Available online at www.sciencedirect.com

ScienceDirect

Computer Communications 31 (2008) 950–961

computer communications

www.elsevier.com/locate/comcom

Vertical handoffs as a radio resource management tool

Abd-Elhamid M. Taha [a], Hossam S. Hassanein [a,*], Hussein T. Mouftah [b]

[a] *School of Computing, Queen's University, Kingston, Ont., Canada K7L 3N6*
[b] *SITE, Ottawa University, Ottawa, Ont., Canada K1N 6N5*

Available online 3 January 2008

Abstract

Vertical handoffs occur when a user changes association from one type of wireless access technology to another while maintaining an active session. Much work has been done in ensuring seamless handoffs that also preserve quality of service (QoS). However, network operators can exploit vertical handoffs through dedicated radio resource management (RRM) modules to relieve congestion, balance the load and uphold QoS requirements. Nevertheless, this exploitation requires rigorous study in order to realize its full potential. In this paper, we advocate the use and study of operator motivated vertical handoffs as a powerful RRM tool. We also discuss the different factors involved in the design and operation of an operator motivated vertical handoff module (OMVHM). Through simulation studies, we show the potential of OMVHs in both the single class and the multi-class setting.

Keywords: Radio resource management (RRM); QoS; Heterogeneous wireless networks; Wireless overlay networks; Vertical handoffs

1. Introduction

A heterogeneous wireless network (HWN) [1] is a composite made of two or more wireless access technologies, each with its own characteristics in terms of coverage, QoS assurance, implementation and operation costs, etc. [1]. HWNs are also called wireless overlay networks (WONs) since their coverage comprises wireless overlays, i.e. coverage overlaps, of different technologies. An example of a wireless overlay is shown in Fig. 1. For a network operator, the objective of creating such a "composition" is to realize a network that is more capable in delivering the service between the service provider and the wireless end user. The operator's HWN would exploit the better characteristics of the different access technologies in terms of coverage, efficiency, or profitability. For users, HWNs present a further degree-of-freedom in selecting the access technology that is most appropriate for the user's terminal capabilities and application requirements – even connect simultaneously to more than one access technology.

Elementary to the operation of HWNs is the existence of multi-mode terminals, i.e. terminals with more than one radio interface and each enabling them to access a different access technology. With such a capability, terminals can initiate connectivity through the technology that most closely matches the user's or the application's requirements. However, without maintaining the session through mobility or varying network conditions, a session must be terminated and re-initiated if the user is interested in continuing the session through a different interface. This capability is called inter-technology handoff or vertical handoff (VH). The term "vertical" counters the term "horizontal handoff" which characterizes a handoff taking place in networks with homogeneous access. For example, a terminal changing association between two 802.16 base stations (BS) or two 802.11 access points (AP) is said to undergo a horizontal handoff, while one changing association between a BS and an AP, or vice versa, is said to undergo a vertical handoff.

The introduction of VHs, therefore, bears certain challenges to the designer of an HWN – especially from a radio

[*] Corresponding author.
E-mail addresses: taha@cs.queensu.ca (A.-E.M. Taha), hossam@cs.queensu.ca (H.S. Hassanein), mouftah@site.uottawa.ca (H.T. Mouftah).

0140-3664/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2007.12.027

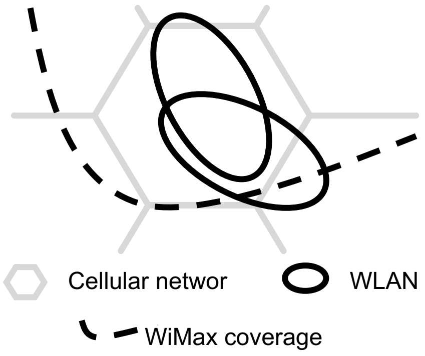

Fig. 1. An overlay of three access technologies.

resource management (RRM) perspective. User requirements and preference will be dependent on user's position and mobility, in addition to the capabilities and attributes of the networks traversed by the user during an active session. Designing RRM frameworks for HWNs will ultimately call for non-traditional modeling methods that include previously not-accounted-for characteristics. The disparity of the different networks overseen by a single operator poses several challenges. The pressing question of defining a network policy and accommodating a user policy, viz. preferences, and how to reach points of compromise also comes into light. However, while VHs motivated by user preferences bear their own challenges, the general notion of VH can be exploited by the operator. Consider, for example, the scenario shown in Fig. 2 where a user with a single mode terminal attempts to access an operator's network that is currently loaded. Under normal circumstances, the operator would not be able to accommodate the user, and the request would be denied. However, if the operator can recognize users that can be migrated to other networks with the objective of "making room" for the user's request, the call would be accepted. In fact, if the operator can recognize an overload instance in a certain network, users can be migrated from the overloaded network to other networks.

It is hence possible to categorize VHs based on the motivator. We shall refer to a VH motivated by user preferences, which can be initiated based on a user's request or upon the operator processing the user's preferences, as a user motivated vertical handoff (UMVH). An operator motivated vertical handoff (OMVH) is one that is driven by the network needs and requirements. We acknowledge that a clear distinction between the two types of handoffs may seem to be difficult to identify, but it should be understood that, while a UMVH works in a user's favor and is usually user initiated, an OMVH, even if beneficial to the user, is ultimately performed to the specific benefit of the operator.

In this paper, we advocate the use of OMVH as a critical RRM tool in next-generation wireless networks. In particular, we are concerned with the elements involved in the design of a reactive module that is dedicated to the selection of users to undergo OMVH. This includes factors affecting user selection such as the number and type of their radio interfaces, the applications running on a user's terminal and the effect of these applications being served in a different network. Against the type of interfaces, the module checks the vacancies in other networks, and whether other networks can deliver the service at an acceptable level. In a setting where multiple grades of services are offered, additional considerations need to be made, e.g. regarding the number of active users in a class or their allocations relative to other classes. In a representative implementation, we employ OMVHs in a setup also employing admission control.

The remainder of this paper is organized as follows. In the next section we overview the relevant literature.

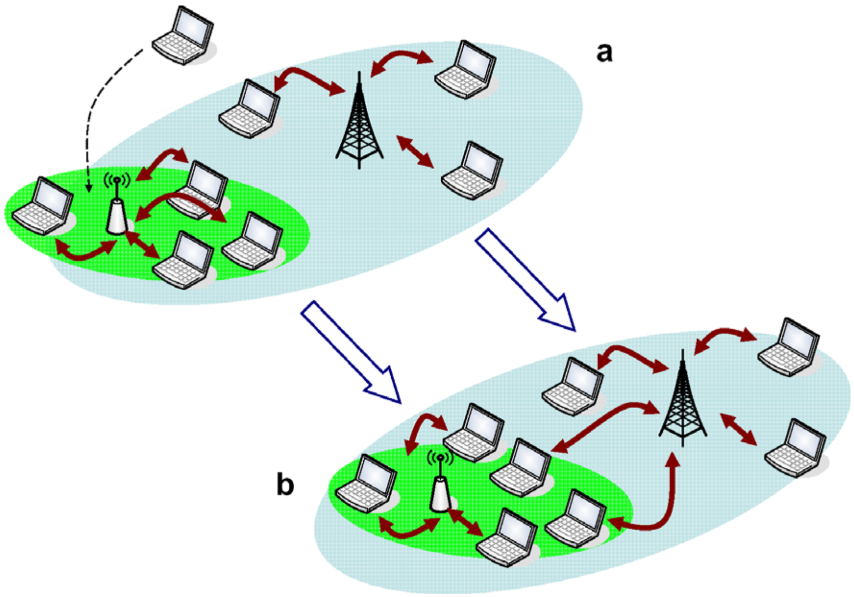

Fig. 2. An instance of employing OMVH: (a) a user makes a call to a congested network; (b) The network toggles the association of two users to another GW and accepts the call.

Next, in Section 3 we elaborate on what triggers an engagement of an OMVH module (OMVHM), and discuss considerations relevant to its operations. In Section 4, we introduce our approach in designing an OMVHM. We then offer a performance evaluation in Section 5. Finally, we conclude and hint at possible future work Section 6.

2. Related work

In this section, we shed some light on the advantages of VH to the operator. This does not deviate from maintaining the user-centricity of HWNs; rather, it complements it. Nevertheless, we maintain that, from the user's perspective, the choice of underlying technology is irrelevant as long as the user's service level agreement is upheld.

VHs are yet to be provided on a commercial scale. However, observable efforts are being made to accelerate their introduction. Multi-mode wireless terminals are beginning to be commercially available. For example, the Nokia E60 [2], provides interfaces for cellular networks (both global system for mobile communication (GSM) or universal mobile telecommunications system (UMTS)) and 802.11 wireless local area network (WLAN). In the future, multi-mode terminals will evolve into using a single interface based on the emerging notion of cognitive radio [3]. The IEEE 802.21 WG is working on enabling continuous seamless connectivity through different access technologies [4]. There are also the efforts by the internet engineering task force (IETF), Third generation partnership project (3GPP) and third generation partnership project 2 (3GPP2) on the IP multimedia subsystem (IMS) framework [5] – a culmination of efforts towards the realization of an all-internet protocol (IP) wireless network, and one that is further being considered as the unifying element for next-generation networks [6]. Through IMS, it becomes possible to evolve the abstraction of access gateways – AP, BS, etc. – to internet gateways (IGW), where each can be described and managed through standardized attributes. An effort of particular interest, exemplified by the work in [7], is one that exploits the features made possible by the introduction of IPv6, especially the enlarged address pool and allowing terminals to hold multiple active interfaces and multiple addresses per each interface. Through these features, which in essence realize multi-homing at the terminal level, it would be possible to have smooth VHs.

For users, VHs make network selection not only viable at session initiation, but also throughout the user's session. Hence, notions such as "always connect to the cheapest network" or "connect only through secure networks" can be implemented in a manner that is seamless to the user. HWNs can therefore be considered as truly user-centric or user-oriented since, by design, they persistently associate the user with the most appropriate connectivity [8].

Network selection is an involved issue in HWNs and takes on a substantial role in RRM frameworks [9,10]. Rightfully, it is receiving substantial attention in research and has been addressed for both new requests and active calls persistently seeking "best" connection [11–20]. Most approaches, e.g. [11,12,15] to network selection resort to a form of "network score function" where the different networks are rated through normalizing and weighing different criteria such as network properties, available interfaces and user preferences. Persistently, the network with the highest score is the one chosen to accommodate the user. For example, if connection cost is more important to the user than the QoS received, then cost would carry a larger weight. Some of the works, however, have taken the received signal strength as the sole element of network choice in HWNs. And while signal strength is mandatory in determining a network's availability, it remains only a single factor in determining the most appropriate network. Notwithstanding, general complexity of network selection have lead researchers to propose heuristics [13,19], the use of logical, rule-based [16] and policy-based [17,18,20] frameworks and neural networks [14].

However, Calvagna and Di Modica in [21] note the ramifications for active sessions to undergo a VH; e.g. service disruption due to handoff latency, and suggest a conservative attitude towards fulfilling user's preferences. The authors recognize that strictly meeting user preferences; e.g. always selecting the network with the best throughput, may not necessarily be of benefit to the user. Given the possible detrimental effects on certain applications, there are times a VH, even to a more capable network, may not be desirable. Hence is the consideration of the cost of VHs which the authors discuss. However, it is worth noting that much effort is being made in making VHs (and (IP) handoffs in general) more seamless. This can be seen, for example, in [22,23] where multihoming at the terminal level is exploited. It can also be observed in the standardization efforts described above. Accordingly, the ramifications of a VH will eventually be limited to the feasibility of meeting the user's service level agreement (SLA) across the different networks and not the VH itself.

Most relevant to the work presented herein is the work by Lincke-Salecker in [24]. Through a Markovian model, the author investigates policies for session selection, transfer and return, with the objective of investigating the effect of different policies on blocking ratios. The author, however, does not take into account the possible user and operational costs of VHs. Also, the objective of the author's work was to realize RRM policies that overlook networks under an operator and not specific modules that can be either deployable or addible to traditional RRM frameworks. Note that in our work we did not consider return policies as we believe that they stand beyond the specific objective of an OMVHM module and that other RRM modules exist to ensure the avoidance of needless VHs and ping-pong effects.

3. Considerations of an OMVH module

In future wireless networks, an operator will assume the provision of networks with various access technologies. It is possible to realize a centralized entity that simultaneously manages the resources of all the networks. However, centralization bears delay, overhead, and general inefficiency, making a distributed solution whereby complementary RRM modules reside in the different networks more desirable. In the following discussions, we make considerations for the latter setting. We further assume the existence of means for exchanging the capabilities, status and demands of each network between such modules.

3.1. RRM framework interactions of OMVHM

The RRM framework into which the OMVHM notion is employed should weigh whether engaging OMVHM is the current best action for both the system and the user. If the framework employs an adaptation and/or a pre-emption module, the procedure for admission control can sequentially check for available bandwidth, adaptable bandwidth, moveable bandwidth, then bandwidth that can be preempted. This "sequence" represents a design emphasis that downgrading a user's allocation is preferred over making the user undergo an OMVH, and so on. On the other hand, the system can simultaneously weigh all the options, i.e. adaptation, OMVH, etc., for each user and select the best option in an individual or class-based manner. The selection of a particular action for each user would be based on status of the system and overlaid networks.

3.2. Triggering OMVHM

There are different triggers that indicate to the operator that a migration of certain users is required. In responding to admission requests, certain users may have special requests that cannot be satisfied in other networks. This could be due to the priority and/or the requirements of the user or the service. There are also instances that depend on the capabilities of the user's terminals. A possible example can be found in transitional scenarios when the user's terminal cannot access all technologies in the overlay. When a certain network nears or reaches congestion, the operator, if feasible, may opt to migrate certain users who are within the coverage of other networks. It is worth noting that congestion may not necessarily arise from traffic dynamics. For example, due to medium variations, increased levels of interference may hinder the network from maintaining the required QoS levels.

To put matters into perspective, it is essential to note that the flexibilities offered by OMVH are intrinsically limited. An operator's capability to migrate a certain user from one network to another is bound by the user's service agreement, the user's terminal capabilities, the capabilities of the networks to which the user can be migrated. However, contemplating the possible enumeration of user applications, preferences, terminal capabilities, network configurations, etc., and the fact that wireless overlays are to comprise a substantial part of any operator's coverage, the potential of OMVHs becomes hard to overlook.

3.3. Elements of OMVHM

When an OMVHM is triggered within a network, the operator begins with identifying users that can be migrated to other networks. The following are some of the considerations that can be made in identifying such users.

- *Interface capability.* What type of interfaces does the user have? Which of interfaces is currently on?
- *Status of other networks in the overlay.* Would the networks available in the overlay area be able to provide an acceptable service delivery at one (or more) of the user's interfaces?
- *Position capability.* Which networks are available at the user's location? Which network would be able to accommodate the user and his behavioral profile; e.g. user's expected mobility pattern?
- *Application sensitivity.* How sensitive are the applications running at the user terminal to vertical handoffs [21]? To which networks would the vertical handoff disruptions; e.g. packet losses, be minimized?
- *Overhead-bound vs. released capacity.* Should the network handoff few users with high allocations, or many users with low allocations? Given that there is a signaling overhead associated with vertical handoffs, is there an upper bound on the number of users that can be handed off to other networks, in total and/or individual?
- *Users satisfaction.* How willing is the user to undergo a vertical handoff? Has this user been recently forcibly handed off? What are the short and long term costs of forcing a user into an OMVH?

These values are general, and stand as applicable when deciding whether an arbitrary user can be migrated to a certain network. Of course, other factors can be involved in the identification. For example, it may matter how long has the service been delivered to a user.

When the operator becomes interested in differentiating users or their services, further considerations can be made. As an example, a certain class of users can be favored or disfavored based on one of the following.

- *Occupied bandwidth.* The target allocation of each class can differ from one class to another. On a finer granularity, users of a certain class may be few but with large allocation, but users of another class may occupy the same allocation with a larger number of users.
- *Releasable bandwidth.* More subtle than the aspect of occupied bandwidth is that of releasable bandwidth. For example, a class may occupy a large bandwidth,

but other networks in the overlay can only accept users of another class that may occupy bandwidth of less significance.

- *Active and releasable connections.* Beyond bandwidth, there are other connection attributes, such as cost, that can be taken into account. The system may be interested in the number of users (connections) currently being serviced, or the number of users (connections) that other networks in the overlay can accommodate.
- *Class satisfaction.* If users of a certain class are generally more capable to be migrated than users of other classes, the network can deplete their "capability" at a rate less than that at which other classes are depleted so that their worth is preserved for more dire situations. The inverse is also possible – where the more "capable" users are transferred first.

Other aspects can be added to consideration. What is important to note, however, is that a multi-class OMVHM may operate under different objectives – including one that satisfies more than one objective. For example, an OMVHM may be set to operate in a manner that depletes classes with the highest occupied bandwidth, yet with the lowest or average worth to be vertically handed off.

4. Designing OMVH modules

In what follows, we detail the considerations involved in designing a module identifying and selecting users to be migrated from one network to another based on a certain trigger.

Readily, we can recognize that there are four distinct stages in the operation, as schematized in Fig. 3, namely Trigger, Identification, Selection and Migration.

In the trigger stage, the system recognizes that there are sufficient reasons to initiate migration. Consider, for example, a scenario when an admission request is made and there are insufficient resources in the requested network. If the network management options are exhausted, e.g. downgradation can no longer be made, then the option of migration becomes attractive.

There are instances where it will be possible to know whether migration is feasible, i.e. whether sufficient resources can be freed, prior to engaging the OMVHM module. However, there are also instances when an inconclusive engagement is unavoidable. The choice is highly dependent on the type of trigger and the number of users and/or networks involved; i.e. at which layer in the network management hierarchy is the OMVHM engaged. Our focus in this paper will be on conclusive engagements. The purpose of the identification stage is to make the network aware of which users can be migrated to which networks overlaying its coverage. This identification depends on the considerations discussed in Section 3.3.

The selection stage is the core of the OMVHM and yields the main thrust in responding to the trigger. In other words, based on the requirements set by the trigger; e.g. certain bandwidth is to be freed, and the sets populated by the identification stage, in addition to the capabilities of the relative networks in the overlay, the selection stage selects users based on a certain criteria while satisfying a specific system objective.

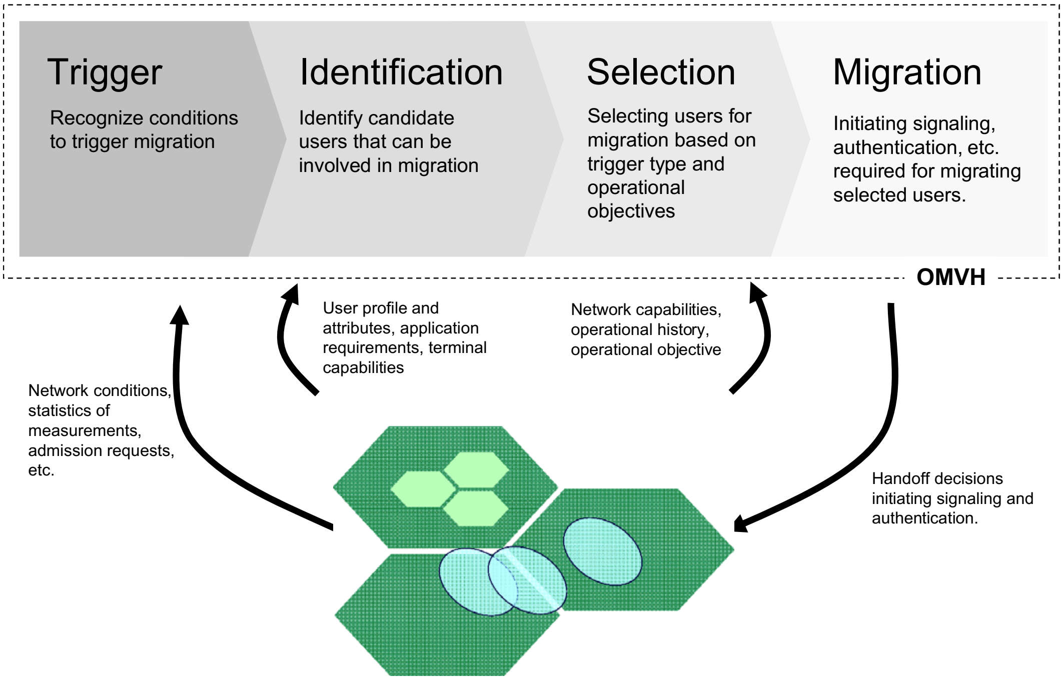
Trigger
Recognize conditions to trigger migration
Identification
Identify candidate users that can be involved in migration
Selection
Selecting users for migration based on trigger type and operational objectives
Migration
Initiating signaling, authentication, etc. required for migrating selected users.
OMVH
User profile and attributes, application requirements, terminal capabilities
Network capabilities, operational history, operational objective
Network conditions, statistics of measurements, admission requests, etc.
Handoff decisions initiating signaling and authentication.

Fig. 3. The four stages in the OMVH operation.

The migration stage oversees all the signaling required between the operator and the selected users. This may involve more than one management layer (physical, network, etc.) and is performed in the means accessible to the network operator. The details of this stage are beyond the scope of this work.

In what follows we will elaborate on the identification and selection stages. However, we digress to introduce a metric that aids the identification stage.

4.1. Evaluating a migration's worth

In order to simplify the migration decision process, a metric is required to indicate the *worth* of a user's migration from one network to another. There are several manners in which this metric can be valuated. Here, we attempt to offer a generalized form. Consider a function, denoted W, that evaluates the worth of migrating user u_{id} from current network to network n. The valuation is performed as follows.

$$W(u_{id}, net\,n) = \Phi_{\Pi}(u_{id}, net\,n) \cdot \Phi_{\Sigma}(u_{id}, net\,n) \tag{1}$$

In essence, the function reflects our recognition that there are two types of factors that affect the valuation of a migration's worth: multiplicative and additive. The function Φ_{Π} valuates the multiplicative components, and can be expanded as follows,

$$\Phi_{\Pi}(u_{id}, net\,n) = \prod_{f=1}^{N_{\Pi}} h_f(u_{id}, net\,n) \tag{2}$$

where N_{Π} is the number of multiplicative factors. The functions h_f may represent imperative factors, i.e. do-or-die factors with binary (0/1) valuations. Examples of these factors include whether the user's terminal has an interface for the receiving network, whether the user is within the coverage of the receiving network or whether available in the receiving network. If the answer to any of these is negative, the worth of migration must be nullified.

The functions h_f may also represent factors that weigh on a migration's worth and, accordingly, hold values between 0 and 1. Here, the number of migrations a user have undergone within a certain time window can be indicated.

The function Φ_{Σ} valuates the additive components of W, and can be expanded as follows,

$$\Phi_{\Sigma}(u_{id}, net\,n) = \sum_{f=1}^{N_{\Sigma}} w_f \cdot g_f(u_{id}, net\,n) \tag{3}$$

where N_{Σ} is the number of additive factors. The functions g_f, valuated between 0 and 1, represent the worth of the value or the change in the value of a certain service attribute; e.g. change in allocations, user location with respect to access gateway in the receiving network, mobility, etc. The weights, w_f, determine the relative significance of each additive element in computing a migration's worth, with a condition that $\sum_{f=1}^{N_{\Sigma}} w_f = 1$.

The reader should note that the worth function can be applied in different manners. For example, at any given time a user is likely to be receiving more than one service. In such a case, the worth of the user's migration can involve the different services in a single expression, or be the summation of the worth of migrating individual services. The latter application lends itself to scenarios where service delivery for a user, or even a single service, can be simultaneously made over more than one network. Notwithstanding, the exact manner in which the worth function is applied is inconsequential to the arguments presented herein.

To give further insight into how different factors are valuated, we offer the following examples. We have already mentioned that the operator should avoid making users undergo several VHs in a certain time frame. For example, in a preceding time window of τ_v, the operator can observe the number of VHs a particular user has undergone, denoted $N_v(t_d - \tau_v)$, where t_d is the decision instance. A function h_v, representing a multiplicative factor, can then be made to the aforementioned effect as follows.

$$h_v = \begin{cases} 1 & N_v(t_d - \tau_v) = 0 \\ 0.4 & N_v(t_d - \tau_v) = 1 \\ 0 & N_v(t_d - \tau_v) > 1 \end{cases} \tag{4}$$

A migration's worth can also be decided by a user's current allocation. The size of allocation relative to the receiving network's capacity affects the number of users that can be migrated. Meanwhile, a user's perceived QoS, and hence short and long term appreciation to the operator, are affected by the ratio of allocations between the target and migrating network. For illustration, we focus on the latter aspect. Let Q_t and Q_s be the allocations in the target and migrating networks, and g_q be the additive factor concerned with a migration's worth relative change in allocation. A crude valuation of g_q could take the following form.

$$g_q = \begin{cases} 1 & Q_t > Q_s \\ 0.8 & Q_t = Q_s \\ 0.2 & Q_t < Q_s \end{cases} \tag{5}$$

Again, we note that such valuations can be considered for the aggregate allocation or the allocations for the different services.

Another example of an additive factor is one that is based on average signal strength received at the user's terminal. Given fluctuations in signal strength measurements, readings can be stabilized through nominal regressive methods (median, FFT, etc.) [25].

4.2. The identification stage

In the identification stage, the network essentially attempts to recognize sets of candidate users that can be migrated to each network n. Specifically, if we denote such a set by $A(net\,n)$, and denote the set of users residing in the current network by U^m, then $A(netn)$ can be defined as follows.

$$A(netn) = \{u_{id} : u_{id} \in U^m, W(u_{id}, netn) \geqslant W_{\text{th}}\} \tag{6}$$

where W_{th} is a minimum threshold of migration worth below which the network or operator judges that a migration should not be considered. It is for simplicity of presentation that we assume that there is a single threshold for all users, services or networks under a single operator. Granted, the threshold's value should more appropriately be set to react to the network conditions and the type of users comprising the networks in addition to their active applications.

4.3. The selection stage

The result of the selection stage is sets of users to be migrated to each individual network in the overlay. These sets are essentially subsets of candidate sets, i.e. A, and their elements are selected according to a certain objective. Denote the set of all users selected for migration from the migrating network by V. We then have the following definitions.

$$V = \bigcup_{n \in O^m} V(netn) \tag{7}$$

where O^m is the set of networks overlaying the migrating network and $V(netn)$, i.e. the set of users selected to migrate to network n, is defined by

$$V(netn) \subseteq A(netn) \tag{8}$$

User selection can be made based on different criteria. One criteria that we perceive as rational and feasible is to maximize the net worth of the OMVHM engagement. Define $W(V)$ to be the total worth of migrating the selected users, i.e.

$$W(V) = \sum_{n \in O^m} \sum_{u_{id} \in V^n} W(u_{id}, netn) \tag{9}$$

where we have utilized V^n as a shorthand for $V(netn)$. The objective then becomes

$$\arg\max W(V) \tag{10}$$

We note that other constraints need to be minded in user selection. For example, it is natural that a user can only be associated with a single network at a given time.

$$V^i \cap V^j = \emptyset \quad \forall i \neq j \tag{11}$$

It is also necessary that the OMVHM releases, at least, the resources required by the trigger. Define $Q(u_{id})$ as the resources required by user u_{id} in the user's current network. Also, define $Q(V)$ as the total resources required by the users in V. Denoting the resources required by the trigger by Q_{req}, then Q_{req} should be the lower bound released by the migration process, i.e.

$$Q(V) \geqslant Q_{\text{req}} \tag{12}$$

Due to the discrete nature of allocations, and in order to bound the allocations released per trigger, we utilize a fragmentation ratio, f, in upper bounding the amount of resources released, i.e.

$$Q(V) \leqslant f \cdot Q_{\text{req}} \tag{13}$$

Naturally, the number of users that can be migrated to a certain network is limited by the amount of available resources in that receiving network. Denote the amount of available resources in network j by Q^j_{avail}. Also, denote that resources occupied by users in V after being migrated by $Q(V^j)^+$, i.e. after having received their allocations in their respective receiving networks. Then, for each receiving network j, the following condition applies.

$$Q(V^j)^+ \leqslant Q^j_{\text{avail}} \tag{14}$$

These are the basic set of considerations required for the operation of a single class OMVHM. However, this model can easily be extended. For example, it is possible that in migrating users to the various receiving networks, users would be migrated equally between them or that migration would be proportional to the capabilities of relative receiving networks.

4.3.1. The multi-class setting

In a multi-class setting further considerations need to be applied in favoring between different classes of users. Prior to describing the operation of multi-class OMVHM, we introduce a notion of class ratios that facilitates differentiation.

We define a class's VH *give*, denoted G, as the extent to which a class can be utilized in migrating users to the receiving network. The units and computation of G depends on the measure utilized in favoring classes over each other. This measure, for example, can be the number of migrated users, the bandwidth associated with the migrated users, or the worth depleted by the migrated users. Regardless of the criterion chosen for selection, a class's give represents its support of the OMVHM operation.

There are different means to calculate G. If the measure was released bandwidth, the network may seek to maintain a specific, fixed ratio for the bandwidth released between different classes. In a more dynamic setting, computation would be based on active measures like the bandwidth occupied by each class. In the following, we detail a possible implementation.

Assume that the system maintains a certain ratio, ρ, between the worth released from two classes i and j. Denote by $W_{x,\text{Tot}}$ the total worth depleted by class x thus far. Hence, OMVHM would seek the following equality.

$$\frac{W_{i,\text{Tot}}}{W_{j,\text{Tot}}} = \rho \tag{15}$$

Consider now the case where $W_{i,\text{Tot}}$ is greater than $W_{j,\text{Tot}}$. For the system to maintain the ratio, OMVHM needs to fixate $W_{i,\text{Tot}}$, and allow more worth to be released from class j. Ratios here becomes the worth to be released from class j in order for OMVHM to maintain ρ. Denote the give of class x by G_x, then

$$G_j = \frac{W_{i,\mathrm{Tot}}}{\rho} - W_{j,\mathrm{Tot}} \tag{16}$$

This computation can be easily extended to more than two classes. At times, the gives computed may not be sufficient to accommodate an incoming call. In this case, the value can be adjusted to the least sufficient give.

We now turn our focus to the description of a multi-class OMVHM. Denote the set of classes defined under the operator by S. The set of users in network n are denoted U^n. The set of class s users in network n are denoted U^n_s. The set of candidate class s users to a receiving network n is denoted A^n_s and is defined as follows.

$$A^n_s = \left\{u_{id} : u_{id} \in U^m_s, W(u_{id}, net\,n) \geqslant W_{\mathrm{th}}\right\} \tag{17}$$

where the superscript m indicates the migrating network. Consequently, the set of class s users selected according to the operating objective to be migrated to n, denoted V^n_s, is defined by

$$V^n_s \subseteq A^n_s \tag{18}$$

Here, too, we note that, in addition to the fact that a user can only be associated with a single class at a given time, a user can only be migrated to one receiving network, i.e.

$$V^i_s \cap V^j_s = \emptyset \quad \forall i \neq j \tag{19}$$

We redefine the set V here as follows.

$$V = \bigcup_{n \in O^m} \bigcup_{s \in S} V^n_s \tag{20}$$

We similarly redefine $W(V)$ as follows.

$$W(V) = \sum_{n \in O^m} \sum_{s \in S} \sum_{u_{id} \in V^n_s} W(u_{id}, net\,n) \tag{21}$$

The objective then becomes

$$\arg\max W(V) \tag{22}$$

Given that the two constraints regarding the total released resources, i.e. $Q(V)$, and resources required by the trigger, i.e. Q_{req} are maintained.

Constraints on the available resources can be defined as for the single class setting, or with individual constraints for each user class, i.e. denoting the amount of available resources for class s in network j by $Q^j_{s,\,\mathrm{avail}}$, the condition becomes

$$Q(V^j_s)^+ \leqslant Q^j_{s,\,\mathrm{avail}} \tag{23}$$

Define V_s as follows.

$$V_s = \bigcup_{n \in O^m} V^n_s \tag{24}$$

Also, define the function $G(V_s)$ as the total give consumed by V_s, and denote the bound on class's give by G_s. Then the OMVHM needs to maintain the following

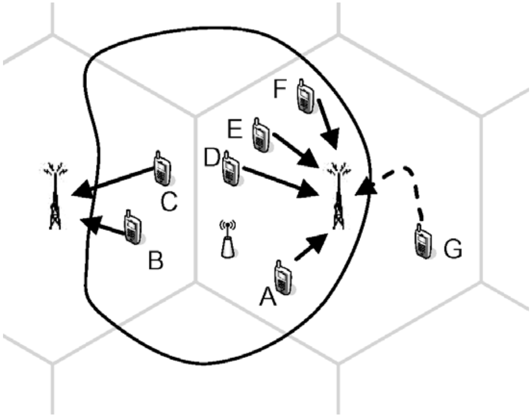

Fig. 4. User G is attempting to enter a loaded cellular network. Users A to F are residing within an overlay of cellular network and a WLAN.

$$G(V_s) \leqslant G_s \tag{25}$$

4.4. Illustrative example

Consider the scenario in Fig. 4. Users A, B and C are class 1 users, and are all allocated 6 effective bandwidth units (ebus). Their worth to be migrated to the WLAN is 0.4. Users D, E and F are class 2 users and are, respectively, allocated 4 ebus, 2 ebus, and 2 ebus. Their respective worth is 0.2, 0.6 and 0.6 User G's request mandates an allocation of 8 ebus. The network's fragmentation ratio is set at 1.3, making the upper bound of releasable allocations to be 10.4 bu. The WLAN can only accept one class 1 user and three class 2 users. Employing the OMVHM without the give budget would result in selecting either user A, B or C, and E and F.

Suppose now that the give of class 1 was zero, and of class 2 was a worth of 0.7. Since 0.7 for class 2 would not result in releasing sufficient bandwidth to accommodate user G, the give is adjusted to the least sufficient worth, which is $(0.2 + 0.6 + 0.6 =)$ 1.4. Hence, users D, E and F would be migrated.

If the give of class 2 was zero, the OMVHM would not be engaged at any value of give for class 1 since the support of the WLAN cannot result in sufficient bandwidth to accommodate G.

5. Performance evaluation

In what follows, we examine the operational aspects of OMVH. Simulation experiments were carried out in an event-driven simulation built utilizing C++ and MATLAB. The OMVHM core was implemented through a mixed integer linear programming (MILP) formulation that was solved using the GLPK package of the GNU project [26].

5.1. Simulation setup

An overlay involving two networks, as shown in Fig. 5, is employed. We will refer to the network with the persis-

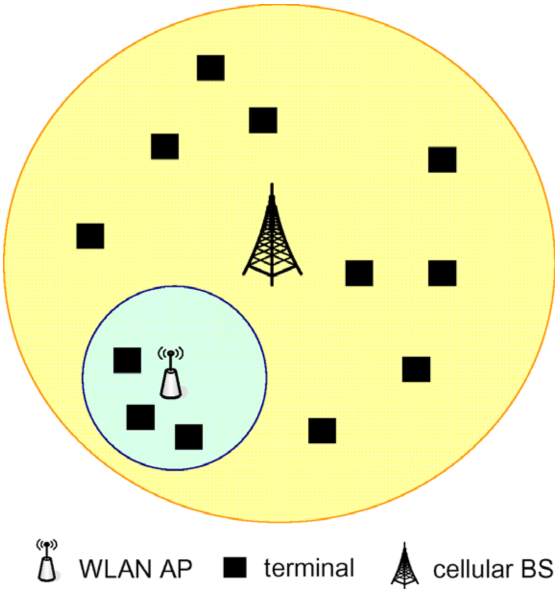

Fig. 5. An overlay of cellular and WLAN coverage.

tently larger coverage by network 1 and the second network by network 2. For purposes of evaluation, the coverages of the two networks are concentric. A fixed number of users is uniformly distributed over the area of the larger coverage. Users make connection requests with an aggregate inter-arrival time that is exponential with controllable mean. The capacities of networks 1 and 2 are, respectively, 80 ebus and 40 ebus. Initially, a single service defined in both networks that is allocated 6 ebus in network 1 and 4 ebus in network 2. The connection holding time exponentially distributed with a mean of 150 s, regardless of the network choice. All users are assumed to be dual mode users, i.e. can request and receive services in both networks. Experiments ran for a simulated time of 3600 s. Each shown result represents the outcome of ten experiments.

Note that the values used here are arbitrary, and that other values were used in the intensive investigation performed displayed similar trends to the ones presented below.

5.2. Blocking probability

When OMVH is not employed, a connection is considered blocked if the available (unallocated) bandwidth cannot satisfy the request. When OMVH is employed, a call is considered blocked if neither the available bandwidth is sufficient nor is migration possible. Blocking probability is the ratio between the number of blocked connections and the number of requests made throughout the simulation time.

Fig. 6 shows the blocking probability for both networks with the arrival rate is varied between 10 and 25 calls per minute in steps of 2.5. In generating the requests, 70% of the calls generated were aimed at network 1. Also, the percentage coverage of network 2 relative to network 1's coverage was set to 60%, i.e. partial overlay. Furthermore, the worth of migration for all users in the overlay in both directions was set to 1. The general goodness of OMVH can be observed, where a reduction of around 10% is observed at 10 calls per minute. The reduction however diminishes as the load increases 25 calls per minute. This is readily understandable since, when possible, each migrating network considers the receiving network as a virtual capacity extension. As the load increases, however, there becomes less and less "room to maneuver" for the OMVHM and the effect of OMVH becomes less noticeable. Note that this and other evaluations to follow are extreme in that every request made in a case of overload is considered to merit a migration. In reality, the decision to migrate will be determined on case by case basis.

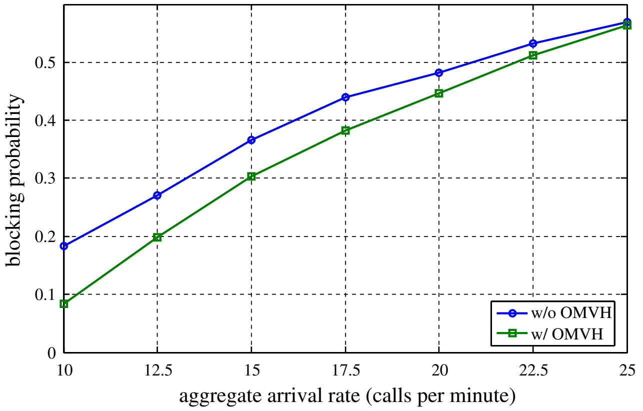

Fig. 6. Total blocking probability (i.e. for both networks) with aggregate load varied.

Also note that Fig. 6 is made for the total blocking probability. Naturally, reductions would be more apparent in network 1 than in network 2 due to the first receiving 70% of call requests. It is natural that the blocking probability for network 2 may actually worsen.

5.3. Effect of overlay percentage

We turn our attention now to study the effect of overlay percentage on the performance of OMVH. In this setup, requests are only made to network 1 with a fixed arrival rate of 10 calls per minute. With users still uniformly distributed over the coverage of network 1, the percentage overlay of the second network was changed from 0% to 100% in steps of 10%. Again, all users within the overlay were transferable. Fig. 7 shows the results. In the figure, the plot with circle bullets represents the case when OMVH is not employed. Naturally, the blocking probability remains the same since migration is not permitted. The plot with the triangle bullets represents the case when the area was changed but capacity was fixed at 40 ebus. This is different than in the case for the plot with square bullets were the area was incremented in steps of 4 ebus. The difference shows the independent effects of capacity limitation and the user distribution between the overlay. Note that the plot

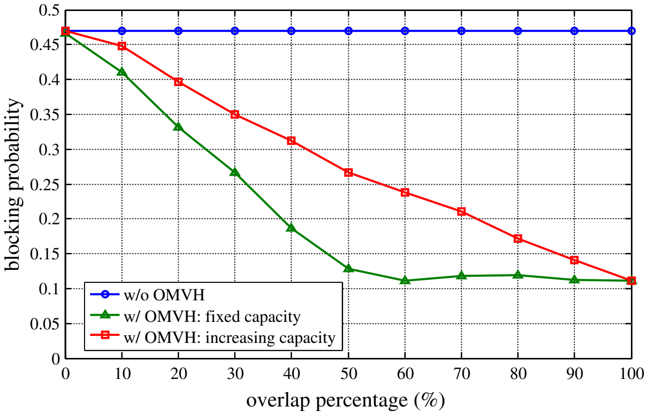

Fig. 7. Blocking probability with overlay percentage varied (relative to larger network).

for fixed capacity saturates between 50% and 60% coverage overlap, indicating that, from this point, an increase in capacity is more useful than an increase in coverage area. Such outcomes can be used when, for example, deploying networks to relieve persistent hotspots.

5.4. Effect of load distribution

The objective of this experiment is to reemphasize again the value of joint of resource management but within the context of OMVH. The percentage of requests made to network 1 was varied between 0 and 1. Migration was allowed in both directions. All users were considered transferable. The results are shown in Fig. 8, where the total blocking probability is plotted against the percentage of requests to network 1.

At the extremes, i.e. towards 0% and 100%, the blocking probability leans towards the blocking probability of the individual network in absence of migration. However, when OMVH is employed, the increase in the blocking probability is steady and less than blocking probability when OMVH is not employed – even at its lowest instance, i.e. at 50%. This affirms that OMVH is another means not only of maintaining admission guarantees, but also for increasing resource utilization.

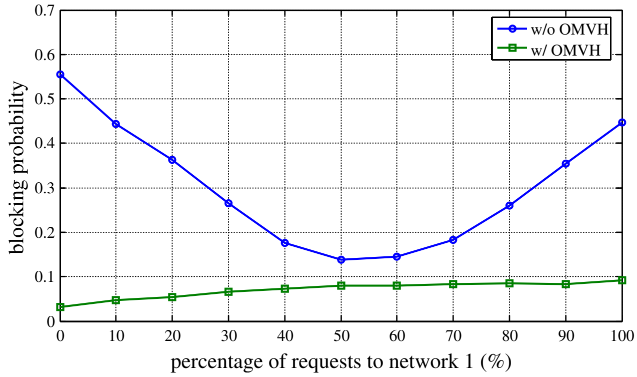

Fig. 8. Blocking probability with percentage of requests to larger networks varied, with and without employing OMVH.

5.5. Limiting average number of handoffs

As aforementioned, the valuation of worth is important as an indicator of a migration's value to the OMVH operation. As an example, we investigated the effect of having a hold-off time once a user is migrated and until the connection terminates. The hold-off time was varied from 0, i.e. no hold-off, to 3600 s, i.e. the length of the simulation time. The latter extreme implies that any connection can be migrated no more than once in its lifetime. In generating the requests, 60% of the calls were directed to network 1. The results are shown in Fig. 9. The four subplots in the figure display different aspects of the effect of the hold-off time. Fig. 9(a) displays the effect on blocking probability, where as the duration of the hold-off time is increased, the blocking probability increases. On the other hand, Fig. 9(b) shows the average number of migrations per user. There are two things to note here. First, that increasing the hold-off time reduces the average number of migrations per user. Second, that instead of the average migrations per user converging to 1, it converges around 1.45 migrations per user. This is because the number of users is limited (200) and, within a single simulation run, it is possible that the same user may make more than one request. Also, once a connection is terminated, the hold-off time is reset. Figs. 9(c) and 9(d) show the constituents of the mean in Fig. 9(b). Naturally, the total number of migrations increases. However, it is informative to note in Fig. 9(d) the increase in the number of users involved per migration between hold-off time 0 and around 150 s. This indicates the byproduct of fairness among users that can be achieved when considering hold-off times in evaluating a migration's worth.

5.6. The multi-class setting

To evaluate the operation of OMVH in the multi-class setting, we implemented a controller on top of the OMVHM to set the bounds on each class's give based on measurements of expended worth per class. The implemented controller can work with other criterion and the choice of worth is only for illustration.

In addition to the service class defined above (exponential duration with mean 150 s, 6 ebus and 4 ebus in networks 1 and 2, respectively), we defined another class with a fixed holding time equal to 300 s and which receives allocations of 4 ebus and 2 ebus in networks 1 and 2, respectively. In generating the requests, the percentage of requests made for each class was varied. Figs. 10–12 show the results of the measured ratios. The percentage of requests made for class 1 are 20% in Fig. 10, 50% in Fig. 11 and 80% in Fig. 12. In each setting, the ratios set between the expended worth for class 1 (exponential hold-

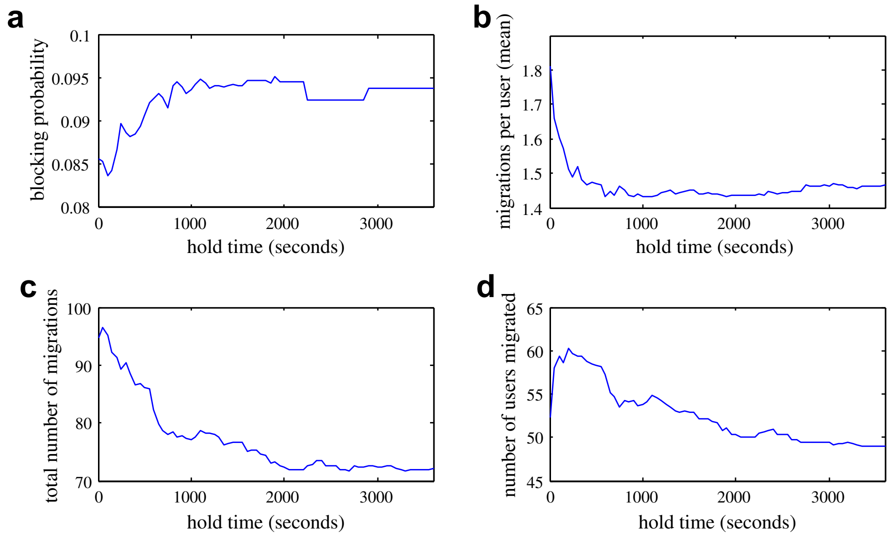

Fig. 9. Effects of varied hold-off time on (a) blocking probability; (b) number of migrations per user; (c) total number of migrations; and (d) number of individual users migrated.

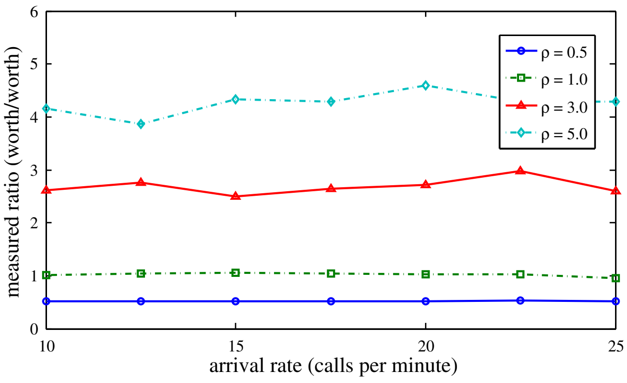

Fig. 10. Measured ratio with class 1 assuming 20% of the arrival rate.

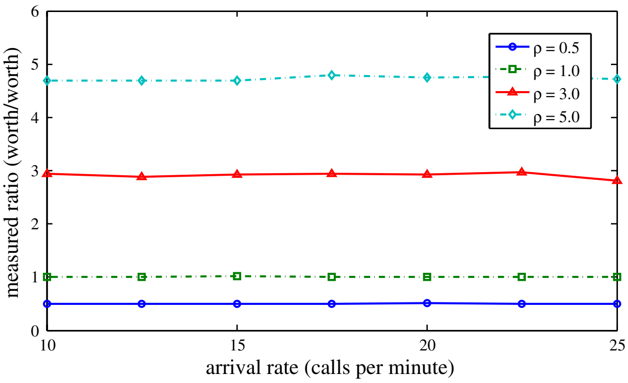

Fig. 11. Measured ratio with class 1 assuming 50% of the arrival rate.

ing time) to that of class 2 (fixed holding time) was set to either 0.5, 1.0, 3.0 and 5.0.

We note that, despite the difficulty to maintain ratios greater than 1 when the connections are mostly class 2 connections, the performance of the control remains satisfactory. From the three figures, too, it seems that the controller can perform well despite varying class occupation.

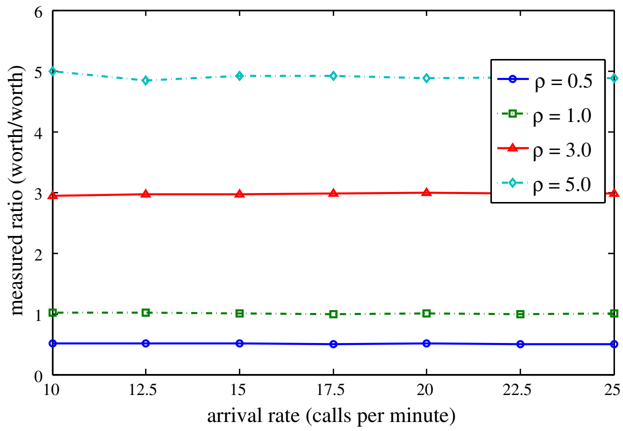

Fig. 12. Measured ratio with class 1 assuming 80% of the arrival rate.

6. Conclusions

In this paper, we introduced the novel notion of OMVH through which the operator can exploit VHs in HWNs. After displaying the feasibility and the advantages from utilizing OMVHs, we enumerated the considerations that must be accounted for when designing a module dedicated to OMVH management. We also showed the stages of operations for such modules, including the identification and selection of users to be migrated. To aid the identification stage, we proposed the notion of a migration's *worth* and showed examples of how it can be computed. In addition, we detailed how differentiation based on different user categorizations can be made.

Our work in this paper focused on reactive applications of OMVHs. Other applications with the objectives of load balancing or reducing the operator's cost of service delivery are also viable, but require further investigations to ensure effectiveness, operational stability and overhead control. The latter application becomes more pressing in light of the emerging efforts towards cognitive radios and spectrum trading.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable input and comments. They also acknowledge the support of Bell University Labs (BUL) and the National Sciences and Engineering Research Council of Canada (NSERC).

References

[1] S. Dekleva, J. Shim, U. Varshney, G. Knoerzer, Evolution and emerging issues in mobile wireless networks, Communications of the ACM 50 (6) (2007) 38–43.

[2] Nokia e60. Available from: <http://europe.nokia.com/A4145124/>.

[3] S. Haykin, Cognitive radio: brain-empowered wireless communications, IEEE Journal on Selected Areas in Communications 23 (2) (2005) 201–220.

[4] The IEEE 802.11 Working Group. Available from: <http://www.ieee802.org/21/>.

[5] 3GPP, IP Multimedia Subsystem (IMS). Available from: <http://www.3gpp.org/ftp/Specs/archive/23_series/23.228/23228-810.zip/>.

[6] A. Cuevas, J.I. Moreno, P. Vidales, H. Einsiedler, The IMS service platform: a solution for next-generation network operators to be more than bit pipes, IEEE Communications Magazine 44 (8) (2006) 75–81.

[7] Mobile nodes and multiple interfaces in ipv6. Available from: <http://www.ietf.org/html.charters/monami6-charter.html/>.

[8] S. Frattasi, H. Fathi, F.H.P. Fitzek, R. Prasad, M.D. Katz, Defining 4g technology from the users perspective, IEEE Network 20 (1) (2006) 35–41.

[9] M.R. Kibria, A. Jamalipour, On designing issues of the next generation mobile network, IEEE Network 21 (1) (2007) 6–13.

[10] P. Magnusson, F. Berggren, I. Karla, R. Litjens, F. Meago, H. Tang, R. Veronesi, Multi-radio resource management for communication networks beyond 3g, in: Vehicular Technology Conference, vol. 3, 2005, pp. 1653–1657.

[11] S. Balasubramaniam, J. Indulska, Vertical handover supporting pervasive computing in future wireless networks, Computer Communications 27 (8) (2004) 708–719.

[12] L.-J. Chen, T. Sun, B. Chen, V. Rajendran, M. Gerla, A smart decision model for vertical handoff, in: Proceedings of the 4th International Workshop on Wireless Internet and Reconfigurability, 2004.

[13] L. Huang, Y. Liu, S. Thilakawardana, R. Tafazolli, Network-centric user assignment in the next generation mobile networks, IEEE Communications Letters 10 (12) (2006) 822–824.

[14] W. Zhang, Handover decision using fuzzy MADM in heterogeneous networks, IEEE Wireless Communications and Networking Conference 2 (2004) 653–658.

[15] X. Liu, V.O.K. Li, P. Zhang, Joint radio resource management through vertical handoffs in 4g networks, IEEE Global Telecommunications Conference (2006) 1–5.

[16] A. Sur, D.C. Sicker. Available from: multi layer rules based framework for vertical handoff, in: 2nd International Conference on Broadband Networks, 2005, pp. 571–580.

[17] J. Prez-Romero, O. Sallent, R. Agustí, Policy-based initial rat selection algorithms in heterogeneous networks, in: Proceedings of the Mobile Wireless Communications Networks, 2005, pp. 1–5.

[18] W. Song, W. Zhuang, Y. Cheng, Load balancing for cellular/WLAN integrated networks, IEEE Network 21 (1) (2007) 27–33.

[19] B. Xing, N. Venkatasubramanian, Multi-constraint dynamic access selection in always best connected networks, in: The Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2005, pp. 56–64.

[20] F. Zhu, J. McNair, Multiservice vertical handoff decision algorithms, EURASIP Journal on Wireless Communications and Networking 2006 (2006) Article ID 25861, 13 p.

[21] A. Calvagna, G.D. Modica, A cost-based approach to vertical handover policies between wifi and gprs: research articles, Wireless Communications and Mobile Computing 5 (6) (2005) 603–617.

[22] F. Siddiqui, S. Zeadally, SCTP multihoming support for handoffs across heterogeneous networks, Fourth Annual Conference on Communication Networks and Services Research (2006) 243–250.

[23] W. Wu, N. Banerjee, K. Basu, S.K. Das, SIP-based vertical handoff between WWANs and WLANs, IEEE Wireless Communications 12 (3) (2005) 66–72.

[24] S.J. Lincke-Salecker, Vertical handover policies for common radio resource management: research articles, International Journal of Communications Systems 18 (6) (2005) 527–543.

[25] Q. Zhang, C. Guo, Z. Guo, W. Zhu, Efficient mobility management for vertical handoff between WWAN and WLAN, IEEE Communications Magazine 41 (11) (2003) 102–108.

[26] GNU linear programming kit. Available from: <http://www.gnu.org/software/glpk/glpk.html/>.