Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks

(Submitted to the IEEE Transactions on Vehicular Technology)

SuKyoung Lee*, Kotikalapudi Sriram†, Kyungsoo Kim, JongHyup Lee, YoonHyuk Kim, and Nada Golmie

Abstract— **There are currently a large variety of wireless access networks, including the emerging Vehicular Ad-hoc Networks (VANETs). A large variety of applications utilizing these networks will demand features such as real-time, high-availability and even instantaneous high-bandwidth in some cases. Therefore, it is imperative for network service providers to make the best possible use of the combined resources of available heterogeneous networks (WLAN, UMTS, VANETs, Wi-MAX, etc.) for connection support. When connections need to migrate between heterogeneous networks for performance and high-availability reasons, then seamless vertical handoff is a necessary first step. In the near future, vehicular and other mobile applications will expect seamless vertical handoff between heterogeneous access networks. With regard to vertical handoff performance, there is a critical need for developing algorithms for connection management and optimal resource allocation for seamless mobility. In this paper, we develop a vertical handoff decision algorithm that enables a wireless access network to not only balance the overall load among all attachment points (e.g., Base Stations (BSs) and Access Points (APs)) but also to maximize the collective battery lifetime of Mobile Nodes (MNs). Moreover, when ad-hoc mode is applied to 3/4G wireless data networks, VANETs and IEEE 802.11 WLANs for more seamless integration of heterogeneous wireless networks, we devise a route selection algorithm to forward data packets to the most appropriate AP in order to maximize the collective battery lifetime as well as maintain load balancing. Results based on a detailed performance evaluation study are also presented here to demonstrate the efficacy of the proposed algorithms.**

Index Terms— **Mobility management, Intersystem handover, QoS management, Simulation modeling, Seamless mobility, High-availability, VANET, WLAN, Wi-MAX, Vertical handoff, Load balancing.**

I. Introduction

There are many types of existing and emerging wireless access networks to support a multitude of mobile applications. Such networks include the emerging Vehicular Ad-hoc Networks (VANETs) as well as the well-known WLANs, UMTS, Wi-MAX, etc. A large variety of applications utilizing these networks will demand features such as real-time, high-availability and even instantaneous high-bandwidth in some cases. The end-user devices will increasingly be equipped with multiple RF interfaces so that it would be feasible to carry and move connections across heterogeneous wireless access networks (e.g., WLAN, UMTS, VANETs, Wi-MAX, etc.) with service continuity and enhancement of service quality. It is imperative for network service providers to make the best possible use of the combined resources of available heterogeneous networks for connection support. When connections need to migrate between heterogeneous networks for performance and high-availability reasons, then seamless vertical handoff is a necessary first step. In the near future, vehicular and other mobile applications will expect seamless vertical handoff between heterogeneous access networks. With regard to vertical handoff performance, there is a critical need for developing algorithms for connection management and optimal resource allocation for seamless mobility.

Different types of access network technologies can be effectively used to enable mobile users to have seamless access to the Internet. Connection handoff is no longer limited to migration between two subnets in Wireless Local Area Network (WLAN), or between two cells in a cellular network, generally known as “horizontal handoff”. In addition to roaming and horizontal handoff within homogeneous subnets (e.g., consisting of only VANETs, or only 802.11 WLANs, or only cellular networks), supporting Quality of Service (QoS) anytime, anywhere, and by any media requires seamless vertical handoffs between heterogeneous wireless access networks. In general the heterogeneous networks can be combinations of many different kinds, e.g., VANET, WLAN, Universal Mobile Telecommunications System (UMTS)/cdma2000, Bluetooth and Mobile Ad hoc Network (MANET). Many new architectures or schemes have been proposed recently for seamless integration of various wireless networks while the integration of WLAN and cellular networks has attracted most attention because currently WLANs and cellular networks coexist and many cellular devices have dual RF interfaces for WLANs and cellular access. Since WLANs and cellular networks are complementary technologies, we focus on these technologies in this paper but our algorithms is widely applicable across any set of access technologies and applications. WLANs have the advantages of low cost and high-speed over cellular networks while cellular networks provide wide-area coverage overcoming the well-known problem in WLANs that their coverage is typically limited to buildings and certain hotspots. Thus, industry as well as academia have started to focus on vertical handoff across wireless LAN and cellular networks.

Several interworking mechanisms have been proposed in [1]-[4] to combine WLANs and cellular data networks into integrated wireless data environments. Two main architectures [2]-[4] have been proposed for interworking between 802.11 WLAN and 3G cellular systems: (1) Tight coupling and (2) Loose coupling (see Fig. 1). With loose coupling the WLAN is deployed as an access network complementary to the 3G cellular network. In this approach, the WLAN bypasses the core cellular networks and data traffic is routed more efficiently to and from the Internet without having to go over the cellular networks which could be a potential bottleneck. Besides, this approach mandates the provisioning of special Authentication, Authorization, and Accounting (AAA) servers

*S.K. Lee is with Yonsei University, Seoul, Korea (Email: sklee@cs.yonsei.ac.kr) and †K. Sriram is with National Institute of Standards and Technology, Gaithersburg, MD, USA (Email: ksriram@nist.gov).

on the cellular operator for interworking with WLANs' AAA services. On the other hand, with tight coupling, the WLAN is connected to the cellular core network in the same manner as any other 3G Radio Access Network (RAN), so that the mechanisms for mobility, QoS and security of the 3G core network such as UMTS can be reused. As a result, a more seamless handoff between cellular and WLAN networks can be expected in the tightly coupled case as compared to the same for the loosely coupled case (attributable to the typical high latency of Mobile IP registrations in the latter case). However, further standardization and development efforts are needed to realize this capability, and this effort will be specific to the 3G RAN technologies.

There have been also several efforts to connect a mobile device equipped with multiple (currently, dual-mode) interfaces to the most optimal one among the heterogeneous access networks covering the mobile device. The optimality criterion is in terms of network performance, and the vertical mobility is achieved by switching the interface of the mobile device to access the appropriate network. The authors of [6] introduced important performance criteria to evaluate seamless vertical mobility, e.g. network latency, congestion, battery power, service type, etc. In [7], the authors proposed an end-to-end mobility management system that reduces unnecessary handoff and ping-pong effect by using measurements on the condition of different networks. In [8], various network layer based inter-network handover techniques have been addressed and their performance is evaluated in a realistic heterogeneous network testbed. The authors of [9] propose a vertical handoff decision method that simply calculates the service quality for available networks and selects the network with the highest quality.

However, there are still more challenges in integrating cellular networks and WLANs (or any combination of heterogeneous network in general). Especially, it is a challenge to design vertical handoff techniques to optimize the overall network performance such as power consumption [6].

As the authors of [5], [6] and [9] have pointed out, known vertical handoff algorithms are not adequate to coordinate the QoS of many individual mobile users or adapt to newly emerging performance requirements for handoff and changing network status. Further, under the current WLAN technology, each mobile device selects an AP for which the Received Signal Strength (RSS) is maximum irrespective of the neighboring network status. Although the attachment to the closest AP is known to consume the least power for the individual mobile device at a given instant, in a situation where many mobile devices try to handoff to the same AP, there would be in effect significantly more power consumption at the mobile devices collectively due to increased congestion delays at the AP. In addition, the power consumed by an AP increases as more power is consumed by all power-on nodes attached to the AP. In this paper, we tackle the following problem: given a network of Base Stations (BSs), Access Points (APs) and Mobile Nodes (MNs), how do we find an appropriate attachment point for an MN to connect to at the time of vertical handoff while optimizing a well defined objective function? We also extend the study for the case when an ad hoc mode such as a MANET or VANET is included in the system. Our objective function includes consideration of collective battery life of MNs and network throughput and capacity. For seamless integration of WLAN and 3/4G wireless networks, we propose a vertical handoff decision algorithm that not only maximizes the overall battery lifetime of MNs in the same coverage but also seeks to maximizes network throughput and capacity in meshed wireless networks where heterogeneous wireless networks coexist. The proposed vertical handoff decision algorithm in effect ensures proper management of network resources so as to minimize power consumption and provide equitable resource utilization among the available access networks. Moreover, when ad hoc mode is applied to 3/4G wireless data networks, VANETs and IEEE 802.11 WLANs for more seamless integration of heterogeneous wireless networks (see Fig. 1), we devise a route selection algorithm to forward data packets to the most appropriate AP/BS in order to maximize the same objective function as stated above. Results based on a detailed performance evaluation study are also presented here to demonstrate the efficacy of the proposed algorithms. It may be mentioned here that route selection algorithms have been previously studied [10]-[11] in the context of WLAN or cellular networks separately.

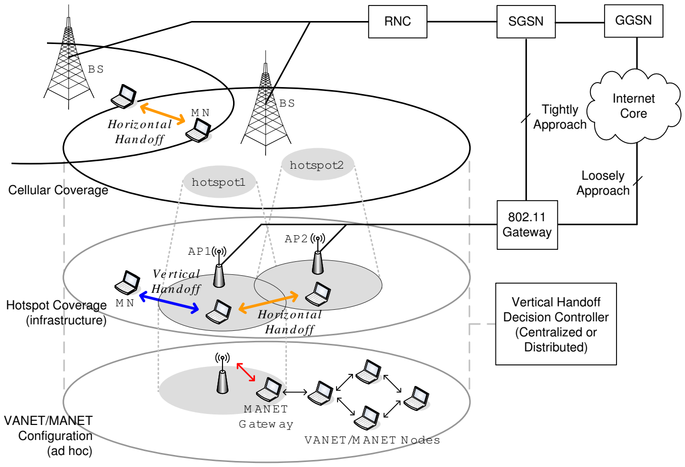


Fig. 1. Architecture of an integrated heterogeneous network consisting of WLAN, cellular and VANET/MANET.

The rest of the paper is organized as follows. We first describe our heterogeneous wireless networking system model in Section II. Then in Section III, we describe algorithms to select an appropriate attachment point while optimizing the system objective function. . In Section IV, we present a route selection algorithm in heterogeneous wireless networks that include an ad hoc network (such as VANET or MANET), while taking into account the amount of traffic to be forwarded and the load at attachment points in the route. In section V, extensive simulation results are presented and the performance of the proposed algorithms is evaluated. Finally, the conclusions are stated in Section VI.

II. Vertical Handoff Decision Algorithm

While wired networks are deterministic in nature, the user experience in a wireless network usually is known to be dependent on radio propagation and other building characteristics changing from one minute to the next. In WLAN environment coexisting with cellular networks, a choice of a more suitable attachment point among BSs as well as APs and the performance of the integrated system have an impact on the user experience, too.

It is not easy to determine the best position of antennas in WLAN, requiring extensive trial and error. Usually a better solution is to install multiple APs so that all work areas are covered strongly. Actually, the 802.11b specification allows for overlapping AP coverage areas. Moreover, these multiple APs coexist with a BS in a cellular coverage area. Here, note that usually one BS is supposed to cover a cellular coverage area but in extremely dense urban areas with high populations, some of the "cellular coverage areas" cannot be controlled by only one BS but by a few BSs. That is, a couple of 3G cells overlap unavoidably due to high user density in the area. Choice of an appropriate attachment point (BS or AP) for each MN together with vertical handoff capability would favorably impact the user QoS experience.

In traditional WLANs, each AP would make configuration decisions based solely on its own view of the network without taking into account any adjacency issues from other APs or BSs - whether on the same network or not. But our system model is based on choosing of an appropriate attachment point based on the characteristics of cellular network as well as each AP within reach. The goal is to optimize the overall performance of the integrated system, in terms of overall battery lifetime and load balancing.

Since the service area covered by one or a few BSs is generally larger than that of a WLAN, each cell in cellular networks is assumed to contain multiple WLAN hotspots [12]. Then, as in [6] and [9], we assume that there exists a Vertical Handoff Decision Controller (VHDC) in the "cellular coverage area" of GPRS/UMTS or cdma2000 with full coverage, wherein WLANs form small hotspots as shown in Fig. 1. That is, the cellular coverage area is defined to be the union of the areas covered by multiple APs while fully covered by one or a few BSs of GPRS/UMTS or cdma2000. Note that only if the cellular coverage area is highly dense, MNs can be serviced by a few BSs. Otherwise, the area is usually covered by one BS. The VHDC collects all the information about the heterogeneous system status and mobile users into some DataBases (DBs) [3][6][9] and decides which attachment point (WLAN or cellular network) MNs requesting a handoff should connect to. There would be client-side software, which is designed for multiple wireless systems and keeps monitoring the available wireless networks by instructing the wireless interface cards to scan for available networks and measure RSS periodically. In the future, distributed implementation of the VHDC may be envisioned with distributed VHDC (D-VHDC) software installed in each MN, especially when ad-hoc environments are supported, such as VANETs. With D-VHDC, the MNs not only obtain the network status information from the APs or the BSs in the available access networks but also share the information with one another. For simplicity, in the integrated WLAN and cellular networking system without ad hoc mode, only the centralized implementation of VHDC is investigated, with assumption of the tight-coupling approach for vertical handoff.

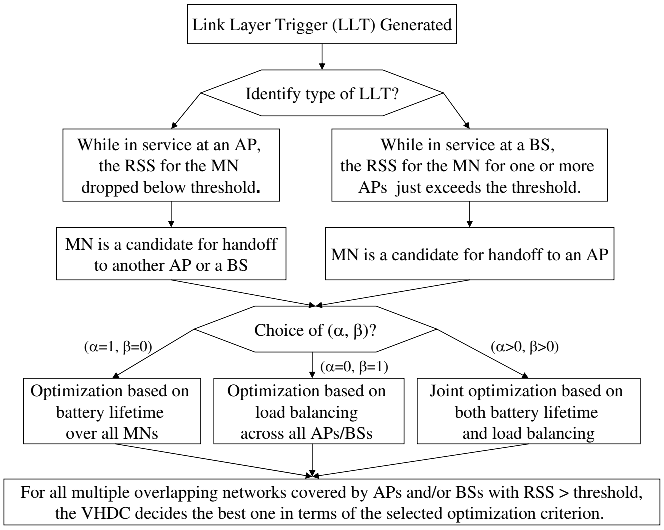

Link Layer Trigger (LLT) Generated
Identify type of LLT?
While in service at an AP, the RSS for the MN dropped below threshold.
While in service at a BS, the RSS for the MN for one or more APs just exceeds the threshold.
MN is a candidate for handoff to another AP or a BS
MN is a candidate for handoff to an AP
Choice of (α, β)?
(α=1, β=0)
(α=0, β=1)
(α>0, β>0)
Optimization based on battery lifetime over all MNs
Optimization based on load balancing across all APs/BSs
Joint optimization based on both battery lifetime and load balancing
For all multiple overlapping networks covered by APs and/or BSs with RSS > threshold, the VHDC decides the best one in terms of the selected optimization criterion.

Fig. 2. Flow chart for vertical handoff algorithm

A handoff may be requested by an MN or it might be triggered by a network (i.e., via VHDC) to optimize the overall network performance as well as mobile users' cost. This could be in the form of a periodic reconfiguration. To deal with handoffs, the VHDC implements the algorithm described in Fig. 2. In the proposed algorithm, if the current wireless network is a cellular network, the VHDC searches to see if a WLAN is available due to its higher speed and lower cost. If the current network is a WLAN and the RSS value is below a threshold, the VHDC tries to search for other networks. In the case that there exist multiple choices of APs of the same type, the VHDC evaluates the APs and then directs a handoff operation to the network with optimal cost. On the other hand, if no other APs are found, the cellular network is then considered the best available wireless network. Thus, the algorithm avoids giving unconditional high priority to WLAN access over cellular networks.

Here, we can choose a performance metric that could be battery lifetime over all MNs or load balancing across all APs/BSs, or a weighted combination of the two. One of these choices can be made by selecting the values of parameters, α and β, which will be explained in the next Section. As we see in Fig. 2, irrespective of current network type, the VHDC decides the most appropriate one from amongst multiple overlapping networks (covered by AP or BS), based on a criterion of optimizing system performance and user's cost.. It is commonly understood that a higher cost would be typically associated with the choice of cellular network over a WLAN (given that the RSS values for both are within threshold).

III. Attachment Point Selection Algorithm to Optimize the System Performance

In this section, details of the optimization techniques used in our vertical handoff decision algorithm and implemented in the VHDC are provided. The WLAN hotspots are typically

TABLE I

GLOSSARY OF VARIABLE DEFINITIONS

Variable	Definition
N	Number of APs
M	Number of BSs
$\bar{U}$	Set of all Mobile Node (MNs)
$\|\bar{U}\|$	Total number of MNs (K)
u_j	MN j $(1 \leq j \leq K)$
r_j	Bandwidth (i.e., data rate) requested by u_j $(1 \leq j \leq K)$
U_t	Set of MNs requesting VHO at time t
$\|U_t\|$	m(t)
V_t	$V_t = \bar{U} - U_t$
$V_t^{(a)}$	Set of MNs that have connection in a WLAN at t
$V_t^{(c)}$	Set of MNs that have connection in a cellular network at t
	$V_t = V_t^{(a)} + V_t^{(c)}$ or $V = V_a + V_c$ (dropping t)
	$U_t, V_t, V_t^{(a)}, V_t^{(c)} \Rightarrow U, V, V_a, V_c$ (dropping t)
$\|V_a\|$	Number of MNs that have a connection in a WLAN at time t
$\|V_c\|$	Number of MNs that have a connection in a a cellular network at time t
$w(i)$	Price/weight for WLAN and cellular network; w_a $(1 \leq i \leq N)$ and w_c $(N+1 \leq i \leq N+M)$
B_i	Maximum bandwidth which an AP a_i can provide
$B_{\hat{i}}^{(c)}$	Maximum bandwidth which a BS c_i $(N+1 \leq i \leq N+M)$ can provide
e_{ik}	Effective bandwidth of MN v_k when it belongs to $V_t^{(a)}$
$e_{ik}^{(c)}$	Effective bandwidth of MN v_k when it belongs to $V_t^{(c)}$
ρ_i	Load at AP a_i $(1 \leq i \leq N)$; Load at BS c_i $(N+1 \leq i \leq N+M)$
z_i	Maximal load which each AP a_i $(1 \leq i \leq N)$ or BS c_i $(N+1 \leq i \leq N+M)$ can tolerate
p_j	Available battery power of MN j
p_{ij}	Power consumption rate per unit time for MN j when attached to AP a_i
$p_{ij}^{(c)}$	Power consumption rate per unit time for MN j when attached to BS c_i
p_j^b	Power consumption amount per byte of transmission at MN j
RSS_{ij}	Received signal strength for MN j from AP a_i or BS c_i
θ_a	RSS threshold to connect to AP
θ_c	RSS threshold to connect to BS
$\hat{i}$	$i - N$

configured as small cells within the aforementioned "cellular coverage area" of GPRS/UMTS or cdma2000 which is relatively larger compared with WLAN hotspots as can be shown in Fig. 1. Since many variables are used in this paper, a glossary of variable names and definitions are provided in Table I.

Let $A = \{a_1, \cdots, a_N\}$ and $C = \{c_1, \cdots, c_M\}$ be the sets of APs in a cellular coverage area and BSs covering the cellular coverage area, respectively. Note that usually $M = 1$ except in the case of a highly dense urban deployment. Even when $M > 1$, M is much smaller than N because typically many APs are deployed within a cellular coverage area. The VHDC maintains the sets A and C covering the cellular coverage area as a list of candidate attachment points. It adds all available WLAN access points (APs) into the set A, and collects the information about load status on every AP in the set A and every BS in the set C. Note that in this section, we take account of only $a_i \in A$ $(1 \leq i \leq N)$ and $c_{\hat{i}} \in C$ $(1 \leq \hat{i} \leq M)$ whereas in the next section, each Mobile Node (MN) in ad hoc networking mode would also be considered as a possible attachment point.

In the cellular coverage area, $\bar{U} = \{u_1, \cdots, u_K\}$ is defined as the set of all MNs. Under mobile-initiated handoff, each MN is either requesting a handoff (or just turned on) or currently serviced by an AP ($\in A$) or BS ($\in C$) with no need for mobility at the time of optimization decision. Thus, the set $\bar{U}$ can be divided into the following two subsets at certain time t:

$$U_t = \{u_{n_1}, u_{n_2}, \cdots, u_{n_{m(t)}}\}$$

where $m(t)$ is the number of MNs requesting handoff at time t and $n_1, ..., n_{m(t)}$ are the corresponding indexes of those MNs, and

$$V_t = \bar{U} - U_t.$$

On the other hand, under network-initiated handoff, each attachment point measures the quality of the radio link channels being used by MNs in its service area. This is done periodically so that degradations in signal strength below a prescribed threshold can be detected and handoff to another attachment point can be initiated. Thus, under network-initiated handoff, U_t can denote the set of MNs that must be handed off, in accordance with VHDC determination, to another appropriate attachment point.

Each AP a_i and BS $c_{\hat{i}}$ are assumed to have a maximum bandwidth, B_i and $B_{\hat{i}}^{(c)}$, respectively. Let $\hat{i}$ denote $i - N$. Let $w(i)$ $(1 \leq i \leq N + M)$ denote the predefined costs or weights for the bandwidths of AP a_i $(1 \leq i \leq N)$ and BS $c_{\hat{i}}$ $(N + 1 \leq i \leq N + M)$. For simplicity, we define two different weights depending on whether the wireless access network is WLAN or cellular network. That is, for APs $a_i \in A$ $(1 \leq i \leq N)$, $w(i) = w_a$, and $w(i) = w_c$ for BSs $c_{\hat{i}}$ $(1 \leq \hat{i} \leq M)$. Each $a_i \in A$ has a limited transmission range and serves only users that reside in its range. If we use the periodic reconfiguration explained in section II, the VHDC will simultaneously consider all the MNs ($\in \bar{U} = U_t \cup V_t$) in the integrated system. The set V_t is divided into subsets $V_t^{(a)}$ and $V_t^{(c)}$ depending on whether $v_k \in V_t$ has a connection in a WLAN or a cellular network, respectively. Note that the $|U_t|$ MNs that are candidates for vertical handoff can belong in a WLAN or a cellular network, subsequent to the handoff decision.

For 802.11 products, it is known that an AP is able to maintain the average bit rate information for the MNs which are currently associated with it. Thus, each AP ($a_i \in A$) and BS ($c_{\hat{i}} \in C$) can maintain the effective data rate, e_{ik} and $e_{\hat{i}k}^{(c)}$ for MN v_k when it belongs to $V_t^{(a)}$ or $V_t^{(c)}$, respectively. However, for each MN $u_j \in U_t$, the AP to which the MN will hand off is not able to evaluate the effective data rate for the MN due to the absence of active signaling between the AP and the MN when they are not connected. Thus, a requested data rate, r_j is defined for each MN $u_j \in U_t$. Otherwise, if we assume that every MN is equipped with client software that periodically collects the bit rate information for every AP/BS in its neighborhood by using beacon messages/pilot bursts, it is possible to evaluate the effective bit rate, e_{ij} and $e_{\hat{i}j}^{(c)}$ from each AP $a_i \in A$ and BS $c_{\hat{i}} \in C$, respectively, to each MN $u_j \in U_t$. The collected information about the effective bit rate is reported to the VHDC where our proposed algorithm is run.

Since our proposed selection algorithms are performed at a certain time instant t, from now on, we shall omit the subscript t from U_t, V_t, $V_t^{(a)}$, and $V_t^{(c)}$ for notational convenience and for the clarity of understanding. Thus, U, V, V_a, and V_c will be used instead. Now, we define the load on AP a_i and on BS $c_{\hat{i}}$, ρ_i in a cellular coverage area as follows:

DEFINITION 1 *For each AP $a_i \in A$ ($1 \leq i \leq N$), the load on AP a_i is*

$$\rho_i = \sum_{v_k \in V_a} e_{ik}, \quad \textit{for } 1 \leq i \leq N \tag{1}$$

while the load on BS $c_{\hat{i}}$ is

$$\rho_i = \sum_{v_k \in V_c} e_{\hat{i}k}^{(c)}, \quad \textit{for } N+1 \leq i \leq N+M. \tag{2}$$

The above definition of load deliberately does not take into account the calls that are requesting handoff and will move away from the AP in consideration at the time decision is made. As a matter of fact, it is possible to compute ρ_i ($1 \leq i \leq N+M$) because an AP or BS is able to maintain the bit rate information for all the MNs connected to itself and also each MN knows its effective bit rate.

Associated with each MN u_j ($1 \leq j \leq K$) is a quantity p_j, denoting the available amount of power or the initial amount of power when it is just attached to a network, that could be maximum when the battery is fully charged. Let p_{ij} denote the power consumption per unit of time needed at MN u_j ($1 \leq j \leq K$) to reach an AP a_i ($1 \leq i \leq N$), that depends on the number of MNs attached to an AP and the data rate requested by MN. That is, the larger the number of power-on nodes attached to the same AP, the more power is consumed by each MN. With greater use of applications requiring higher data rate, the MN will consume power at higher rates. Thus, the amount of load at AP has an impact on the power consumed by MNs as $p_{ij} \propto \rho_i$. Similarly, $p_{\hat{i}j}^{(c)}$ ($\propto \rho_{N+\hat{i}}$) stands for the power level needed at MN u_j to reach BS $c_{\hat{i}}$.

When each MN u_j ($1 \leq j \leq K$) is associated with a certain AP a_i ($1 \leq i \leq N$) or BS $c_{\hat{i}}$ ($1 \leq \hat{i} \leq M$), a formal definition of battery lifetime matrix for MNs with respect to each attachment point in the cellular coverage area is given as follows:

DEFINITION 2 *Let $\mathbf{L} = \{l_{ij}\}_{(N+M)\times K}$ be the battery lifetime matrix where the matrix element, l_{ij} ($1 \leq i \leq N+M$) denotes the battery lifetime of u_j supposing that MN u_j hands off to AP a_i ($1 \leq i \leq N$) while $l_{(N+\hat{i})j}$ ($1 \leq \hat{i} \leq M$) is the battery lifetime of u_j in case that MN u_j hands off to BS $c_{\hat{i}}$. Then, for each MN u_j ($1 \leq j \leq K$), we have*

$$l_{ij} = \frac{p_j}{p_{ij}}, \qquad \textit{for } 1 \leq i \leq N \tag{3}$$

and

$$l_{ij} = \frac{p_j}{p_{\hat{i}j}^{(c)}}, \qquad \textit{for } N+1 \leq i \leq N+M \tag{4}$$

where it is assumed that every $l_{ij} > 0$ in this study.

Once the matrix $\mathbf{L}$ is computed and reported to VHDC, the VHDC decides which attachment point should be selected among the set A and C optimizing the overall battery lifetime cost for all MNs, to be defined formally later in this section. Based on the decision by VHDC, MNs requesting a handoff are covered by the selected attachment point with the optimal battery lifetime cost.

To formulate the optimal vertical handoff decision problem, binary variable x_{ij} is defined to have a value one ($x_{ij} = 1$) if user u_j is associated with AP a_i ($1 \leq i \leq N$) or BS $c_{\hat{i}}$ ($N+1 \leq i \leq N+M$) and zero ($x_{ij} = 0$) otherwise. Let RSS_{ij} ($1 \leq i \leq N+M$) be Received Signal Strength (RSS) for MN j from AP a_i and BS $c_{\hat{i}}$, respectively while θ_a and θ_c denote RSS thresholds for AP and BS, respectively. Then, we can define an association matrix $\mathbf{X}$ consisting of x_{ij} as follows:

DEFINITION 3 *Let $\mathbf{X} = \{x_{ij}\}_{(N+M)\times K}$ be an association matrix for a cellular coverage area such that*

$$\sum_{1 \leq i \leq N+M} x_{ij} = 1, \qquad \textit{for } 1 \leq j \leq K \tag{5}$$

$$x_{ij} \in \{0, 1\} \tag{6}$$

and

$$x_{ij} = 0 \textit{ if } RSS_{ij} < \begin{cases} \theta_a & \textit{for } 1 \leq i \leq N \\ \theta_c & \textit{for } N+1 \leq i \leq N+M \end{cases} \tag{7}$$

where x_{ij} ($1 \leq i \leq N$, $1 \leq j \leq K$) and $x_{(N+\hat{i})j}$ ($1 \leq \hat{i} \leq M$, $1 \leq j \leq K$) are a binary indicator which is 1 if and only if user u_j hands off to AP a_i and BS $c_{\hat{i}}$, respectively. Moreover, let $\mathcal{X}$ be the set of all association matrices.

The BSs collectively provide full coverage for the entire region of interest. The above DEFINITION 3 accordingly assures that each MN requesting handoff is covered by either a BS or an AP.

DEFINITION 4 *The battery lifetime of MN u_j for an association matrix $\mathbf{X} = \{x_{ij}\}$, $lt_j(\mathbf{X})$ is defined as*

$$lt_j(\mathbf{X}) = \sum_{1\le i\le N+M} l_{ij}x_{ij} \tag{8}$$

We can define the requested data rate on AP a_i, and on BS $c_{\hat{i}}$ for an arbitrary association matrix as follows:

DEFINITION 5 *Let γ_i ($1 \le i \le N+M$) denote requested data rate on AP a_i ($1 \le i \le N$) and BS $c_{\hat{i}}$ ($1 \le \hat{i} \le M$). Then, for any $\mathbf{X} = \{x_{ij}\} \in \mathcal{X}$,*

$$\gamma_i(\mathbf{X}) = \sum_{u_j\in U} r_j x_{ij} \tag{9}$$

In case each MN $u_j \in U$ is able to evaluate the effective bit rate e_{ij} and $e^{(c)}_{\hat{i}j}$ from the candidate APs (i.e., those with $RSS_{ij} > \theta_a$) and the candidate BS (i.e., those with $RSS_{ij} > \theta_c$), Eq. 9 in DEFINITION 5 is replaced by $\gamma_i(\mathbf{X}) = \sum_{1\le i\le N}\sum_{u_j\in U} e_{ij}x_{ij} + \sum_{1\le\hat{i}\le M}\sum_{u_j\in U} e^{(c)}_{\hat{i}j}x_{ij}$.

For the given battery lifetime matrix $\mathbf{L}$, we formulate the vertical handoff decision problem to maximize the battery lifetime (network wide) as follows:

$$\textbf{Max-}L : \text{Max}_{\forall\mathbf{X}\in\mathcal{X}} \sum_{1\le j\le K} lt_j(\mathbf{X}) \tag{10}$$

subject to

$$\rho_i + \gamma_i(\mathbf{X}) \le \begin{cases} B_i, & \text{for } 1 \le i \le N \\ B^{(c)}_{\hat{i}}, & \text{for } N+1 \le i \le N+M \end{cases} \tag{11}$$

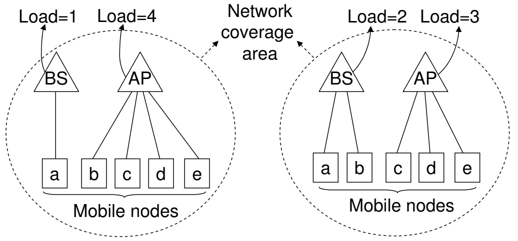
where the constraint in Eq. 11 ensures that the total load on each attachment point cannot exceed the maximum bandwidth supported by each AP or BS.

However, in the problem formulation **Max-**L, the total battery lifetime of the system is maximized without considering fairness with regard to individual battery lifetime of different MNs. Thus the max-min fairness is taken account of as follows:

$$\textbf{Max/Min-}L : \text{Max}_{\forall\mathbf{X}\in\mathcal{X}} \ \big(\ \text{Min}_{1\le j\le K} lt_j(\mathbf{X})\ \big) \tag{12}$$

subject to the same constraint as stated for **Max-**L in Eq. 11. While the earlier formulation of **Max-**L in Eq. 10 increases the total battery lifetime, it may in some situations compromise MNs with already lower remaining power. So we mention this alternative **Max/Min-L** formulation, but in this paper our focus is more towards joint optimization of battery lifetime and fairness in terms of distributedness of load at APs/BSs.

We now turn to the problem of distributing the overall load in a cellular coverage area. LEMMA 1 in the Appendix captures the fact that minimizing the sum of squared numbers is equivalent to minimizing the standard deviation of the numbers when the mean is constant. Since the standard deviation represents the degree of variation, we aim for the load per AP or BS in the cellular coverage area to stabilize around a mean value M with small deviations.



(a) p=1 (b) p=2

Fig. 3. Examples of load distribution when **Opt-**F is applied

PROPERTY 1 *The total load from all MNs of U, $\sum_{1\le i\le N+M}\gamma_i(\mathbf{X})$ does not change irrespective of what values $\mathbf{X}$ has. Thus, in a cellular coverage area, the expression $\frac{1}{N+M}\sum_{1\le i\le N+M} w(i)\Big(\frac{\rho_i+\gamma_i(\mathbf{X})}{z_i}\Big)$ also becomes invariant to the decision (i.e., $\mathbf{X}$) at the time of performing the optimization algorithm, where z_i is a maximal load which each AP or BS can tolerate.*

In the above PROPERTY 1, $z_i = B_i$ for $1 \le i \le N$ andor $z_i = B^{(c)}_{\hat{i}}$ for $N+1 \le i \le N+M$ (as in Eq. 11), noting that $\hat{i} = i - N$.

On the basis of LEMMA 1 and PROPERTY 1, we define a load-based cost function, F, and formulate the following optimization for distributedness of load:

$$\textbf{Opt-}F : \text{Min } F = \text{Min}_{\forall\mathbf{X}\in\mathcal{X}} \sum_{1\le i\le N} w(i)\Big(\frac{\rho_i+\gamma_i(\mathbf{X})}{z_i}\Big)^p \tag{13}$$

subject to

$$\rho_i + \gamma_i(\mathbf{X}) \le z_i, \qquad \text{for } 1 \le i \le N+M \tag{14}$$

where $p = 2$. Minimizing the cost function in Eq. 13 results in preventing BSs and APs with already higher load from being more congested.

Fig. 3 shows the examples of load distribution status resulting from introducing p $(= 2)$ into Eq. 13. Consider a cellular coverage area with two attachment points and 5 mobile users named from ‘a’ to ‘e’, where it is assumed that the weights of each MN for cellular network and WLAN are the same (i.e., $w_c = w_a$) and the maximal load at the AP and the BS are 5 (i.e., $z_1 = z_2 = 5$) for simplicity . Assume that the data rate to every MN is 1. When $p = 2$, the attachment points are selected as in Fig. 3-(b) because $(\frac{1}{5})^2 + (\frac{4}{5})^2 (= 0.68) > (\frac{2}{5})^2 + (\frac{3}{5})^2 (= 0.52)$ on the basis of Eq. 13. However, when $p = 1$, the attachment point selection may result in Fig. 3-(a) because $\frac{1}{5} + \frac{4}{5} (= 1) = \frac{2}{5} + \frac{3}{5} (= 1)$. That is, when $p = 1$, there is no difference between the two cases in Figs. 3-(a) and 3-(b) because the total load in the two cases is the same.

Thus, the cost function in Eq. 13 provides fairness from load balancing point of view when deciding an attachment point for an MN that requires handoff. Due to the fact that the wireless users associated with an AP share the buffer and bandwidth at the AP, the consideration of fairness works towards mitigating user congestion at APs.

In order to accomplish a joint optimization of the total battery lifetime and the fairness of load in a cellular coverage area, we formulate a combined cost function with parameters α and β as follows:

$$G(\mathbf{X},\alpha,\beta) = \alpha\sum_{1\leq j\leq L} lt_j(\mathbf{X}) - \beta \sum_{1\leq i\leq N+M} w(i)\Big(\frac{\rho_i+\gamma_i(\mathbf{X})}{z_i}\Big)^2 \quad (15)$$

Minimizing the cost function in Eq. 13 is equivalent to maximizing the negative of the same cost function because $\frac{\rho_i+\gamma_i(\mathbf{X})}{z_i} < 1$. Thus, we have the joint optimization statement of the total battery lifetime and the fairness of the load as follows:

$$\textbf{Opt-}G : \text{Max } G(\mathbf{X},\alpha,\beta) \quad (16)$$

with the constraints of Eq. 14. In Eq. 16, when $\alpha = 1$ and $\beta = 0$, it is evident that the Eq. 16 is an equivalent optimization problem of Eq. 10. Further, the optimization problem, $\text{Max}_{\forall\mathbf{X}\in\mathcal{X}}\ G(\mathbf{X},0,1)$ subjected to the constraint in Eq. 14, is equivalent to **Opt-**F.

The aforementioned algoritahms in this Section are normally meant to be performed at the time of link layer triggers for MNs (i.e., when handoffs are requested or desired). However, they can also be performed periodically to optimize the battery lifetime and the fairness of load in an area, resulting in network-initiated handoff.

IV. Optimization of Battery Lifetime of Heterogeneous Networks Including Ad Hoc mode

A. Integrated WLAN and Cellular Networking System Including Ad Hoc Networking Mode

Now we consider network architectures where, in addition to cellular networks and WLAN, peer-to-peer communications is further enabled using the IEEE 802.11 ad hoc mode [10]-[11] (also see Fig. 1). Now see seek to generalize the algorithm to select the most appropriate attachment point by considering the further selection of intermediate MNs to relay data packets to that attachment point.

In this system with ad hoc networking, cooperating MNs form a MANET/VANET using the IEEE 802.11 interface in an ad hoc mode. When an MN which is actively receiving data frames from a BS of cellular network or an AP, experiences low downlink channel rate and the VHDC cannot find an alternative direct attachment point (i.e., BS or AP) for the MN, a route will be selected via the MANET/VANET to allow the MN to access an appropriate attachment point.

The Dynamic Source Routing (DSR) technique [13][14] is used with suitable modification as the underlying route discovery protocol in our system. As shown in Fig. 4, the MN (i.e., source labeled as *src*) sends out a *route request* message using its IEEE 802.11 interface. This *route request* message is broadcast through the ad hoc network according to the *r*oute discovery protocol. The objevtive is to find an optimal relay route in terms of overall battery lifetime, to reach a node (i.e., proxy node) with high downlink channel rate to an attachment point. In order to prevent the *route request* message from being sent to all APs and the BS in a cellular coverage area, the number of hops is limited by using the Time-to-Live (TTL) field in the *route request*. Thus, the spread of a *route request* is controlled only to nearby attachment points located within pre-designated number of hops over the ad hoc network.

Again referring to Fig. 4, the candidate proxy node sends a *relay* message to its BS or AP. Once the BS or AP receives the *relay* message, the VHDC selects an attachment point and the best route to the attachment point based on the algorithm that is presented in Section IV-B. Here, the VHDC can be implemented in a distributed manner into every AP and BS, or can operate in a centralized way to make decisions about vertical handoffs on behalf of all APs and BS in a cellular coverage area. The selected attachment point updates its routing table entry for the MN while sending a *relay ack* message to the proxy node. And then the proxy node returns a *route reply* to the MN that initiated the *route request*. Accordingly, the selected BS or AP transmits the data frames to the proxy node over the downlink route maintained in its own routing table. When the proxy node receives a data frame from the BS or AP, it forwards the frame to the next relay node. This forwarding process continues via the IEEE 802.11 interfaces of all the relay nodes on the selected route till the MN receives the frame. And for the case when the downlink channel rate of the proxy node goes below a certain level due to its mobility, DSR is modified for the proxy node to piggyback its degraded downlink channel rate in data frames that are forwarded to the source MN so that it could begin another round of route discovery to find another optimal relay route.

We will now proceed to present the details of the algorithm for selection of the relay and proxy nodes that constitute the source MN's route toward a BS or an AP. This is the algorithm that the VHDC in Fig. 4 implements for selection of the best route in response to the *relay* messages received from the candidate proxy nodes.

B. Route Selection Algorithm to Optimize Battery Lifetime of System Including Ad Hoc Mode

For heterogeneous wireless networks, which include ad hoc networking, we aim to evenly balance over all MNs the battery power consumed in relaying traffic for others. As presented in Section III, in a heterogeneous wireless network without ad hoc support, the battery lifetime of each MN is considered to be related only to the RSS and the congestion (i.e., load) at its attachment point. However, in a heterogeneous wireless network supporting ad hoc mode, the amount of traffic each MN relays has a great impact on the MN's battery lifetime, and hence all MNs in the network must participate about equally in relaying each other's data frames. Thus, in this section, taking account of the amount of traffic load to be forwarded, we develop a route selection algorithm that maximizes the remaining battery life of "bottleneck" node which has the lowest residual energy. This results in maximizing the overall battery lifetime of the system as well.

We consider a finite population of K MNs in a cellular coverage area as in Section III. Let D be the amount of traffic in bytes that has to be routed via some MNs in the cellular coverage area. For MN $u_j \in U$ which experiences low downlink channel rate while receiving data frames from

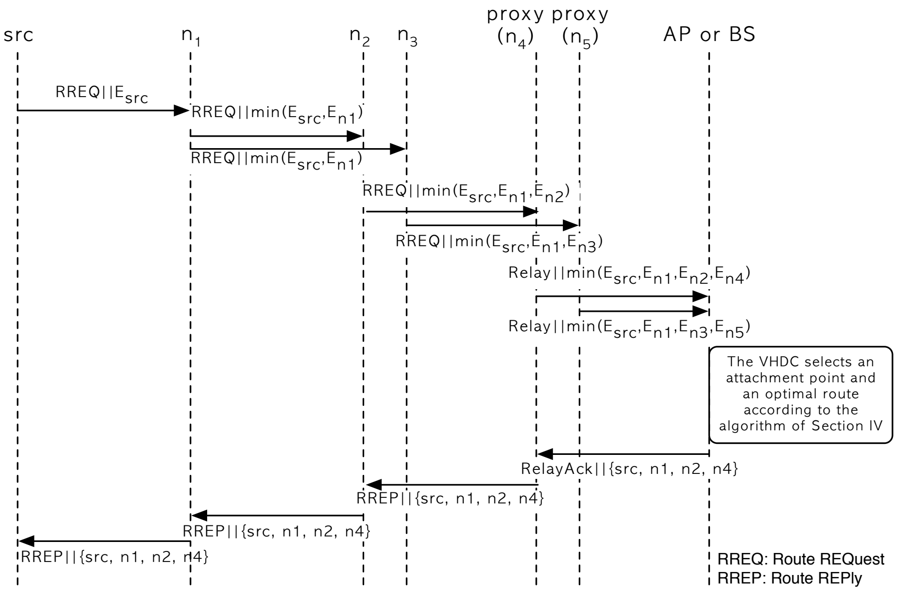


Fig. 4. Example procedure to discover a relay route to a proxy node, where a relay route {source, node n_1, node n_2, proxy node x_1} is selected.

a BS of cellular network or an AP, unless the VHDC can find an alternative point-to-point attachment point, a route will be selected by using ad hoc networking. Let p_j^b be the power consumption amount per byte of transmission at a given node u_j. Then, the cost function is defined as:

$$E_j = \frac{p_j}{p_j^b D}. \tag{17}$$

The maximum battery lifetime resulting from selection of a given route, r_s is determined by the minimum value of E_j over the path, that is:

$$L_s = \text{Min}_{\forall u_j \in r_s} E_j. \tag{18}$$

Let R be the set of all possible routes between the MN u_j that is experiencing degraded downlink channel rate and candidate attachment point. Thens, we select the route r_{max} with the maximum battery lifetime value from the set R as follows:

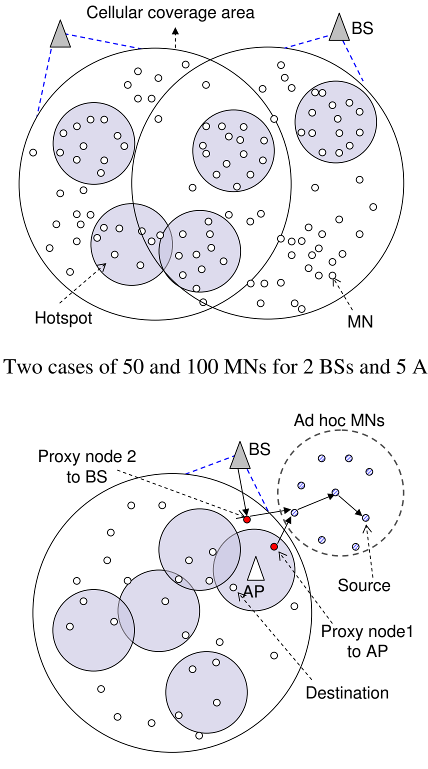
$$\begin{aligned} r_{max} : \ & \text{Max}_{\forall r_s \in R} L_s \\ & = \text{Max}_{\forall r_s \in R}\Big(\text{Min}_{\forall u_j \in r_s} \frac{p_j}{p_j^b D} \Big) \end{aligned} \tag{19}$$

When the route discovery process is triggered for an MN u_j that is experiencing low downlink channel rate, the battery lifetime information i.e. E_j, is sent encapsulated in the header of *rout request* message as a *cost* field. When a relay node u_i ($i \neq j$) receives the *route request* message, it calculates the value of E_i and compares it with the *cost* field in the received *route request*. If the calculated E_i is less than the value of the *cost* field, then E_i is copied into the *cost* field. This process is repeated until the *route request* message reached a BS or AP (see Fig. 4).

The above algorithm results in selecting the most appropriate proxy node and an associated optimal route (characterized by maximum battery life for the bottleneck node) to an attachment point via that proxy node.

V. Performance Evaluation

Eqs. 10, 13 and 16 in Section III are Mixed Integer Programming (MIP) formulations for battery lifetime maximization and load balancing. These MIP problems can be solved using the well known branch and bound algorithm [15]. First, we describe the simulation setup. Then we present the simulation results detailing the total battery lifetime over all MNs and the



(a) Two cases of 50 and 100 MNs for 2 BSs and 5 APs

(b) Network topology with ad hoc mode

Fig. 5. Simulation topologies of heterogeneous wireless networks.

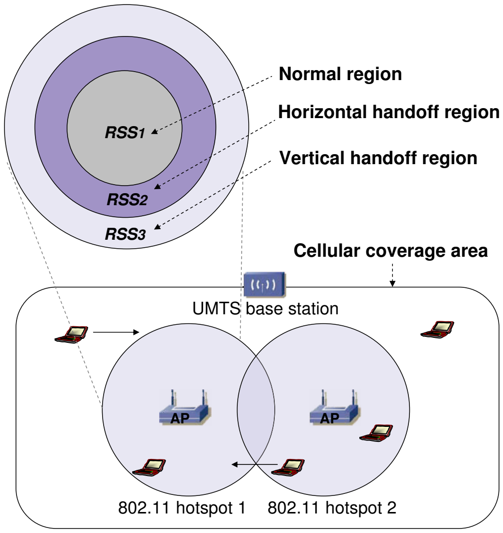


Fig. 6. An expanded look of test cellular coverage area for simulation

load distribution across all APs and BSs.

A. Simulation Environment

We conducted simulations for a cellular coverage area that is covered by two overlapping BSs and five hotspots as shown in Fig. 5-(a). Further, Fig. 5-(b) shows our simulation topology for the case when MANET/VANET is used as an enhancement to the cellular network and the hotspots. We simulated two test scenarios in which 50 and 100 MNs are dispersed, respectively, over the combined coverage area of the two BSs in the topology of Fig. 5-(a). Within the cellular coverage area, each hotspot area is conceptually divided into three different concentric areas as shown in Fig. 6. The innermost area, RSS_1, has the strongest RSS while the second area, RSS_2, which is outside RSS_1, has lower RSS than RSS_1. And the third area, RSS_3, representing the remaining portion of the hotspot area has the weakest RSS. As depicted in Fig. 6, RSS_2 region is potentially the horizontal handoff region, whereas RSS_3 is potentially the vertical handoff area. It should be noted that in realistic WLAN environments, RSS is highly variable over time even at a fixed location, depending on several known/unknown parameters such as multi-path, interference, local movements, etc. Therefore, in our simulation tests, we do not strictly go by Fig. 6, and hence each MN, regardless of whether it exists in RSS_1 or RSS_2 or RSS_3 regions, has its own randomly simulated RSS value, RSS_{ij}, where i and j denote the AP and the MN, respectively.

At the beginning of the simulation run, MNs are evenly distributed over all WLAN areas, and hence 10 MNs (first test case) or 20 MNs (second test case) are serviced by each AP. The MNs move around during the entire simulation time. A random mobility model is used to characterize the movement of MNs inside a cellular coverage area. The RSS_{ij} values for all pairs of MN and AP association are varied over time according to a pre-selected distribution.

The requested data rate of each MN, r_j can be one of the values from the set $\{$64 kbps, 128 kbps, 192 kbps$\}$. When a new connection arrives, the associated data rate is uniformly selected from the three allowed data rates. The battery power of an MN, p_j is initialized at the onset of its connection to the value of 3×10^3 Joules (J). The rates of consumption of MN j's battery power in association with AP i and with BS i are p_{ij} and $p_{ij}^{(c)}$, respectively and each is assumed to be exponentially distributed with a mean of 5 mJ/s [18]. The bandwidth capacities of each AP and each BS, B_i and $B_i^{(c)}$ are set to 20 Mbps and 2 Mbps, respectively. We set the weights (or prices) associated with AP and BS bandwidth usage, w_a and w_c, to values 1 and 10, respectively.

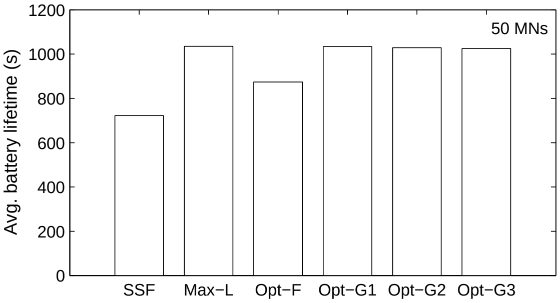


Fig. 7. Average battery lifetime for 50 MNs

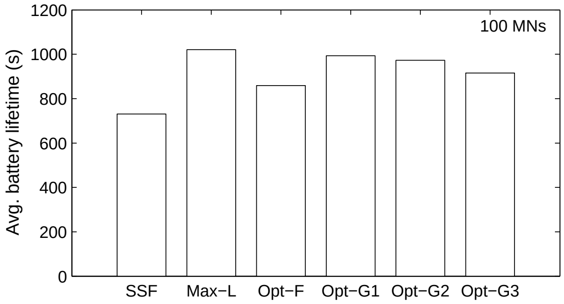


Fig. 8. Average battery lifetime for 100 MNs

In our experiment, we used the TOMLAB optimization package [16] and from the libraries thereof, CPLEX was used to solve the problem formulations described in Section III. We use the branch-and-bound algorithm in the CPLEX optimization package for solving the MIP optimization problems. We studied the battery lifetime and the evenness of load distribution for the two test cases. Ten independent simulation runs of duration 10,000s each were performed, measurements were taken at intervals of 1000s, and the results reported were averaged over the ten runs. Our two key performance metrics were measured over the simulation time considering all MNs, APs, and BSs involved in the two test cases.

B. Simulation Results

In this section, we present and discuss simulation results for the two topologies and two test cases described in Section V-A. We compare the performance of our methods that are based on new optimization criteria with that of an existing method, namely the Strongest-Signal First (SSF) method. The latter is the default user-AP association method in the IEEE 802.11 standard. The comparisons are presented in terms of overall system battery lifetime averaged over all MNs and

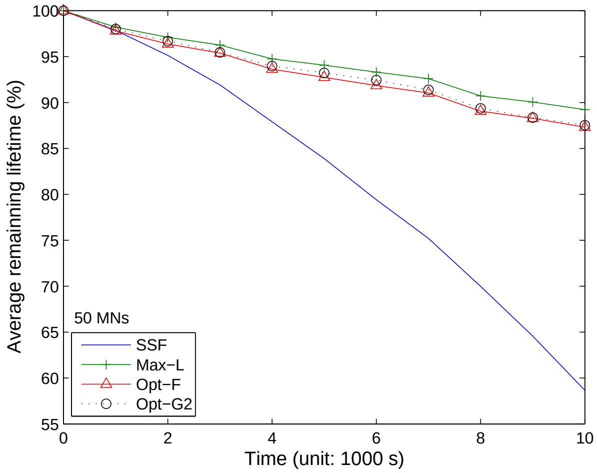


Fig. 9. Average battery lifetime versus time when there are 2 BSs, 5 APs and 50 MNs

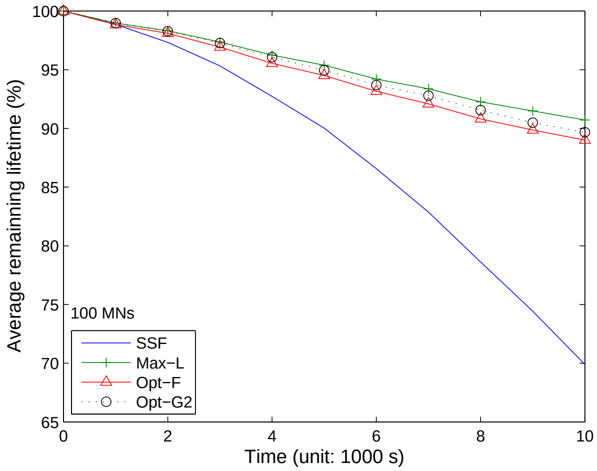


Fig. 10. Average battery lifetime versus time when there are 2 BSs, 5 APs and 100 MNs

distributedness of load among the attachment points (APs and BSs).

As stated earlier, for a given set of loads and MNs' battery lifetimes, the values of α and β in solving the joint optimization problem in Eqs. 15 and 16 can be selected appropriately to put different emphases on battery lifetime and load balancing. The values of the weights, α and β, would be typically supplied by the network operator or carrier responsible for maintenance of the network. For this study, based on some preliminary simulation runs with typical system and load parameters, we have determined that the first term in Eq. 15 (corresponding to battery lifetime) is typically about 5 orders of magnitude greater than the second term (corresponding to normalized load). This is naturally dependent on the measurement units used as well for each of the terms. Hence, for the joint optimization to work meaningfully, we must select β values to be in the ballpark of 10^5 those of α, and vary each in its respective range to study performance sensitivity to their values.

Figs. 7 and 8 show the average battery lifetime per MN for the two test cases (50 and 100 MNs, respectively) for various optimization methods. These methods include solving the battery lifetime optimization problem, **Max-L**, the load fairness optimization problem, **Opt-F**, and the joint optimization problem, **Opt-G**. For the joint optimization function, **Opt-G**, α and β are set such that $\frac{\beta}{\alpha}$=10^5, 3×10^5, and 5×10^5, which are denoted as **Opt-G_1**, **Opt-G_2**, and **Opt-G_3**, respectively, in Figs. 7-12. As we would expect, **Max-L** achieves the longest battery lifetime among all the cost functions or optimization methods in consideration (see Figs. 7 and 8).

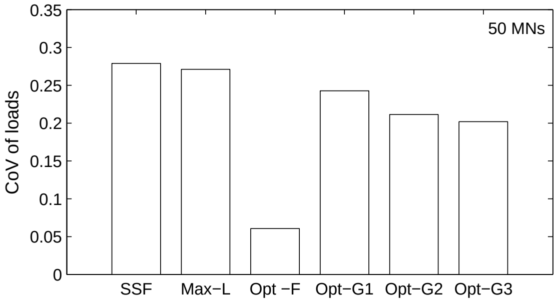


Fig. 11. Distributedness of load for 50 MNs

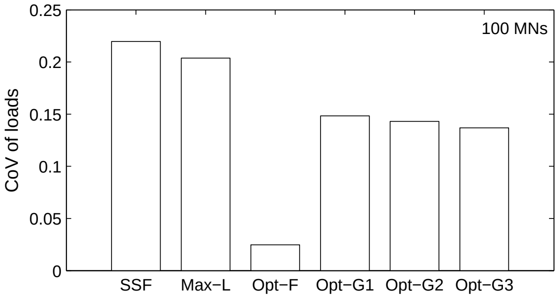


Fig. 12. Distributedness of load for 100 MNs

In Figs. 9 and 10, we plot the percentage remaining battery lifetime averaged over all MNs versus simulation time for the two test cases of 50 MNs and 100 MNs, respectively. We observe the same phenomenon as in Figs. 7 and 8. The battery lifetime for all the four schemes decreases with time. However, **Max-L** achieves the the best performance in terms of average remaining battery lifetime, while SSF performs the worst. In Figs. 11 and 12, we plot the Coefficient of Variation (CoV) of loads which is defined as the standard deviation of loads observed at the APs divided by the mean load. This definition has been used extensively as a fairness metric in the literature for illustration of the distributedness of load (i.e., load balancing) [17]. Figs. 11 and 12 show that **Opt-F** performs best among all the optimization methods as expected because **Opt-F** aims to evenly distribute the load among attachment points accessible by MNs in a cellular coverage area. However, for **Opt-F** method, the average battery lifetime is shorter compared to those for **Max-L**, **Opt-G_1**, **Opt-G_2**, and **Opt-G_3** as was noted in Figs. 7 and 8. SSF achieves the worst performance in terms of distributedness of load as well as battery lifetime. The weighted combined optimization method, **Opt-G** provides performance that lies in between those for **Max-L** and **Opt-F** in terms of either of the two performance

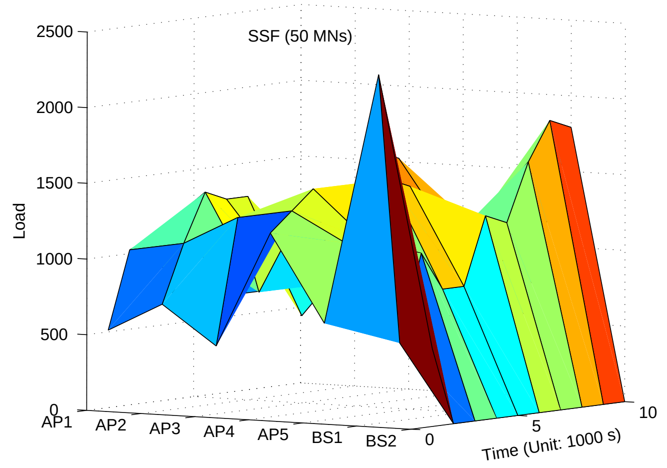


Fig. 13. For SSF, load status at APs versus simulation time

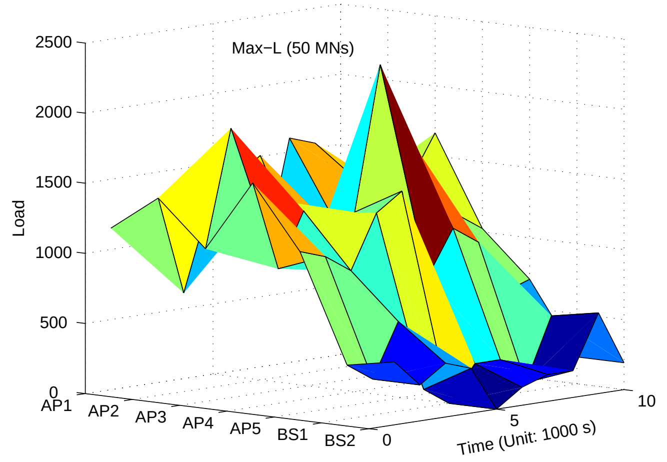


Fig. 14. For **Max-**L, load status at APs versus simulation time

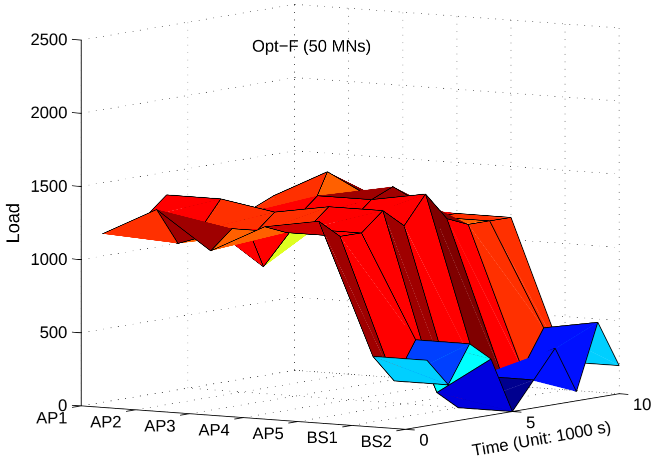


Fig. 15. For **Opt-**F, load status at APs versus simulation time when there are 50 MNs

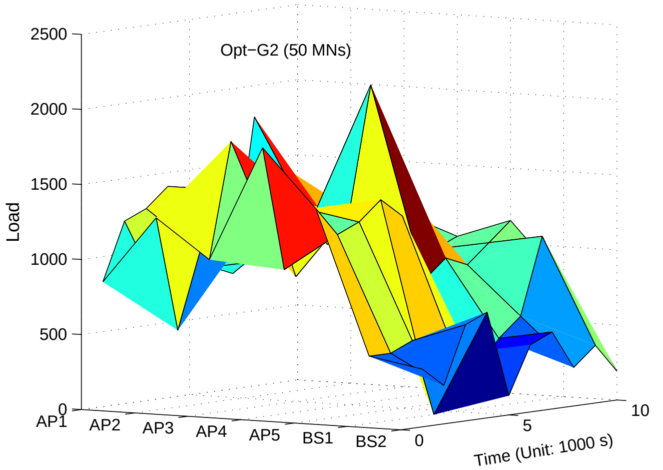


Fig. 16. For $G(\mathbf{X}, \alpha, \beta)$, load status at APs versus time when there are 50 MNs

metrics, i.e., battery lifetime or load fairness.

In Figs. 13-16, we plot the overall load at each AP and each BS versus simulation time for the first test case with 50 MNs active in the test coverage area. Similarly, the overall load for the second test case with 100 active MNs are plotted in Figs. 17-20. Theses figures show how the load is distributed among APs and BSs by the proposed cost functions as well as the SSF approach during the whole simulation time. As mentioned in section III, it is known that the price level for WLANs is cheaper than that for cellular networks. Thus, in our simulation tests, $w_c \gg w_a$ so that APs are selected in preference to BSs when an attachment point needs to be selected. That is, we aim to use cheaper WLAN bandwidth (especially, for multimedia traffic) in preference to the BS bandwidth. Through our proposed joint optimization method, the VHDC has the flexibility to manipulate the relative emphasis on extending battery lifetime vs. load balancing. We observe from Fig. 13 that under the SSF scheme, one of the five APs (AP4 in the graph) carries the maximum load of 1664 kbps at the simulation unit time 6000s and the maximum load of 2304 kbps is associated with one BS (BS1) at the time 1000s. We observe the SSF method does a poor job of not only distributing the load very unevenly across APs but also it favors BS1 at the expense of BS2 in the simulation test case with 50 MNs. On the other hand, for **Opt-**F method the load is quite evenly distributed over the APs within an approximate narrow range of 1400-1500 kbps, as we see in Fig. 15. Similar observations can be made for the test case with 100 MNs from the plots shown in Figs. 17 and 19. Based on the two sets of plots shown in Figs. 12-15 and Figs. 16-19, the following two other important observations can be made about the advantages of our proposed methods **Max-**L, **Opt-**F, and Max $G(\mathbf{X}, \alpha, \beta)$ over the SSF method: (1) These methods show lower preference for BSs over APs which are desirable since APs are better suited to carry higher-bandwidth multimedia calls, and (2) The parameters α and β can be suitably tuned by the network operator to achieve pure load balancing optimization or pure battery lifetime optimization or a suitable

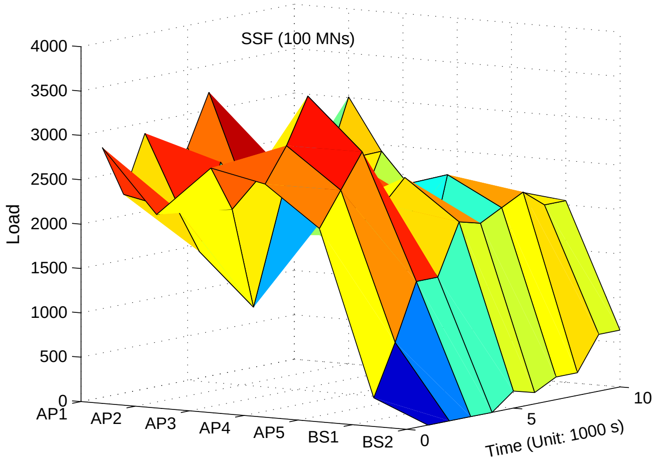


Fig. 17. For SSF, load status at APs versus time when there are 2 BSs, 5 APs and 100 MNs

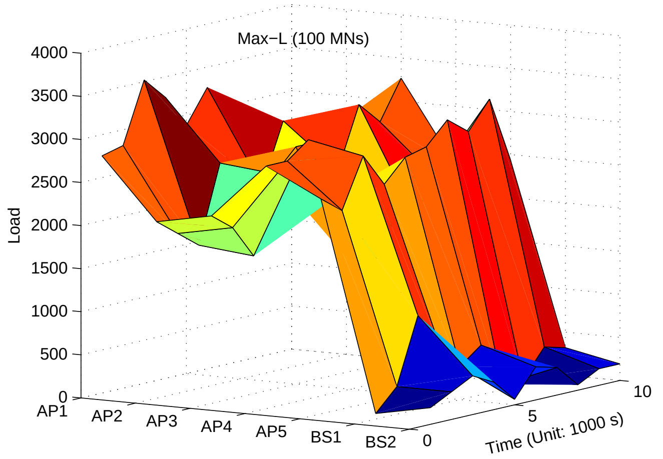


Fig. 18. For **Max-**L, load status at APs versus time when there are 2 BSs, 5 APs and 100 MNs

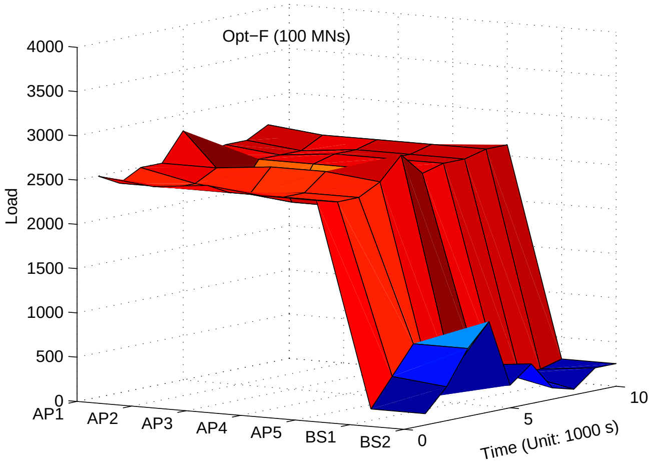


Fig. 19. For **Opt-**F, load status at APs versus time when there are 2 BSs, 5 APs and 100 MNs

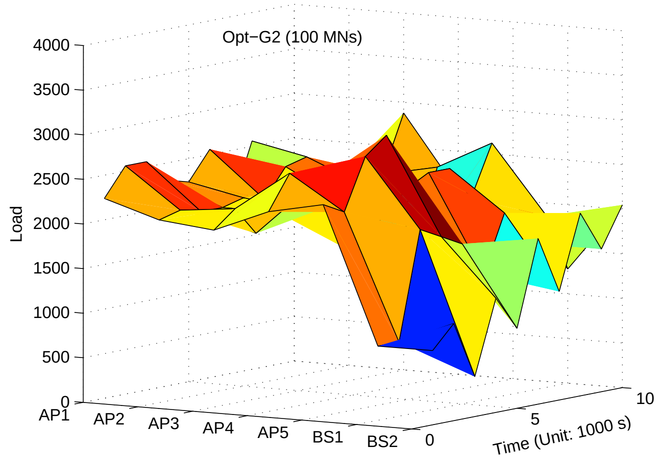


Fig. 20. For Max $G(\mathbf{X}, \alpha, \beta)$, load status at APs versus time when there are 2 BSs, 5 APs and 100

weighted combination of the two.

C. Battery Lifetime Results for Heterogeneous Networks Including Ad Hoc Mode

In this subsection, we compare the performance of the proposed route selection algorithm which is described in Section IV with that of DSR. The results presented here are obtained from the simulation model described in Section V-A (see Fig. 5 (b)), wherein the number of MNs in an ad hoc area is set to 40. As shown in Fig. 5 (b), the MNs operating in ad hoc mode are not within the coverage of any AP or BS, but are in range of each other via their short-range radios. This figure also shows two example routes from a source node (i.e., ad hoc mode MN) to two candidate proxy nodes; proxy node 1 reaches the destination via an AP and proxy node 2 does the same via a BS. The route selection algorithm, r_{max}, proposed in Section IV may typically select a different route than that selected by DSR algorithm, because our r_{max} algorithm is enhanced to take into account the battery lifetime of the route.

In our simulation runs, a pair of nodes consisting of one each in the ad hoc and cellular coverage areas are selected randomly as the source and destination nodes, respectively. Five such pairs of nodes are selected per 1000s of time, and one connection is generated each time. All the MNs are randomly distributed and move randomly. When they move, a new route is selected between the pair of nodes if the current route becomes unusable due to the movement and power considerations. The amount of data sent per connection from the source node is exponentially distributed with mean D Kbytes per connection. D is set to one of these three values: 5, 10, and 15 Kbytes. The initial energy of each MN is 1000 mJ. We use the power consumption model developed in [18] for the WLAN interface, where the energy consumed by a network interface as it sends and receives point-to-point messages, is described as $0.8\text{mJ} + 2.4\text{mJ/Kbyte} \times D$. Ten independent simulation runs of duration 20,000s each are performed, measurements are taken at intervals of 1000s, and the results reported are averaged over the ten runs.

TABLE II

(A) The remaining energy (mJ) of each proxy node after 20,000s of simulation time, and (b) The average CVE for the proxy nodes, measured at 1000s intervals and averaged over simulation time

(a) Remaining energy (mJ) of each proxy node

		$D = 5$	
	Proxy 1	Proxy 2	Proxy 3
DSR	743.3	1000	1000
r_{max}	913.8	910.3	919.3
		$D = 10$	
	Proxy 1	Proxy 2	Proxy 3
DSR	456.8	1000	1000
r_{max}	812.4	806.4	838
		$D = 15$	
	Proxy 1	Proxy 2	Proxy 3
DSR	254	1000	1000
r_{max}	702.6	783.5	768

(b) Average CVE

D	5	10	15
DSR	0.08	0.18	0.27
r_{max}	0.018	0.043	0.077
$\frac{\text{CVE of DSR}}{\text{CVE of } r_{max}}$	4.44	4.19	3.51

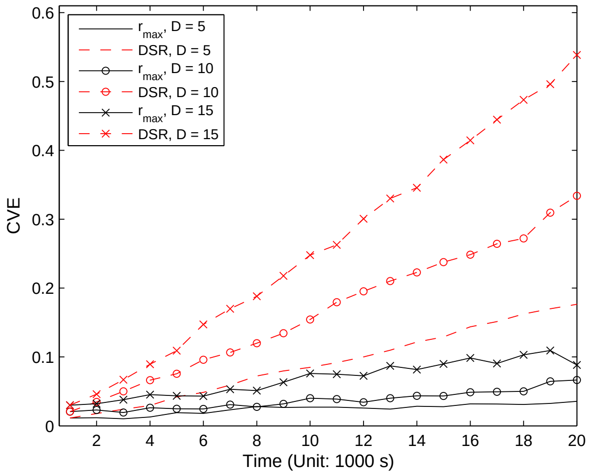


Fig. 21. CVE for the remaining energy of the three proxy nodes

Here, our focus is on the power consumed by proxy nodes while forwarding packets on behalf of other nodes in the ad hoc area. The proposed algorithm, r_{max}, aims to improve the longevity of the network by making the battery life for proxy MNs last longer. Accordingly, in this algorithm the load gets balanced over all accessible proxy nodes so that MNs in the ad hoc area would be able to sustain connectivity for longer period with MNs outside the ad hoc area. The performance of our proposed r_{max} algorithm and that of DSR are compared in Table II. This table compares the two algorithms using two separate metrics: (1) The remaining energy of each of the three available proxy nodes at the end of simulation run length of 20,000s, and (2) The covariance of remaining energy, CVE, for the three proxy nodes (measured at intervals of 1000s and averaged over the simulation run). It is evident that the r_{max} algorithm performs consistently better than DSR. As an example, in Table II-(a) we see that for the case of D = 15 Kbytes, the remaining energy of Proxy 1 is 254 mJ and 702.6 mJ, respectively, for the DSR and r_{max} algorithms. This effectively means that the probability that Proxy 1 MN will shut off is much higher for DSR as compared to that for the r_{max} algorithm. In essence, the proposed r_{max} algorithm distributes the load evenly across the three proxy MNs so that each has about equal remaining energy. This is further illustrated in Table II-(b) and Fig. 21 by comparing the CVE values for DSR and r_{max} algorithms. The improvement in CVE values for the r_{max} algorithm over DSR are quite significant, and are lower by factors of 4.44, 4.19, and 3.51 for the cases of D = 5, 10, and 15 Kbytes, respectively.

VI. Conclusion

When connections need to migrate between heterogeneous networks for performance and high-availability reasons, then seamless vertical handoff is a necessary first step. In the near future, vehicular and other mobile applications will expect seamless vertical handoff between heterogeneous access networks, which will include VANETs/MANETs.

New metrics for vertical handoff continue to emerge and the use of new metrics make the vertical handoff decision process increasingly more complex. In this paper, we tried to highlight the metrics best suited for the vertical handoff decisions. We also proposed a generalized vertical handoff decision algorithm that seeks to optimize a combined cost function involving battery lifetime of MNs and load balancing over APs/BSs. We further proposed an enhanced algorithm for the case when ad hoc mode MNs forming VANET/MANET are included in the heterogeneous networks. This latter algorithm allows the proxy nodes, which provide connectivity to the nearest AP or BS for the ad hoc mode MNs, to share transit loads with the goal of balancing their consumption of battery power. Our performance results based on detailed simulations illustrate that the proposed algorithms perform much better than the conventional optimization based on the SSF method, which is based on RSS alone. Our proposed method gives the network operator the leverage to easily vary the emphasis from maximizing the overall system battery lifetime for MNs to seeking fairness of load distribution over APs and BSs, with weighted combinations in-between.

References

[1] 3GPP TR 23.234 v7.1.0, "3GPP System to WLAN Interworking; System Description (Release 7)", March 2006, http://www.3gpp.org/specs/specs.htm.

[2] A.K. Salkintzis, "Internetworking Techniques and Architectures for WLAN/3G Integration Toward 4G Mobile Data Networks", *IEEE Wireless Communications*, June 2004.

[3] N. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, S. Miller, and L. Salgarellim, "Integration of 802.11 and Third-Generation Wireless Data Networks", *IEEE Proceedings of Infocom*, 2003.

[4] V. Varma, S. Ramesh, K. Wong, and J. Friedhoffer, "Mobility Management in Integrated UMTS/WLAN Networks", *IEEE Proceedings of ICC*, 2003.

[5] T.B. Zahariadis, "Guest Editorial: Migration toward 4G Wireless Communications", *IEEE Wireless Communications Magine*, June 2004.

[6] J. McNair and F. Zhu, "Vertical Handoffs in Fourth-Generation Multinetwork Environments", *IEEE Wireless Communications Magine*, June 2004.

[7] Chuanxiong Guo, Zihua Guo, Qian Zhang and Wenwu Zhu, "A Seamless and Proactive End-to-End Mobility Solution for Roaming across Heterogeneous Wireless Networks", *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 834-848, June 2004.

[8] R. Chakravorty, P. Vidales, K. Subramanian, I. Pratt, and J. Crowcroft, "Performance Issues with Vertical Handovers - Experiences from GPRS Cellular and WLAN Hot-spots Integration", *IEEE Proceedings Percom: Pervasive Computing and Communications*, 2004

[9] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in Fourth Generation Heterogeneous Networks", *IEEE Communications Mag.*, vol. 44, no. 10, pp. 96-103, October 2006.

[10] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated Cellular and Ad Hoc Relaying Systems: iCAR", *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, October 2001.

[11] D. Cavalcanti, D. Agrawal, C. Corderio, B. Xie and A. Kumar "Issues in Ingetrating Cellular Networks, WLANs, and MENETs: A Futuristic Heterogeneous Wireless Network", *IEEE Wireless Communcations Magazine*, vol. 12, no. 3, pp. 30-41, June 2005.

[12] A. Doufexi, S. Armour, and A. Molina, "Hotspot Wireless LANs to Enhance the Performance of 3G and Beyond Cellular Networks", *IEEE Communications Magazine*, vol. 41, no. 7, pp. 58-65, July 2003.

[13] D. Johnson, D. Maltz, and Y. Hu, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)", *Internet Draft*, draft-ietf-manet-dsr-10.txt, July 2004.

[14] H. Luo, R. Ramjee, P. Sinha, L. Li, and S. Lu, "UCAN: A Unified Cellular and Ad-Hoc Network Architecture", *Proceedings of ACM Mobicom'03*, September 2003.

[15] R. Fletcher and S. Leyffer, "Numerical Experience with Lower Bounds for MIQP Branch-And-Bound", *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 604-616, 1998.

[16] TOMLAB: A General Purpose MATLAB Environment for Optimization, http://tomlab.biz.com

[17] R. Jain, "The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling," *Wiley- Interscience*, New York, NY, April 1991.

[18] L. Feeney and M. Nilsson, "Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment," *Proceedings of IEEE Infocom'01*, 2001.

APPENDIX

In order to take into account the fairness of load distribution, a simple but useful lemma is provided as follows:

LEMMA 1 *Let $\{b_i\}_{i=1}^{I}$ be a finite sequence of real numbers and $A = \frac{1}{I}\sum_{i=1}^{I} b_i$ the mean value of the sequence. Then,*

$$\sum_{i=1}^{I} b_i^2 = \sum_{i=1}^{I} (b_i - A)^2 + IA^2. \tag{20}$$

PROOF. Since $\sum_{i=1}^{I}(b_i - A) = 0$,

$$\begin{aligned}\sum_{i=1}^{I} b_i^2 &= \sum_{i=1}^{I}[(b_i - A) + A]^2 \\ &= \sum_{i=1}^{I}[(b_i - A)^2 + 2A(b_i - A) + A^2] \\ &= \sum_{i=1}^{I}(b_i - A)^2 + 2A\sum_{i=1}^{I}(b_i - A) + \sum_{i=1}^{I} A^2 \\ &= \sum_{i=1}^{I}(b_i - A)^2 + IA^2.\end{aligned} \tag{21}$$