predicted number is made equal to C_{112}, and the DSRA and the CORA algorithms achieve the same call blocking probability.

B. Performance of The DSRA Algorithm

In the following, we study the performance of the DSRA algorithm versus its two design parameters, namely the upper bound on call blocking probability $\epsilon_{l\upsilon k}$ and the prediction duration τ. Due to space limitation, we only show the results for multi-homing service and the same conclusions hold for single-network service.

Figure 6 shows the performance of the DSRA algorithm in terms of the amount of allocated resources per call and call blocking probability versus ϵ_{122}, with call arrival rate $\upsilon_{122} = 1.4$ call/minute and $\tau = 1$ minute. A small ϵ_{122} value results in a low call blocking probability. However, this corresponds to a large predicted number of calls and hence large BS/AP link access price values, which results in a small amount of allocated resources per call. On the other hand, a large ϵ_{122} value results in a high call blocking probability and a large amount of allocated resources per call. It is observed that the call blocking probability does not exceed its upper bound ϵ_{122}. The upper bound ϵ_{122} should be chosen to balance the trade-off existing between the allocated resources per call and the call blocking probability.

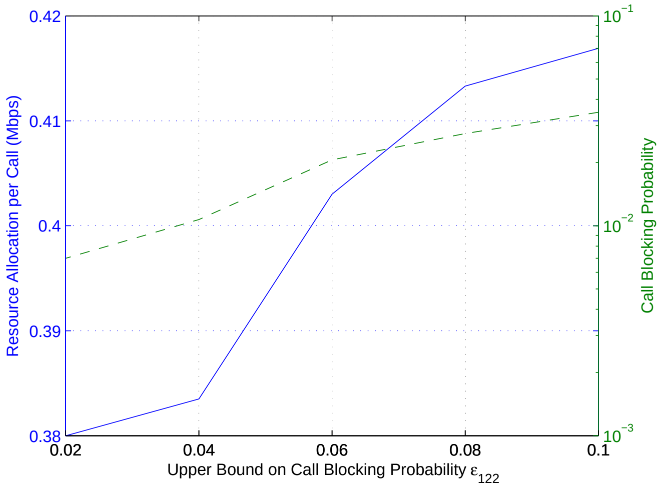

Fig. 6. The DSRA algorithm performance versus ϵ_{122}.

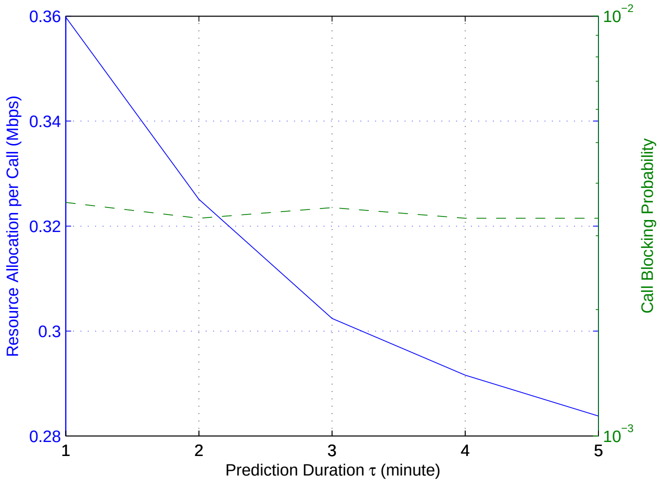

Fig. 7. The DSRA algorithm performance versus τ.

Figure 7 shows the performance of the DSRA algorithm in terms of the amount of allocated resources per call and call blocking probability versus the prediction duration τ, with $\upsilon = 1.4$ call/minute and $\epsilon_{122} = 1\%$. As τ increases, the DSRA algorithm updates the BS/AP link access price less frequently and a larger number of simultaneously present calls is predicted. As a result, the resource allocation per call is reduced. Also, simulation results indicate that the call blocking probability does not exceed its upper bound ϵ_{122}.

VII. Conclusion

In this paper, a decentralized resource allocation algorithm is proposed for a heterogeneous wireless access medium to support MTs with single-network and multi-homing services. The proposed solution gives the MTs an active role in the resource allocation operation, such that an MT with single-network service can select the best available network at its location and asks for its required bandwidth, while an MT with multi-homing service can determine the required bandwidth share from each network in order to satisfy its total required bandwidth. The resource allocation relies on short-term call traffic prediction and network cooperation in order to perform the decentralized resource allocation in an efficient manner. The two design parameters, namely ϵ_{lvk} and τ, can be properly chosen to strike a balance between the desired performance in terms of the allocated resources per call and the call blocking probability, and between the performance and implementation complexity. Different service providers are expected to cooperate with each other if they get paid from the mobile users in exchange of the cooperative service.

References

[1] M. Ismail, W. Zhuang, and M. Yu, "Radio resource allocation for single-network and multi-homing services in heterogeneous wireless access medium," *IEEE VTC 2012*, to appear.

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

[3] W. Zhuang and M. Ismail, "Cooperation in wireless communication networks," *IEEE Wireless Communications*, vol. 19, no. 2, pp. 10-20, April 2012.

[4] S. Mohanty and I. F. Akylidiz, "A cross-layer (layer 2 + 3) handoff management protocol for next generation wireless systems," *IEEE Trans. Mobile Computing*, vol. 5, no. 10, pp. 1347-1360, 2006.

[5] W. Shen and Q. Zeng, "Resource management schemes for multiple traffic in integrated heterogeneous wireless and mobile networks," *Proc. 17th Int. Conf. ICCCN*, pp. 105-110, August 2008.

[6] E. S. Navarro, Y. Lin, and W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Trans. Vehicular Technology*, vol. 57, no. 2, pp. 1243-1254, March 2008.

[7] C. Luo, H. Ji, and Y. Li, "Utility based multi-service bandwidth allocation in the 4G heterogeneous wireless access networks," *Proc. IEEE WCNC'09*, April 2009.

[8] M. Ismail and W. Zhuang, "A distributed resource allocation algorithm in heterogeneous wireless access medium," *Proc. IEEE ICC 2011*, June 2011.

[9] M. Ismail and W. Zhuang, "A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium," *IEEE J. Select. Areas Communications*, vol. 30, no. 2, pp. 425-432, Feb. 2012.

[10] M. Ismail, A.Abdrabou, and W. Zhuang, "Cooperative decentralized resource allocation in heterogeneous wireless access medium," *IEEE Trans. Wireless Communications*, submitted.

[11] S. J. Gerasenko, A. Rayaprolu, S. Ponnavaikko and D. K. Agrawal, "Beacon signals: what, why, how, and where?" *Computer*, vol. 34, pp. 108-110, 2001.

[12] W. Song, Y. Cheng, and W. Zhuang, "Improving voice and data services in cellular/WLAN integrated network by admission control," *IEEE Trans. Wireless Communications*, vol. 6, no. 11, pp. 4025-4037, Nov. 2007.

[13] M. Li, M. Claypool, R. Kinicki, and J. Nichols, "Characteristics of streaming media stored on the Web," *IEEE/ACM Trans. Networking*, vol. 5, no. 4, pp. 601626, Nov. 2005.

[14] W. Song and W. Zhuang, "Resource allocation for conversational, streaming, and interactive services in cellular/WLAN interworking," *Proc. IEEE Globecom'07*, pp. 4785-4789, Nov.-Dec. 2007.

[15] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," *Performance Evaluation*, vol. 31, no. 3-4, pp. 245-279, Jan. 1998.

[16] H. Shen and T. Basar, "Differentiated Internet pricing using a hierarchical network game model," *Proc. IEEE ACC'04*, pp. 2322-2327 vol.3, 2004.

[17] D. P. Bertsekas, *Non-linear programming*, Athena Scientific, 2003.

[18] I. E. Grossmann, "Review of nonlinear mixed-integer and disjunctive programming techniques," *Optimization and Engineering*, vol. 3, no. 3, pp. 227-252, Sept. 2002.

[19] M. Tawarmalani and N. .V. Sahinidis, "Global optimization of mixed integer nonlinear programs: a theoritical and computational study," *Math. Program.*, vol. 99, no. 3, April 2004.

[20] M. Schluter, J. A. Egea, and J. R. Banga, "Extended ant colony optimization for non-convex mixed integer nonlinear programming," *Comput. Oper. Res.*, vol. 36, no. 7, pp. 2217-2229, 2009.

[21] M. R. Bussieck and S. Vigerske, "MINLP Solver Software," *Wiley Encyclopedia of Operations Research and Management Science*, Apr. 2011.

[22] N. V. Sahinidis and M. Tawarmalani, "BARON: GAMS solver manual," May 2011.

[23] A. Neumaier, O. Shcherbina, W. Huyer, and T. Vinko, "A comparison of complete global optimization solvers," *Math. Program.*, vol. 103, pp. 335-356, 2005.

[24] M. R. H. Mandjes and P. Zuraniewski, "M/G/infinity transience, and its applications to overload detection," *Performance Evaluation*, vol. 68, no. 6, pp. 507-527, Feb. 2011.

[25] L. S. Lasdon, *Optimization theory for large systems*, Macmillan series in operations research, 1970.

[26] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Select. Areas Communications*, vol. 24, no. 8, pp. 1439-1451, Aug. 2006.

[27] M. Naghshineh and A. Acampora, "QoS provisioning in micro-cellular networks supporting multiple classes of traffic," *Wireless Networks*, vol. 2, no. 3, pp. 195-203, Aug. 1996.

[28] M. C. Ferris, R. Jain, and S. Dirkse, "GDXMRW: interfacing GAMS and MATLAB," Feb. 2011.

Muhammad Ismail (S'10) received the BSc. and MSc. in Electrical Engineering (Electronics and Communications) from Ain Shams University, Cairo, Egypt in 2007 and 2009, respectively. He is a research assistant and currently working towards his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include distributed resource allocation, quality-of-service provisioning, call admission control, green wireless networks, and cooperative networking. He served as a TPC member in the ICWMC in 2010, 2011 and 2012. He is serving in the IEEE INFOCOM 2014 organizing committee as a web chair. He joined the International Journal On Advances in Networks and Services editorial board since January 2012. He has been an editorial assistant for the IEEE Transactions on Vehicular Technology since January 2011. He has been a technical reviewer for several conferences and journals (IEEE Communications Magazine, IEEE Transactions on Mobile Computing, IEEE Transactions on Wireless Communications, IEEE Communications Letters, International Journal in Sensor Networks, and IET Communications).

Weihua Zhuang (M93-SM01-F'08) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, since 1993, where she is a Professor and a Tier I Canada Research Chair in Wireless Communication Networks. Her current research focuses on resource allocation and QoS provisioning in wireless networks. She is a co-recipient of the Best Paper Awards from the IEEE Multimedia Communications Technical Committee in 2011, IEEE Vehicular Technology Conference (VTC) Fall 2010, IEEE Wireless Communications and Networking Conference (WCNC) 2007 and 2010, IEEE International Conference on Communications (ICC) 2007, and the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine) 2007 and 2008. She received the Outstanding Performance Award 4 times since 2005 from the University of Waterloo, and the Premier's Research Excellence Award in 2001 from the Ontario Government. Dr. Zhuang is the Editor-in-Chief of IEEE Transactions on Vehicular Technology, and the Technical Program Symposia Chair of the IEEE Globecom 2011. She is a Fellow of the Canadian Academy of Engineering (CAE), a Fellow of the IEEE, and an IEEE Communications Society Distinguished Lecturer.