A QoS Oriented Vertical Handoff Scheme for WiMAX/WLAN Overlay Networks

Dong Ma, *Student Member*, *IEEE*, and Maode Ma, *Senior Member*, *IEEE*

Abstract—Recently, a number of wireless communication technologies are migrating toward heterogeneous overlay networks. The integration of Mobile WiMAX and WLAN seems to be a promising approach due to their homogeneous nature and complementary characteristics. In this paper, we investigate several important issues for the interworking of Mobile WiMAX and WLAN networks. We address a tightly coupled interworking architecture. Further, a seamless and proactive vertical handoff scheme is designed based on the architecture with aims to provide always the best quality of service (QoS) for users. Both the performance of applications and network conditions are considered in the handoff process. Moreover, we derive evaluation algorithms to estimate the conditions of both WiMAX and WLAN networks in terms of available bandwidth and packet delay. A simulation study has demonstrated that the proposed schemes can keep stations always being best connected.

Index Terms—WiMAX, WLAN, tightly coupled, vertical handoff, available bandwidth, packet delay.

1 Introduction

The trend in fourth-generation wireless networks (4G) is the coexistence of heterogeneous technologies. During the past few years, wireless local area networks (WLANs) have been widely deployed due to its low cost and high capacity. On the other hand, mobile worldwide interoperability for microwave access (Mobile WiMAX) networks become a fast growing technology for its promised high bandwidth over long-range transmission with quality of service (QoS) supports. The integration of WiMAX and WLAN has been seen as a promising approach toward 4G.

In the design of heterogeneous overlay systems, one of the most important issues is vertical (intersystem) handoff (VHO) support. Generally, traditional horizontal (intrasystem) handoffs are initiated only by mobility to maintain the connectivity of the station. However, more metrics may be considered in VHOs especially when more than one network is available. These metrics can be classified into two categories [1]. One category is *QoS*. If the service provided by the connected network cannot satisfy the requirements, the station may switch to another network for better performance. The other category is *user preference* which reflects the user's special requirements on price, power consumption or speed, etc. Therefore, VHO plays a significant role in achieving the main goal of 4G networks—allowing users to profit always best connected (ABC) service.

For VHO schemes, "seamless" and "proactive" are two desirable features [2]. A proactive handoff means that the handoff process (i.e., initiation, decision, and execution) is controlled by the stations. Hence, if QoS metrics such as bandwidth and packet delay are considered in a VHO scheme, the stations should be able to detect network conditions for a handoff decision. Consequently, the network condition detection algorithms need to be tightly integrated into QoS oriented VHO schemes [3]. On the other hand, a seamless handoff denotes that the handoff execution is transparent to upper layer applications. Indeed, this depends on the interworking architecture of heterogeneous networks. In the existing cellular/WLAN overlay systems, there are two types of interworking architectures: tightly coupled where WLAN works as a radio access network of cellular system, and loosely coupled where different networks are independently deployed but integrated at network layer. Comparably, a more seamless VHO can be expected in the tightly coupled networks, where the handoff execution follows the protocols of cellular networks and Mobile IP is not necessary while which is commonly deployed in the loosely coupled networks [4]. As a result, undesirable signaling cost induced by Mobile IP can be avoided in the tightly coupled cases.

Due to the newly developed WiMAX, there have been some, but still limited proposals made for VHOs in WiMAX/WLAN overlay networks. The works in [5] and [6] addressed the protocol related issues, both of which used Mobile IP to maintain the active connections during handoff execution process. The networks were integrated in a loosely coupled manner although it was not definitely addressed. User preference related metrics have been studied relatively well in literatures [7], [8], [9], [10]. In [11], [12], and [13], QoS metrics were taken into account in the handoff decisions. However, the network condition detection schemes have not been provided. In [14] and our paper [15], QoS aware handoff solutions were addressed. But only bandwidth was taken as the QoS metric. Meanwhile, the details on the network integration as well as the handoff management have not been fully addressed.

In this paper, we investigate the integration and VHO issues in WiMAX/WLAN overlay networks. The major

- *The authors are with the Division of Communications, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. E-mail: mado0001@e.ntu.edu.sg, Maode_Ma@pmail.ntu.edu.sg.*

Manuscript received 1 Aug. 2010; revised 15 Apr. 2011; accepted 16 Apr. 2011; published online 5 Aug. 2011.
Recommended for acceptance by S. Papavassiliou.
For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number TPDS-2010-08-0459.
Digital Object Identifier no. 10.1109/TPDS.2011.216.

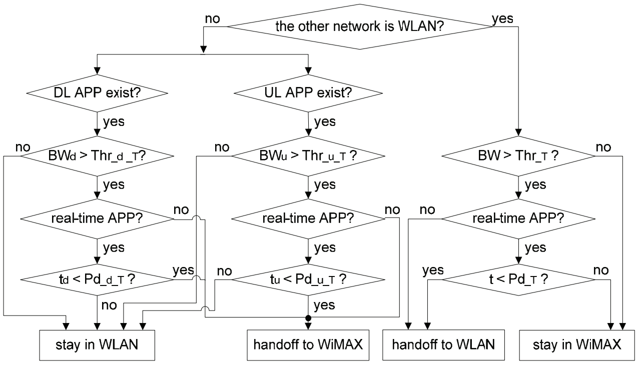

Fig. 1. The QoS-triggered VHO decision algorithm.

TABLE 1
Parameters Used in VHO Decision Algorithm

APP	current application
$BW_d/BW_u/BW$	estimated available bandwidth of DL WiMAX/UL WiMAX/WLAN
$Thr_{\_d\_T}$/ $Thr_{\_u\_T}$/ $Thr_{\_T}$	threshold for the available bandwidth of DL WiMAX/UL WiMAX/WLAN
$t_d/t_u/t$	estimated average packet delay of DL WiMAX/UL WiMAX/WLAN
$Pd_{\_d\_T}$/ $Pd_{\_u\_T}$/ $Pd_{\_T}$	threshold for the average packet delay of DL WiMAX/UL WiMAX/WLAN

contributions of our work are threefold: 1) a QoS oriented VHO scheme is proposed for the tightly coupled WiMAX/WLAN networks to provide the ABC services for both mobile and fixed users; 2) in order to achieve proactive handoffs, network condition detection algorithms are derived for stations to estimate the available bandwidth and the packet delay of WiMAX and WLAN networks, respectively; and 3) since to our knowledge, there is still no tightly coupled architecture dedicatedly designed in literatures for WiMAX/WLAN systems, we address an architecture to support our VHO scheme. To avoid losing focus, the interworking mechanisms are provided in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.216.

The remainder of this paper is organized as follows. In Section 2, the VHO scheme of our design is described. And then the available bandwidth and packet delay estimation algorithms are derived in Section 3. Section 4 presents the performance evaluation to show the effectiveness and the feasibility of our schemes. Finally, the paper is concluded and summarized in Section 5.

2 Proposed VHO Scheme

In existing QoS oriented VHO approaches (e.g., [3], [13], [16]) for overlay networks, QoS metrics are considered in handoff decisions. However, the handoff procedures are normally initiated when the stations move across the border of WLANs. As a result, both the fixed stations and the mobile stations within overlapped areas cannot benefit from VHOs. We highlight that a VHO could be initiated by two factors: mobility when a station moves out of the coverage of its connected WLAN, and QoS when the connected network cannot satisfy the requirements. Therefore, the QoS-triggered handoffs should be designed with an objective to provide ABC services for both mobile and fixed stations.

To achieve a proactive handoff, we design a VHO manager (VHOM) to control the whole handoff process, which works on the medium access control (MAC) layers of WiMAX and WLAN interfaces at the station. In the following analysis, we focus on basic operations for the QoS-triggered VHOs. The components of VHOM and the implementation details of the complete scheme, e.g., mobility-triggered VHO and methods to save power and avoid packet loss and ping-pong effect, are given in Appendix B available in the online supplemental material.

2.1 Service Evaluation and Handoff Initiation

Once an application is established at the station, VHOM will detect each packet of this application. Based on the delay sensitivity characteristics, the applications are classified into real-time applications and nonreal-time applications, respectively. Since a real-time application is sensitive to latency, both the throughput and packet delay of the traffic are measured. For a nonreal-time application, the amount of transmission data is more important, and then only the traffic throughput is measured.

Here, the transmission direction of the application should be taken into account. For an uplink (UL) application, VHOM can record the moment that the packet arrives at the MAC layer buffer of the station and the moment that the packet is successfully transmitted by the connected network. Therefore, the calculated UL traffic throughput and packet delay can well reflect the performance of the local connected network (WiMAX or WLAN). If an UL application continually violates the QoS requirements for a given period, the handoff decision process will be started.

For the downlink (DL) traffic, however, the station cannot obtain the time information that the packet arrives at base station (BS) or access point (AP), and then an end-to-end delay will be calculated in this case rather than the delay purely introduced by the local connected network. Meanwhile, if it is the DL traffic throughput or packet delay violates the QoS requirements, the poor performance may be introduced by the local connected network or by other networks on the path between two end nodes of this application. To avoid performing an unnecessary VHO within the local network, VHOM needs to evaluate the conditions of the local connected network first in this case. Only when the local connected network is proved to be working in a bad condition, the following handoff decision process can be started.

2.2 Network Condition Detection and Handoff Decision

In this phase, a decision of whether to perform a VHO will be made by VHOM. The main work is to decide whether the conditions of the other network that is not serving the station can satisfy the QoS requirements. The flowchart is shown in Fig. 1 with the used parameters listed in Table 1. The available bandwidth of the network is evaluated first. If the calculated result is larger than the threshold and a real-time application is running on the station, the average packet delay of the network will be further estimated. If the other network is WLAN, the network conditions are estimated based on the total radio resource since the medium of WLAN is contended by all stations including AP. But in WiMAX networks, the radio resource has been allocated into a DL portion and an UL portion, which allows the network conditions be evaluated for DL and UL, separately. Therefore, only when both the DL and UL network conditions satisfy the requirements, a decision of handoff to WiMAX can be made.

To make an effective handoff, it is required that the conditions of the target network must be good enough. This is guaranteed by accurately estimating the network conditions and setting suitable thresholds. We have designed novel algorithms for both WiMAX and WLAN networks to detect the network conditions in terms of the available bandwidth and the packet delay, which will be described in Section 3 and proved to be accurate in Section 4. By the proposed scheme, it is required that the thresholds for evaluating the available bandwidth (i.e., $\mathrm{Thr}_{\_d\_T}, \mathrm{Thr}_{\_u\_T}$ and $\mathrm{Thr}_{\_T}$) should not be less than the total throughput of the corresponding applications. Meanwhile, the thresholds for the packet delay (i.e., $\mathrm{Pd}_{\_d\_T}, \mathrm{Pd}_{\_u\_T}$ and $\mathrm{Pd}_{\_T}$) should not be larger than the required lowest packet delay for the corresponding applications.

2.3 Handoff Execution

Once a decision of handoff to the other network is made, VHOM needs to transfer current connections at the station to the target network. By existing VHO schemes (e.g., [5], [6]), this operation is usually executed by network layer approach—Mobile IP where signaling overhead cannot be neglected. However, our work is based on the proposed tightly coupled interworking architecture, which is described in Appendix A available in the online supplemental material. Under this architecture, AP connects to the central gateway called ASN GW in the access service network (ASN) of WiMAX, just as its overlay BS. Therefore, AP belongs to the same subnetwork at IP layer with its overlay BS. As a result, the IP address of the station needs not to be changed after a VHO, which makes a MAC layer handoff possible.

Based on this consideration, we deploy an address resolution protocol (ARP) [17] method to execute VHOs. When a handoff decision is made, VHOM issues a gratuitous ARP message which will be transmitted by the target interface at the station. The message conveys the IP address of the station and the MAC address of the target interface. The AP or BS in the target network relays the message to ASN GW, which then updates its ARP cache by binding the IP address with the MAC address contained in the message. Then, ASN GW issues an ARP reply message to the station. Hereafter, the data packets destined to the station will be transferred by ASN GW via the newly selected network.

It can be found that the handoff in our scheme is executed by a round message trip from the station to ASN GW, which is much faster than Mobile IP-based methods since the signaling overhead to the home network is avoided. Meanwhile, a seamless handoff is achieved by our approach since the operations are performed at MAC layer and then transparent to upper layers.

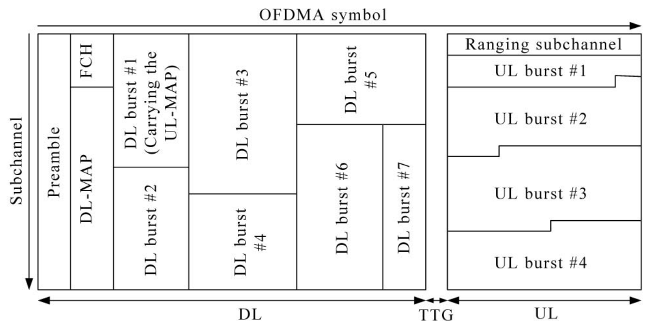

Fig. 2. A sample of OFDMA-TDD frame structure.

3 Proposed Estimation Algorithms

The available bandwidth of a link equals to the difference between total capacity and the traffic load over the link [18]. Usually, when the modulation and coding methods are known, the total capacity can be calculated. Therefore, the key idea of estimation is to find the utilization information of the link. In this section, we develop our scheme to estimate the utilization of WiMAX and WLAN networks and further derive the available bandwidth and the packet delay in the networks based on this information.

3.1 Estimation in WiMAX

Mobile WiMAX is specified by IEEE 802.16e [19] standard which uses orthogonal frequency division multiple access (OFDMA) technique. Both the time division duplexing (TDD) and frequency division duplexing (FDD) are supported by IEEE 802.16e. In the following algorithms, TDD is selected as an example. An OFDMA-TDD frame structure is shown in Fig. 2. In the standard, an adaptive split between DL and UL subframes are allowed. But it is usually fixed or remained unchanged for a long period in practical applications [20]. Hence, we take the fixed split case as an example for analysis. However, our algorithms can be extended to be used in FDD as well as adaptive split TDD operations.

3.1.1 Available Bandwidth Estimation

By the OFDMA technique, the bandwidth is allocated in the form of data bursts where an integer number of slots are included. The BS determines the number of DL and UL slots that a station obtains in one frame, and then broadcasts the resource allocation results through DL-MAP and UL-MAP messages at the beginning of each DL subframe. Therefore, the station can easily obtain the utilization of WiMAX link by aggregating the number of allocated slots stated in DL-MAP/UL-MAP messages.

Let AAS_d and AAS_u denote the number of allocated DL/UL slots in one frame, which are averaged from n frames. T_f, T_{df}, and T_{uf} denote the duration of a frame, a DL subframe, and an UL subframe, respectively. Then, the available bandwidth in DL and UL can be calculated by

$$\begin{cases} B_d = \left(1 - \dfrac{AAS_d}{s_d}\right) \dfrac{\delta_d s_d}{T_f} \\ B_u = \left(1 - \dfrac{AAS_u}{s_u}\right) \dfrac{\delta_u s_u}{T_f} \end{cases}, \quad (1)$$

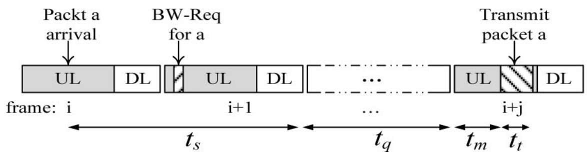

Fig. 3. An example of UL packet transmission.

where $s_d(s_u)$ denotes the total slots in a DL (UL) subframe and $\delta_d(\delta_u)$ is the number of bits transmitted in a DL (UL) slot which depends on the modulation and coding schemes deployed.

3.1.2 Packet Delay Estimation

We define the packet delay as the latency from the time that the packet arrives at the MAC layer buffer to the time that the packet is successfully transmitted. We divide the total delay into four components for analysis

$$t = t_s + t_q + t_m + t_t. \tag{2}$$

- The scheduling delay t_s is taken from the moment the packet arrives at the MAC layer buffer, to the moment this arrival information is obtained by BS.
- The queuing delay t_q is the time to be waited for the beginning of the frame allocated for transmitting the packet.
- The mapping delay t_m is taken from the beginning of the allocated frame to the first time slot appointed to the station.
- The transmission delay t_t is the time required to transmit the packet.

UL delay estimation. IEEE 802.16e provides a number of ways for stations to request bandwidth from BS. Here, we take a typical example for analysis as shown in Fig. 3, where each station can obtain an opportunity to send Bandwidth Request (BWR) message to BS for its newly arrived packets. In analysis, it is assumed that the packet arrival follows Poisson distribution, which is a valid assumption because it has given sufficiently accurate results for WiMAX networks [21].

- *Calculation of* t_s. The scheduling delay t_s in the UL consists of two parts. The first part is the delay incurred by the arrival till the start of the next frame. It is known that Poisson arrivals occur completely randomly in time, and then the first part of t_s can be approximated by $0.5T_f$. The second part is just the duration of the next frame, where BWR message is transmitted. Then, we write the scheduling delay t_s as

$$t_s = 0.5T_f + T_f = 1.5T_f. \tag{3}$$

- *Calculation of* t_q. After receiving BWR message, BS has known the arrival of the packet at the station. The packet can be imagined to enter a common queue at BS, which is shared by UL packets from all stations. If the next frame after sending BWR message is granted to this packet, the queuing delay t_q will be zero. Otherwise, it will be increased by frame duration till one frame is granted.

Here, we model this common queue by $\mathrm{D^X/D^m/1}$ queuing system [22], [23]. The service time of this queue is the frame duration T_f. The explicit formula for the equilibrium queue length of $\mathrm{D^X/D^m/1}$ system is

$$E(N_q) = \sum_{k=1}^{m-1} \frac{1}{1-z_k} + \frac{A''(1) - m(m-1)}{2(m-\mu_A)}, \tag{4}$$

where *A(z)* is the probability generating function of newly arrival batch size, μ_A is the mean arrival batch size, m is the fixed number of customers that can be served during one serving period. Since the packet arrival at a station follows Poisson distribution, the packet arrival at the common queue of BS also follows Poisson distribution based on the merging property. We use λ (packets/frame) to denote the mean packet arrival rate at the common queue. Then, we have $\mu_A = \lambda$, and *A(z)* can be expressed as

$$A(z) = e^{\lambda(z-1)}, \tag{5}$$

where $z_1, \ldots, z_{m-1}$ together with $z_0 = 1$ are the m roots of $z^m = A(z)$ in $|z| \le 1$. An explicit expression for these roots is given by

$$z_k = \sum_{l=1}^{\infty} c_l w_k^l, k = 0, \quad 1, \ldots, m-1, \tag{6}$$

where $w_k = e^{2\pi ki/m}$. For Poisson traffic, c_l can be expressed by the following expression:

$$c_l = e^{-l\rho} \frac{(l\rho)^{l-1}}{l!}, \tag{7}$$

where $\rho = \lambda/m$ is the utilization of the queue server with $0 \le \rho < 1$.

In our scheme, the number of packets that can be served in one UL subframe is

$$m = \frac{\delta_u s_u}{L}, \tag{8}$$

where L is the average packet length at the desired station. After the desired station accesses to the network, the radio resource will be utilized by it and other existing stations. Therefore, the utilization ρ will be expressed as

$$\rho = \frac{AAS_u}{s_u} + \frac{B}{\delta_u s_u / T_f}, \tag{9}$$

where B is the expected bandwidth of the desired station. Subsequently, we can get $\lambda = \rho m$. By combining (5) and (9), we can solve $E(N_q)$ for a given AAS_u. Finally, from the Little's theorem, we can obtain the queuing delay t_q

$$t_q = \frac{E(N_q)T_f}{\lambda}. \tag{10}$$

- *Calculation of* t_m *and* t_t. As shown in Fig. 2, all UL data bursts are allocated as horizontal stripes. If the packet length L is larger than σ which is the number of bits that can be transmitted in one UL subchannel, the transmission will definitely use the time slots at the head of the next allocated subchannel. In this case, the mapping delay t_m will be zero and the transmission delay t_t will be the total UL subframe duration T_{uf}. Then, the sum of t_m and t_t in this case is

$$t_m + t_t = 0 + T_{uf} = T_{uf}. \tag{11}$$

Conversely, if L is less than σ, the transmission will start from the end of the previous burst. Here, there may be two scenarios. In scenario S1, the residual slots of that subchannel are adequate and then the data transmission will only use that subchannel. Otherwise, the data transmission will be continued to the next subchannel in scenario S2. Let P_{s1} and P_{s2} denote the probability of scenario S1 and S2, respectively. We can calculate the mean of the sum of t_m and t_t by the following expression:

$$\begin{aligned} t_m + t_t &= E[t_m + t_t|S1] \cdot P_{s1} + E[t_m + t_t|S2] \cdot P_{s2} \\ &= \frac{1}{2}\left(\frac{L}{\sigma} + 1\right)T_{uf} \cdot \left(1 - \frac{L}{\sigma}\right) + T_{uf} \cdot \frac{L}{\sigma} \\ &= \left[\frac{1}{2} + \frac{L}{\sigma} - \frac{1}{2}\left(\frac{L}{\sigma}\right)^2\right]T_{uf}. \end{aligned} \tag{12}$$

Totally, we have

$$t_m + t_t = \begin{cases} T_{uf}, & L \geq \sigma \\ \left[\frac{1}{2} + \frac{L}{\sigma} - \frac{1}{2}\left(\frac{L}{\sigma}\right)^2\right]T_{uf}, & otherwise. \end{cases} \tag{13}$$

DL delay estimation. In the DL transmission, the BS has complete knowledge about the traffic. There is no need to send BWR messages, and then the scheduling delay t_s will be $0.5T_f$. The calculation of the DL queuing delay t_q is similar to that of the UL. It only needs to change the UL specific parameters including δ_u, s_u, and AAS_u to corresponding DL parameters δ_d, s_d, and AAS_d.

The mapping methods used in DL are quite different from those in UL. The standard requires all DL data bursts to be rectangular. According to the scanning order, current mapping algorithms can be divided into two types. One type is scanning and assigning slots from symbol to subchannel and the other is from subchannel to symbol [24]. Here, we take the second type as an example. The average mapping delay can be written as

$$t_m = T_p + \frac{1}{2}\rho(T_{df} - T_p), \tag{14}$$

where T_p denotes the time duration of the preamble, ρ is utilization of DL resource which can be estimated just as (9) in the UL case.

The transmission delay t_t in DL depends on the burst size and the number of bits transmitted by all subchannels in one slot duration, ϑ. It is supposed that one burst contains one packet. If the packet size L is larger than ϑ, the packet will occupy the subsequent time slots and pack to a rectangular based on its length. Otherwise, it will only occupy one slot duration. Then, the transmission delay can be approximated as

$$t_t = \begin{cases} T_s, & L \leq \vartheta \\ \left\lceil \frac{L}{\vartheta} \right\rceil T_s, & otherwise \end{cases}, \tag{15}$$

where T_s is the duration of one DL time slot.

3.2 Estimation in WLAN

WLANs are specified by IEEE 802.11 standards, where the fundamental access method is distributed coordination function (DCF) known as carrier sense multiple access with collision avoidance (CSMA/CA). Network allocation vector (NAV) is the main scheme used to avoid collision by setting a busy duration on hearing frame transmissions from other stations. Hence, the utilization of the WLAN channel is well reflected by NAV.

Our work evolves from the work in [3], where the available bandwidth has been derived from NAV as

$$BW = B_0 - L\frac{NAV}{T_n + \frac{1}{2}T_{n,c}(N-1)}, \tag{16}$$

where B_0 is the total system bandwidth, L is the mean frame size, T_n is the NAV duration for a successful frame transmission, $T_{n,c}$ is the NAV duration for a collision, and N is the average number of trials of a transmission. In [3], a mean access delay has also been derived as

$$\begin{aligned} t_a = &\frac{b(p_c T_{col} + (1 - p_c)T_s)}{2} + b\frac{v(CW_{\min})}{2} \\ &+ p_0 T_s \sum_{j=1}^{S-1} p_1^{j-1} + T_{col}\sum_{j=1}^{S-1} p_1^j + \sum_{j=1}^{S-1} p_1^j v(2^j CW_{\min}). \end{aligned} \tag{17}$$

All the variables in (16) and (17) can be estimated by NAV. However, this access delay is defined as the time used from the moment when a packet becomes the head of the queue at the MAC layer to the moment when this frame is successfully transmitted. It is not the total delay since the time spent in the queue has not been included.

Here, we further extend the model based on the above expressions. Each station is regarded as a queuing system. The total delay experienced by a packet consists of two parts: the queuing delay and the service time. We notice that the access delay obtained in (17) is just the service time in this model, and thus we can estimate the queuing delay from this access delay with a precondition that the distribution of the service time is known. Zhai et al. [25] has investigated the queuing characteristics of WLAN and concluded that the service time distribution can be well approximated by exponential distribution, and M/M/1 provides a good approximation for the queuing system especially when the station works in nonsaturated states, which are usually encountered in real implementations. Thus, we model this queuing system as M/M/1. With the mean packet arrival rate λ at the station, the queuing delay t_q can be obtained from the service time t_a by the following expression:

$$t_q = \frac{\lambda t_a^2}{1 - \lambda t_a}. \tag{18}$$

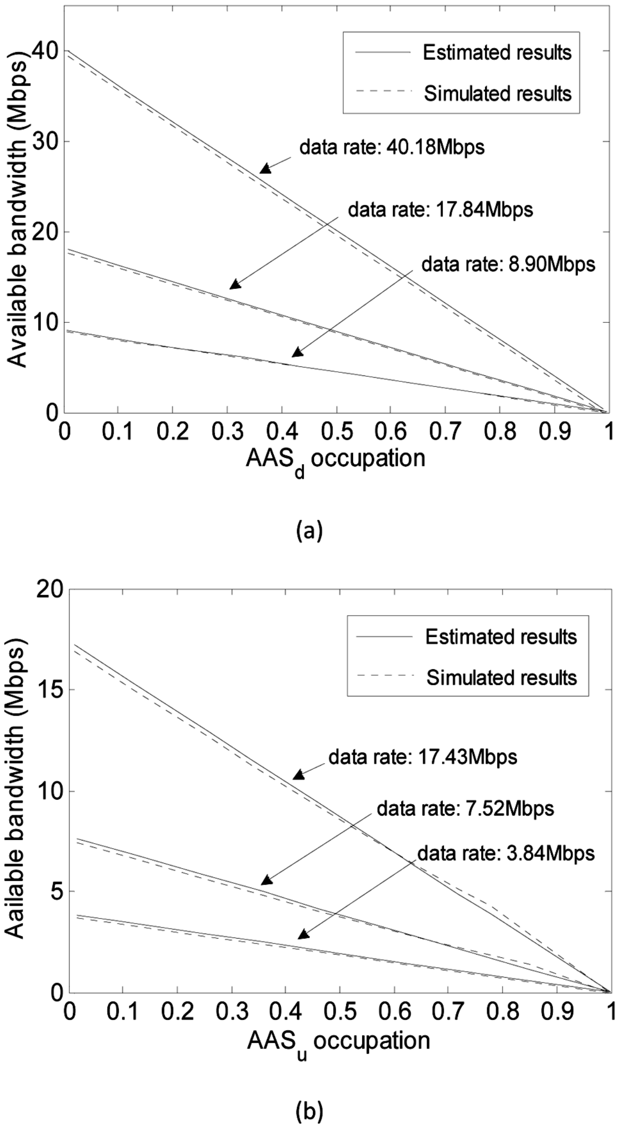

Fig. 4. Available bandwidth in WiMAX. (a) DL. (b) UL.

Finally, the total delay can be expressed as

$$t = t_q + t_a = \frac{\lambda t_a^2}{1 - \lambda t_a} + t_a. \tag{19}$$

4 Performance Evaluation

To demonstrate the effectiveness and feasibility of our proposed approach, we have built a simulation model by the simulation software QualNet 4.0. IEEE 802.11b was used as the WLAN standard. The channel bandwidth of WiMAX was 20 MHz. The DL to UL time radio in WiMAX was 2:1, with default frame duration of 5 ms.

4.1 Evaluation of Network Condition Estimation Methods

4.1.1 Available Bandwidth Estimation in WiMAX

We tested the DL and UL available bandwidth in different data rate cases, which were determined by the modulation schemes and given by $\delta_d s_d / T_f$ and $\delta_u s_u / T_f$. The results estimated by the proposed algorithms have been compared with the corresponding simulation results in Fig. 4, for DL and UL, respectively. The AAS_d (AAS_u) occupation denotes the utilization of the medium, which equals to AAS_d/s_d (AAS_u/s_u). It is shown in the figure that the simulated results and the estimated results match very well for both the DL and UL transmissions.

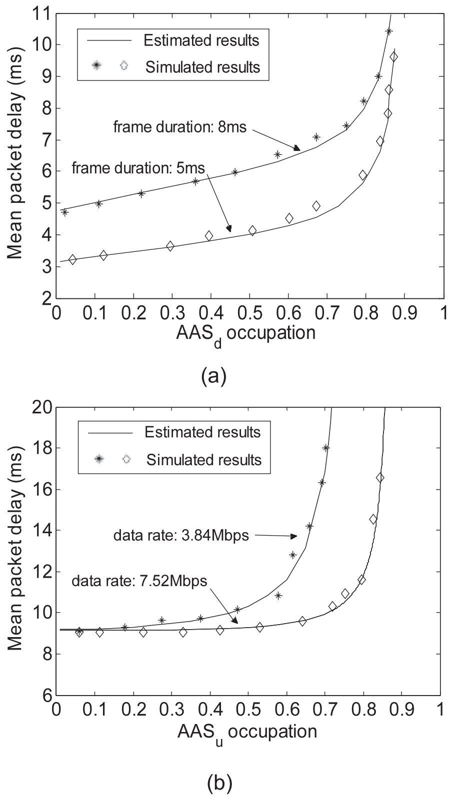

Fig. 5. Mean packet delay in WiMAX. (a) DL. (b) UL.

4.1.2 Packet Delay Estimation in WiMAX

From the analysis in Section 3.2, we can see that the estimated packet delay is determined by several parameters such as frame duration and the data rate. In the simulation, a DL or UL application ran on the desired station with a constant packet size of 500 bytes. The expected throughput was 800 kbps. For DL, the results were tested for different frame duration cases. T_s was 0.2 ms and ϑ was 2,880 bits. For UL, the results were tested for different data rate cases. σ was 360 bits or 720 bits for two cases.

The results estimated by the proposed algorithm are compared with corresponding simulated results in Fig. 5. It can be seen that the DL delay increases slowly with the increasing of AAS_d occupation. This is indicated in (14) that the mapping delay increases with the utilization. In addition, we can see that the DL delay is much lower than the UL delay of the same frame duration (5 ms). That is caused by the fact that the station needs to spend an extra frame to send BWR message at UL. Moreover, the packet transmission extends from time domain to the frequency domain in UL, while the case in DL is converse.

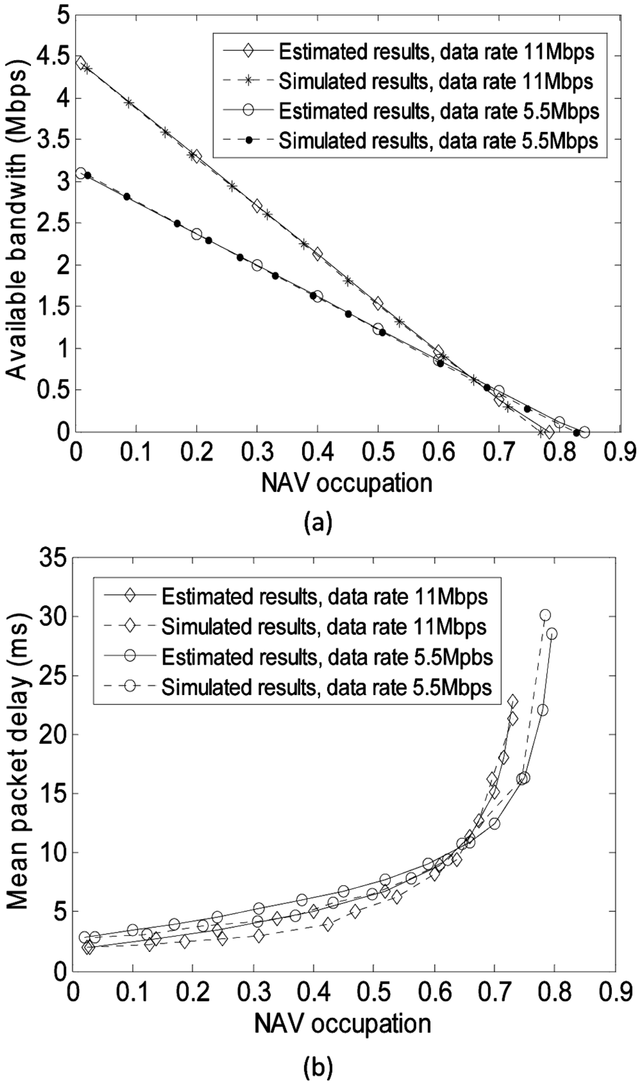

Fig. 6. WLAN estimation. (a) Available bandwidth. (b) Mean packet delay.

4.1.3 Estimation in WLAN

The available bandwidth and packet delay of WLAN network estimated by our algorithms are compared with corresponding simulation results in Fig. 6. The results are shown for different data rate type cases. The packet size of the desired station is 1,000 bytes and the packet arrival rate λ is 50 (packets/second). In the packet delay estimation, we found that the difference between estimated and simulated results was mainly caused by t_a when the network was nonsaturated, and then caused by both t_a and t_q when the network tended to be saturated.

4.2 Evaluation of Proposed VHO Scheme

On the proposed VHO scheme, we have conducted simulations in an interworking system which consists of one WiMAX network and two overlapped WLANs, as shown in Fig. 7. We have compared our proposed VHOM scheme with two other reference schemes. The first one is a QoS-oriented VHO scheme designed in [16] for WLAN-based overlay networks, where the QoS provided by a WLAN is considered in VHO decisions when a mobile station moves across it. The second one is a WLAN-preferred scheme that is widely deployed in cellular/WLAN systems. As long as the station enters the coverage of a WLAN, a VHO to the WLAN will be performed immediately. The performances of the schemes have been tested for both mobile and fixed stations.

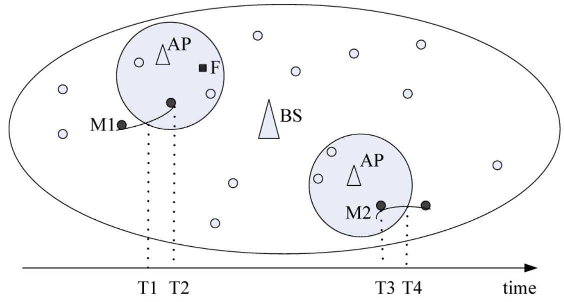

Fig. 7. Simulation topology of WiMAX/WLAN interworking system.

4.2.1 For Mobile Stations

We have performed simulations for mobile stations in two scenarios. In the first scenario, a mobile station M1 moved toward an overlapped region as shown in Fig. 7. There was an UL connection running at M1 which was a nonreal-time variable bit rate (VBR) application with an expected throughput of 800 kbps. Initially, the VHOM found that the average throughput of the application fell below the accepted 600 kbps. But there was no other network available at that time. Since both reference schemes trigger handoffs only by mobility, the performance of applications was not detected by them periodically. At the moment of T1, a WLAN network was found, and then a handoff was executed by the second reference scheme immediately. By our scheme, VHOM initiated the available bandwidth estimation process first. To guarantee the conditions of WLAN sufficiently good, $\mathrm{Thr}_{\_T}$ was set to be 900 kbps. The estimated available bandwidth could not satisfy this requirement until T2, and then a handoff was performed at T2 as shown in Fig. 8a. By the first reference scheme, the station did not execute handoff at T1 either for the bad performance of WLAN.

In the second scenario, a mobile station M2 was initially served by the WLAN with the same application at M1. At the moment of T3, VHOM found that the throughput fell below the accepted 600 kbps for three times, and then a handoff was executed since the estimated results showed WiMAX working in a good condition ($\mathrm{Thr}_{\_u\_T} = 900$ kbps). It is shown in Fig. 8b that a handoff was executed at T4 by the second reference scheme when the station moved out of the border of WLAN, while a little earlier by the first reference scheme because a higher exit signal-to-noise ratio (SNR) threshold was set for the bad conditions of WLAN. But this scheme makes effects only around the border of WLANs. When the station stay inside the WLAN (before 80 s), the handoff cannot be initiated to proactively improve the QoS of applications. The simulation results show that always the best service can be achieved by the proposed VHOM scheme.

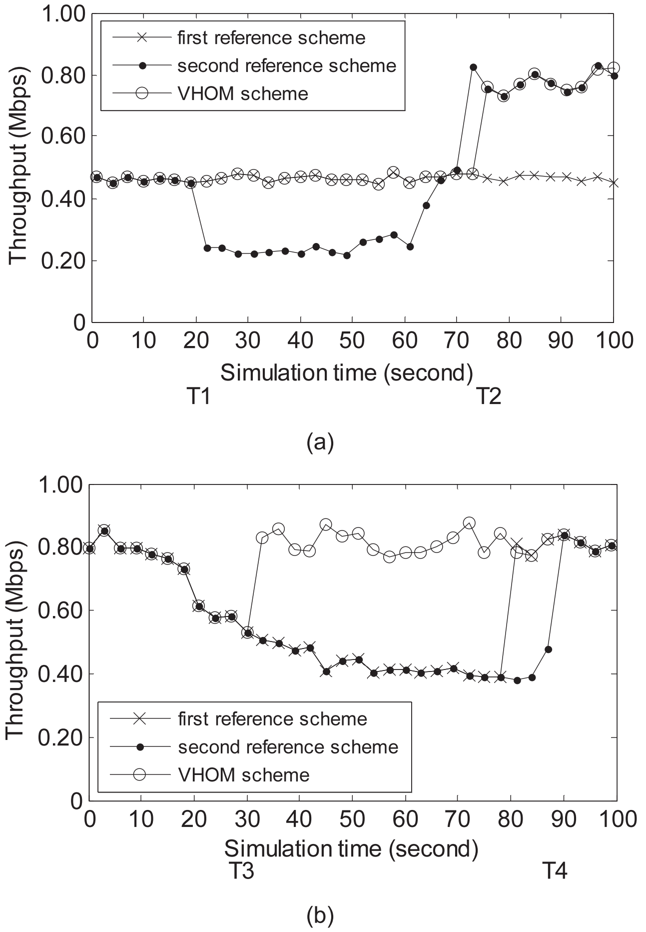

Fig. 8. Throughput comparison in the movement scenario.

4.2.2 For Fixed Stations

In the simulation, a station F located in an overlapped region without movements, which worked with a DL real-time constant bit rate (CBR) application. The expected throughput was 500 kbps with a threshold of 450 kbps. The accepted end-to-end packet delay is 20 ms. Initially, the station was served by the WLAN. At around 27 s, the average end-to-end delay detected by VHOM could not fulfill the requirement for three times although the throughput was fairly well. By the VHOM scheme, the condition of the WLAN was evaluated first. The result showed that the latency introduced by the WLAN operations exceeded the 8 ms, which is the threshold. Then, the VHOM began to estimate the status of the WiMAX network. At about 30 s, a decision of handoff was made based on the fact that the DL conditions of the WiMAX network could satisfy the requirements ($Pd_{d\_T} = 8$ ms). The throughput and end-to-end packet delay obtained by the schemes have been compared and shown in Fig. 9. Both reference schemes could not initiate handoffs for the fixed stations. The results show that a much lower packet delay and more stable throughput can be obtained by the proposed VHOM scheme.

5 Conclusion

In this paper, we investigate several important issues for the interworking of WiMAX and WLAN networks. We address a tightly coupled interworking architecture as the platform of our scheme. Based on the tightly coupled architecture, we propose a novel seamless and proactive VHOM scheme for stations to control the vertical handoff operations in the interworking networks, which aims to provide ABC service for both mobile users and fixed users. In order to make stations be able to proactively evaluate network conditions for making handoff decisions, we develop algorithms to estimate the available bandwidth and packet delay in WiMAX and WLAN, respectively. By the simulation experiments, we have proven the feasibility and effectiveness of our proposed schemes.

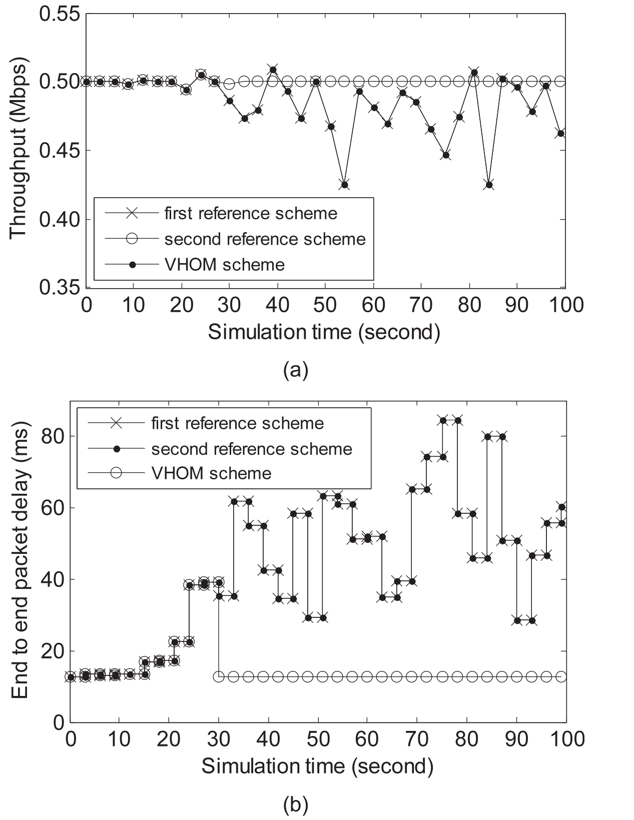

Fig. 9. Simulation results for the fixed scenario. (a) Throughput. (b) End-to-end packet delay.

References

[1] C.W. Lee, L.M. Chen, M.C. Chen, and Y.S. Sun, "A Framework of Handoffs in Wireless Overlay Networks Based on Mobile IPv6," *IEEE J. Selected Areas Comm.*, vol. 23, no. 11, pp. 2118-2128, Nov. 2005.

[2] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in Fourth Generation Heterogeneous Networks," *IEEE Comm. Magazine*, vol. 44, no. 10, pp. 96-103, Oct. 2006.

[3] C. Guo, Z. Guo, Q. Zhang, and W. zhu, "A Seamless and Proactive End-to-End Mobility Scheme for Roaming across Heterogeneous Wireless Networks," *IEEE J. Selected Areas Comm.*, vol. 22, no. 5, pp. 834-848, June 2004.

[4] G. Lampropoulos, N. Passas, L. Merakos, and A. Kaloxylos, "Handover Management Architectures in Integrated WLAN/Cellular Networks," *IEEE Comm. Surveys and Tutorials*, vol. 7, no. 4, pp. 30-44, Oct.-Dec. 2005.

[5] A.B. Pontes, D.D.P. Silva, J. Jailton, O. Rodrigues, and K.L. Dias, "Handover Management in Integrated WLAN and Mobile WIMAX Networks," *IEEE Wireless Comm.*, vol. 15, no. 5, pp. 86-95, Oct. 2008.

[6] W.S. Lim, D.W. Kim, Y.J. Suh, and J.J. Won, "Implementation and Performance Study of IEEE 802.21 in Integrated IEEE 802.11/802.16e Networks," *Computer Comm.*, vol. 32, no. 1, pp. 134-143, Jan. 2009.

[7] C. Youngkyu and C. Sunghyun, "Service Charge and Energy-Aware Vertical Handoff in Integrated IEEE 802.16e/802.11 Networks," *Proc. IEEE INFOCOM*, pp. 589-597, 2007.

[8] C.C. Yang, C.S. Tsai, J.Y. Hu, and T.C. Chuang, "On the Design of Mobility Management Scheme for 802.16e-Based Network Environment," *Computer Network*, vol. 51, no. 8, pp. 2049-2066, June 2007.

[9] W. Lee, E. Kim, J. Kim, I. Lee, and C. Lee, "Movement-Aware Vertical Handoff of WLAN and Mobile WiMAX for Seamless Ubiquitous Access," *IEEE Trans. Consumer Electronics*, vol. 53, no. 4, pp. 1268-1275, Nov. 2007.

[10] W. Yang, Y. Wang, Y. Tseng, and B.P. Lin, "Energy-Efficient Network Selection with Mobility Pattern Awareness in an Integrated WiMAX and WiFi Network," *Int'l J. Comm. System*, vol. 23, no. 2, pp. 213-230, Feb. 2010.

[11] J.J.J. Roy, V. Vaidehi, and S. Srikanth, "Always Best-Connected QoS Integration Model for the WLAN, WiMAX Heterogeneous Network," *Proc. First Int'l Conf. Industrial and Information Systems (ICIIS)*, pp. 361-366, Aug. 2006.

[12] S.F. Yang and J.S. Wu, "Handoff Management Schemes across Hybrid WiMAX and Wi-Fi Networks," *Proc. IEEE TENCON*, pp. 680-683, 2007.

[13] M.J. Kim, S.W. Son, and B.H. Rhee, "QoS Based Provisioning Vertical Handover between IEEE 802.11 and 802.16," *Proc. 11th Int'l Conf. Advanced Comm. Technology (ICACT)*, pp. 1415-1418, Feb. 2009.

[14] Z. Dai, R. Fracchia, J. Gosteau, P. Pellati, and G. Vivier, "Vertical Handover Criteria and Algorithm in IEEE 802.11 and 802.16 Hybrid Networks," *Proc. IEEE Int'l Conf. Comm. (ICC)*, pp. 2480-2484, May 2008.

[15] D. Ma and M. Ma, "A QoS-Based Vertical Handoff Scheme for Interworking of WLAN and WiMAX," *Proc. IEEE Global Telecomm. Conf. (GLOBECOM)*, pp. 2260-2265, Nov./Dec., 2009.

[16] A.V. Garmonov et al., "QoS-Oriented Intersystem Handover between IEEE 802.11b and Overlay Networks," *IEEE Trans. Vehicular Technology*, vol. 57, no. 2, pp. 1142-1154, Mar. 2008.

[17] D.C. Plummer, "An Ethernet Address Resolution Protocol (ARP)," IETF RFC 826, Nov. 1982.

[18] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy, "Bandwidth Estimation: Metrics, Measurement Techniques, and Tools," *IEEE Network*, vol. 17, no. 6, pp. 27-35, Nov./Dec. 2003.

[19] IEEE Std 802.16 - 2009 "IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Broadband Wireless Access Systems," May 2009.

[20] B. Kim, J. Yun, Y. Hur, C. So-In, R. Jain, and A.K. Al Tamimi, "Capacity Estimation and TCP Performance Enhancement over Mobile WiMAX Networks," *IEEE Comm. Magazine*, vol. 47, no. 6, pp. 132-141, June 2009.

[21] D. Chuck and J. Chang, "Bandwidth Recycling in IEEE 802.16 Networks," *IEEE Trans. Mobile Computing*, vol. 9, no. 10, pp. 1451-1464, Oct. 2010.

[22] Y.Q. Zhao and L.L. Campbell, "Equilibrium Probability Calculations for a Discrete-Time Bulk Queue Model," *Queueing Systems*, vol. 22, nos. 1/2, pp. 189-198, 1996.

[23] A. Janssen and J.S.H. Van Leeuwaarden, "Analytic Computation Schemes for the Discrete-Time Bulk Service Queue," *Queueing Systems*, vol. 50, nos. 2/3, pp. 141-163, July 2005.

[24] T. Ohseki, M. Morita, and T. Inoue, "Burst Construction and Packet Mapping Scheme for OFDMA Downlinks in IEEE 802.16 Systems," *Proc. IEEE Global Telecomm. Conf. (GLOBECOM)*, pp. 4307-4311, 2007.

[25] H. Zhai, Y. Kwon, and Y.G. Fang, "Performance Analysis of IEEE 802.11 MAC Protocols in Wireless LANs," *Wireless Comm. and Mobile Computing*, vol. 4, no. 8, pp. 917-931, Dec. 2004.

Dong Ma received the BE degree in communication engineering and the MSc degree in cryptography from Southwest Jiaotong University, China, in 2002 and 2006, respectively. From 2006 to 2007, she was working in Beijing Xinwei Telecom Technology Inc. as an information system engineer. Currently, she is working toward the PhD degree in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Her research interests include interworking of heterogeneous networks, mobility management, quality of service and video transmission in wireless networks. She is a student member of the IEEE.

Maode Ma received the BE degree from Tsinghua University in 1982, the ME degree from Tianjin University in 1991, and the PhD degree in computer science from Hong Kong University of Science and Technology in 1999. Now, he is a tenured associate professor in the School of Electrical and Electronic Engineering at Nanyang Technological University in Singapore. He has extensive research interests including wireless networking and wireless network security. He has led and/or participated in around 20 research projects funded by government, industry, military, and universities in various countries. He has been a member of the technical program committees for more than 100 international conferences. He has been a general chair, technical symposium chair, tutorial chair, publication chair, publicity chair, and session chair for more than 50 international conferences. He has more than 180 international academic publications including about 70 journal papers, more than 110 conference papers and/or book chapters, and 3 academic books. He is the editor-in-chief of *International journal of Electronic Transport*. Currently, he serves as an associate editor for *IEEE Communications Letters*, a senior editor for *IEEE Communications Surveys and Tutorials*, and an associate editor for International *Journal of Network and Computer Applications*, International Journal of *Security and Communication Networks*, and International Journal of *Wireless Communications and Mobile Computing*. He is a senior member of *IEEE Communication Society* and a member of a few technical committees in the *IEEE Communication Society*. He is a senior member of the IEEE.