A Cooperative Two-Step Vertical Handoff Scheme with Mobility Prediction

Shih Yu Chang[1], and Pin-Han Ho[2]
[1]Department of Applied Data Science, San Jose State University, San Jose, U.S.A.
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada

As mobile communication evolves into 3G beyond, the interworking of multiple heterogeneous networks serves as the major effort for taking the best advantage of different technologies available in supporting various emerging services, such as VoIP, Video on Demand (VoD), and IP Television (IPTV), etc. Vertical handoff is one of the key mechanisms in achieving Always Best Connected (ABC) for the mobile users by leveraging the benefits of deploying different types of networks for provisioning seamless handoff/roaming services in presence of user mobility. This paper aims to introduce a novel cooperative two-step vertical handoff scheme for the integration of 3G Wireless Wide-Area Networks (WWAN) and the IEEE 802.11 Wireless Local-Area Networks (WLANs). The proposed scheme is based on the cooperation based access point (AP) and mobile station (MS), where the AP manipulates the sensed signal strength to determine whether a pre-handoff action should be initiated. To improve the accuracy of user mobility prediction, a Markov model that incorporates with a novel parameter training process is developed at the AP for acquiring the hotspot geographic arrangement, such as the location of aisles, walls, and entrances/exits, etc., which is considered as the major factor of determining the user mobility patterns in an indoor hotspot. We will justify feasibility and discuss the operation complexity of the proposed cooperative vertical handoff. Moreover, error propagation due to inaccurate signal strength measurement is studied through Maximum Likelihood estimation. Finally, we will clearly demonstrate the merits gained by using the proposed two-step vertical handoff mechanism through extensive simulation, where the derived analytical models are verified.

***Index Terms*—WWAN, WLAN, pre-handoff, vertical handoff, Video on Demand (VoD), and IP Television (IPTV)**

I. Introduction

In the 3G beyond and 4G wireless communication systems [1], [2], [3], integration of heterogeneous network domains is a technological trend leading to a ubiquitous and pervasive communication environment. A Mobile Station (MS) is expected to support multiple air interfaces and Media Access Control (MAC) protocols in order to be Always Best Connected (ABC) for the Internet access through one of the available interfaces according to a cross-layer decision making process. To achieve this goal, the 802.21 working group has addressed extensive efforts on the standardization process for 802 media access-independent mechanisms. The first draft on Media-Independent Handoff Function (MIHF) and Media-Independent Information System (MIIS) has emerged in March 2006, and the third draft has quickly came over in December 2006 [4], which has successfully defined the service models along with a unified signaling/software framework for achieving seamless interworking and information exchange.

The service model defined in 802.21 for the interworking of 802.11 and WWAN is illustrated in Fig. 1, where a shim layer performing MIHF between the IP (layer 3) and link layer (layer 2) is formed. With the media-independent handoff service access point (MIH-SAP), the MIHF is equipped with a service access point (SAP) to the higher layer entities, such as transport, handoff policy function, and layer-3 mobility management. Meanwhile, LLC-SAP and 3GLink-SAP are two media-dependent SAPS that allow the MIHF to use the services from the lower layers of the mobility management protocol stack and their management planes. Since the link layer parameters and software artifacts could be very vendor-specific, the adoption of the MIHF serving as the 2.5 layer can greatly improve the system interoperability and facilitate the emergence of new services and applications. With the unified network information exchange mechanisms, the vertical handoff schemes can be designed with better intelligence.

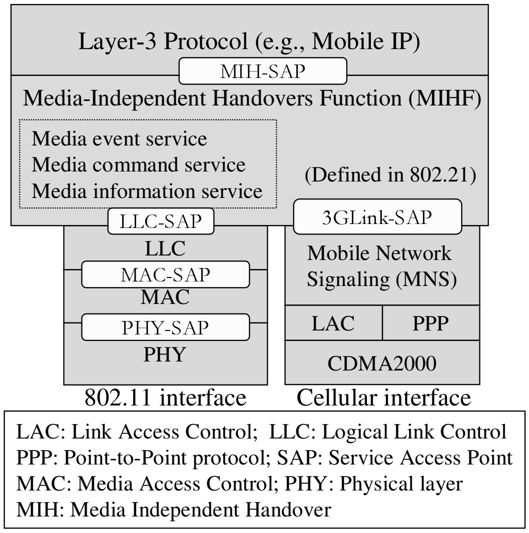
Layer-3 Protocol (e.g., Mobile IP)
MIH-SAP
Media-Independent Handovers Function (MIHF)
Media event service
Media command service
Media information service
(Defined in 802.21)
LLC-SAP
3GLink-SAP
LLC
MAC-SAP
MAC
PHY-SAP
PHY
Mobile Network Signaling (MNS)
LAC
PPP
CDMA2000
802.11 interface
Cellular interface
LAC: Link Access Control; LLC: Logical Link Control
PPP: Point-to-Point protocol; SAP: Service Access Point
MAC: Media Access Control; PHY: Physical layer
MIH: Media Independent Handover

Fig. 1. Layers infrastructure of 802.21.

IEEE 802.21 helps with handoff initiation, network selection and interface activation during vertical handoffs. Most important of all, IEEE 802.21 enables co-operative handoff decision making between the clients and the network. With such a standard platform, the network operation and management

Manuscript received Aug. 8, 2021; revised Nov. 21, 2021. Corresponding author: Shih Yu Chang (email: shihyu.chang@sjsu.edu).

can optimize handoffs between heterogeneous 802 systems (such as 802.3, 802.11, and 802.16) and between 802 systems and cellular systems (such as 3GPP with CDMA 2000) with intelligence and inter-operability, which is considered to be one of the most challenging tasks in the interworking of heterogeneous wireless communication networks. The major challenge is on how to take advantage of the advertised network state information and determine whether the MS should switch to another available air interface for better supporting those highly interactive real-time services under user mobility. This process is also referred to as vertical handoff, which plays an important role in the effort of integrating heterogeneous networks for achieving ABC. Seamless and swift handoffs become a particularly important design requirement especially in the presence of extensive emerging real-time and multimedia services such as VoIP, video phone, and Mobile TV, etc.

IEEE 802.11 based wireless Local Area Networks (WLANs) and 3GPP based cellular systems (WWAN) are two popular service-oriented wireless communication technologies that have been widely deployed in the world. The integration of WLANs and WWANs is envisioned to be the first step in the evolution to the next-generation all-IP wireless Internet [5], [6], [7]. To realize such an application scenario, mobile communication device manufacturers have to make their mobile stations (MSs) (such as mobile phone and PDA) with compatible with both digital cellular systems and WLANs. With this, a MS can acquire up to 54 Mbps high-speed data streams in a WLAN hotspot with a much cheaper fare, while maintaining a 56 to 115 kbps data rate when roaming between disjoint hotspot areas covered by a cellular system. A vertical handoff action is required when the MS moves across the border of a WLAN and a WWAN. In particular, when the MS is leaving the WLAN, the timing for setting up a new link to the WWAN is important and should be subject to careful design such that the system can take the best advantage of the interworking.

To implement vertical handoff, Mobile IP (MIP) is the most commonly adopted protocol on top of the 802.21 shim layer, which manages user mobility by introducing two network elements: Home Agent (HA) and Foreign Agent (FA). With MIP, when a MS is in the communication range of the home agent, the MS accesses the Internet through regular IP protocol. When the MS moves and switches to the communication range of a FA, it is authenticated by the FA and registers the FA's address to its HA as a Care of Address (CoA). The data packets destined to the MS are then encapsulated with a new IP header from the FA and delivered through a tunnel between the HA and the MS. In order to solve the inherent triangular routing problem [8], [9] incurred in the conventional MIP, optimized MIP [7] sets the starting point of the tunnel as the Correspondent Station (CS) instead of the HA.

Some unique features exist in the design of a vertical handoff scheme in the interworking of a cellular system and 802.11 based WLANs, and are summarized in the following.

1) Since a WWAN usually covers the whole geographic area while each WLAN has a small coverage which is entirely overlapped with that of the WWAN, the upward (from WLAN to WWAN) and downward (from WWAN to WLAN) handoffs must be considered differently. The downward handoff is subject to less problems since the old link to the WWAN can be cut when the new link to the WLAN is completely set up for achieving zero handoff delay. On the other hand, in the upward handoff, the SINR of the old link to the WLAN could become unqualified due to user mobility, where a non-zero handoff delay is introduced in case the new link to the WWAN is not ready in time.
2) The received signal strength (RSS) in a WLAN and a WWAN is not comparable for making a handoff decision due to totally different MAC protocols and air interfaces adopted in the two networks.
3) Since WLANs have low cost and high rate, the developed handoff scheme is preferred to have the MS to consume the resources in the WLAN as long as possible.

In order to reduce the vertical handoff delay and packet loss, most previously proposed schemes took either one of the following two approaches: (1) Implementing different steps of vertical handoff in parallel or implementing some steps of handoff in advance when a specific condition is met. The representative scheme based on this approach is the one with the IETF pre-registration mechanism [10], [11]. (2) Decreasing signaling latency by re-assigning a new network component closer to the MS to serve the HA function. By using this approach, HAWAII [12] and Cellular IP [13] were proposed to reduce the vertical handoff delay.

This paper focuses on the a new implementation of a two-step vertical handoff process in the interworking of WWAN and WLAN. Based on the framework and service supports defined in IEEE 802.21, we introduce a novel cooperative handoff decision making process, where the access point (AP) of the WLAN is equipped with a suite of interoperable mobility prediction mechanisms for each accessing MS. We first define the two-step handoff process of interest in the study, which is comprised of the *pre-handoff* step and *handoff* step. Different from the study in [11], with the proposed pre-handoff step the MS not only completes the whole registration process such as signaling, authentication, bridging, and database updates, etc., but also can receive data from the new link. Furthermore, instead of making all the decisions at a MS (which is the basic assumption taken by most of the previously reported studies), it is the AP to notify an accessing MS to start pre-handoff. Motivated by the fact that the indoor user mobility is mainly dominated by the geographic arrangement of the hotspot, such as aisles, walls, and entrances/exits, etc., we propose a Markov chain based parameter training mechanism at the AP to gain the knowledge of user mobility pattern in the hotspot, in order to improve the accuracy in user mobility prediction. Since the obtained parameters through the training process is specific to the geographic arrangement of the hotspot, we expect a much better accuracy in achieving the desired performance in the vertical handoff process can be yielded. In addition, since the pre-handoff decision is made at the AP, the design complexity and power consumption of the MSs can be significantly reduced.

We will justify that the proposed mechanism can be easily performed under the service supports of the 802.21 framework.

As one of the most distinguished features from the previously reported counterparts, the performance of the proposed scheme is independent of which user mobility model is assumed since the AP derives the user mobility pattern through the historical user mobility record and the proposed parameter training mechanism. Extensive analysis is conducted for modeling the system performance, including the error propagation from the signal detection inaccuracy to the induced error in the measured handoff delay. Although there are many existing works about handoff [14], [15], [16], the proposed method outperforms these existing ones in vertical handoff delay since we take pre-handoff decision to save time.

The rest of this paper is organized as follows. In Section II, the proposed two-steps vertical handoff scheme is presented. In Section III, the proposed parameter training process is introduced. The performance analysis is given in Section. IV, which involves in a training process in determining the transition probability of each pair of position states. The effects of estimation error upon the measured handoff delay is discussed in Section V. In Section VI, simulation is conducted to verify the proposed analytical model and demonstrate the effectiveness of the training model. Section VII concludes this paper.

II. Proposed Two-Steps Vertical Handoff Scheme

As a review, this section first presents the standard handoff scheme that has been widely adopted in the current implementation, followed by the introduction of the proposed two-step handoff mechanism. We will focus on the case where the MS intends to hand over from WLAN to WWAN.

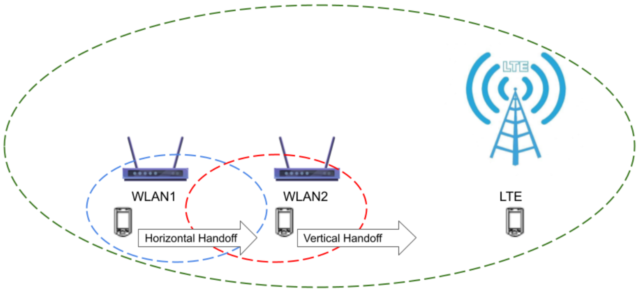

Fig. 2. Horizontal handoff and vertical handoff.

Handoff can be classfied into either vertical or harizontal handoff based on the different number of network interfaces involved during the handoff process, as shown by Fig. 2. A horizontal handoff happes between network access points that utilizes the same wireless protocol. For example, when a mobile device migrates between WLAN network domains, the handoff event would be considered as horizontal since the connection is reconnected by the change of WLAN domain but not of the wireless protocol. A vertical handoff will use two different network interfaces, which are used by different wireless protocols. For example, when a mobile device travels out of an WLAN network and move into a LTE network, such handoff is considered as vertical handoff. In this work, we will focus on vertical handoff.

The standard handoff process defined in mobile IP (MIP) includes the following procedures (also shown in Fig. 3): (a) the MS detects the signal of a new network domain, (b) the signal condition of the network domain meets the handoff requirement such that the MS decides to initiate a handoff, (c) the MS sends a handoff request message to HA (steps 1-8 of Fig. 3) and disconnects from FA$_{\text{old}}$ (d) The HA requests the AAA server (step 9) to authenticate the handoff request; (e) The HA receives the response from the AAA [1] server (step 10), (f) The HA removes the current connection to the FA$_{\text{old}}$ and delivers the data to FA$_{\text{new}}$ (steps 11 and 12).

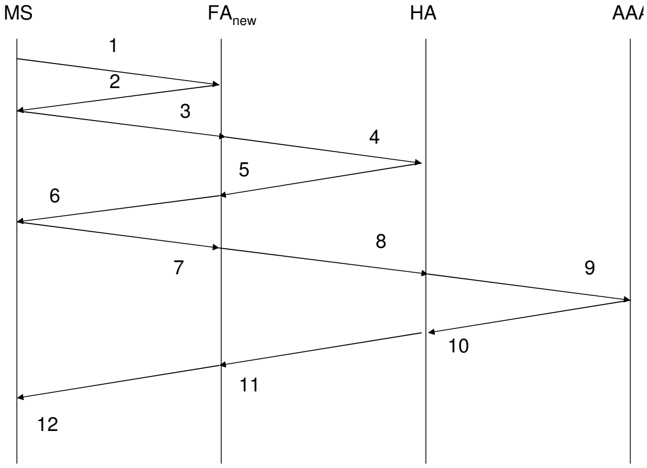

Fig. 3. Signaling for a vertical handoff procedure.

The timing diagram for a conventional vertical handoff process without the pre-handoff mechanism is shown in part (A) of Fig. 4. t_0 denotes the time instant at which the MS can sense the signal from the FA$_{\text{new}}$, and t_{ho} denotes the time instant when the handoff condition is met. At this moment, the MS is disconnected from the FA$_{\text{old}}$, and starts to establish a new link to HA. Let t_{fset} be the time instant that the MS starts to receive data through the FA$_{\text{new}}$. It can be observed that handoff delay can be easily incurred when t_{fset} is latter than t_{ho}.

The proposed two-step vertical handoff scheme is characterized by having a pre-handoff action, in which the AP notifies an accessing MS to start establishing a new link to FA$_{\text{new}}$ when the AP finds that the *pre-handoff condition* of the MS is met. Therefore, the link to the new network domain could be hopefully established before the MS cuts the old link (or called *handoff action*) in case a proper timing for setting up the new link is taken. Note that the MS will experience packet loss and handoff delay in the event that the SINR to the FA$_{\text{old}}$ becomes unqualified before the link to FA$_{\text{new}}$ is completely set up.

The timing diagram for the proposed two-step vertical handoff scheme is illustrated in Fig. 4(B). In the first step, the MS is notified by the AP for pre-handoff. Let t_{ipre} denote

[1] AAA is denoted for authentication, authorization, and accounting

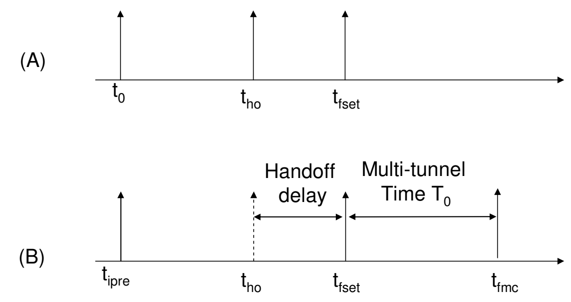

Fig. 4. Two vertical handoff scheme. (A) The conventional scheme; (B) Two steps vertical handoff scheme. The vertical handoff delay is introduced if $t_{fset} > t_{ho}$. After the timer of multi-tunnel counting down to zero, there are two possibilities. If $t_{fmc} > t_{ho}$, the multi-tunnel redundant time is incurred. If $t_{ho} > f_{fmc}$, the event of undo happens.

the time instant at which the pre-handoff condition is met, and t_{fset} denote the time instant that the new link with the FA$_{\text{new}}$ is successfully established. The time duration for setting up the new link in the first step, denoted as T_{set}, is defined as $T_{set} = t_{fset} - t_{ipre}$. In the second step, the MS activates a timer (denoted as T_{timer}) with an initiate value T_0, where both the new and old links are kept active until the timer is counted down to zero. This is referred to as the *multi-tunnel stage*.

At the end of the multi-tunnel stage, the MS either enters the new network domain or returns back to the old network domain, which corresponds to a handoff event and an *undo event*. In the former, the old link is cut by the MS since the link is subject to unqualified SINR when the timer is counted down to zero. Thus, the MS is handed over to the WWAN. In the latter, on the other hand, since the SINR of the FA$_{\text{old}}$ is still qualified at the end of the timer, the MS ends up the multi-tunnel stage simply by cutting the new link and then updating its status in the AP through the 802.21 remote event service.

Let the instant of ending the multi-tunnel period be denoted as t_{fmc}. According to Fig. 4(A), the handoff delay with the conventional MIP break-after-make scheme is $t_{fset} - t_{ho}$, while with the two-step vertical handoff, the handoff delay is determined by how early the MS initiates the pre-handoff procedure. The timing diagram for the proposed pre-handoff scheme is shown in Fig. 4(B). In case $t_{fset} > t_{ho}$), non-zero handoff delay $\tau_{delay} = t_{fset} - t_{ho}$ is introduced since the MS cannot establish the new link successfully before the SINR with the FA$_{\text{old}}$ becomes unqualified. The value of τ_{delay} is a random variable determined by the mobility behavior of the MS and how the pre-handoff condition is defined. In case of $t_{fset} < t_{ho} < t_{fmc}$, the MS can receive the data from the newly established link before the old link is unqualified, which leads to zero vertical handoff delay. The time duration of $T_0 - (t_{ho} - t_{fset})$ is also referred to as the *multi-tunnel time*. Finally, in case $t_{ho} > t_{fmc}$, the event of undo happens since the MS is still in the old network domain after the MS initiating its pre-handoff procedure for time duration $T_{set}+T_0$. In this case, the new link has to be cut, and the MS ends the multi-tunnel stage.

It is clear that both handoff delay and multi-tunnel time should be as small as possible since the former stands for the QoS impairment due to the handoff event, while the latter reflects the additional resource consumed in the handoff event and how easily/frequently each MS starts to initiate a new link to the WWAN. Furthermore, when the handoff delay is reduced by initiating the pre-handoff action earlier, the multi-tunnel time is increased, which results in a much higher frequency of undo events. Note that the high frequency of undo events may significantly increase the operational overhead and reduce the system stability due to frequent link setup/disconnection, authentication, and reconfiguration of network entities, etc. Thus, one of the major targets of this study is to achieve a graceful tradeoff between the two performance metrics by properly tuning the related parameters in determining whether or not a MS should initiate pre-handoff. For this purpose, a user mobility training process based on a Markov chain model for user position state prediction is developed (which will be presented in the next section) and implemented at the AP according to the historical user mobility patterns. Based on the obtained parameters, a whole picture on the user mobility in the hotspot can be sketched, by which a high accuracy in the user mobility prediction can be achieved at the AP in the subsequent pre-handoff decision making process since the user mobility pattern could be dominated by the geographic arrangement of the hotspot instead of user customs. With the obtained user mobility prediction model, the IP layer of the AP can be equipped with the intelligence of estimating the time instant of losing the WLAN signal by analyzing the RSS, by which the MS is notified by the AP to make a pre-handoff decision.

As the pre-handoff procedure starts, a request for setting up a link to the FA$_{\text{new}}$ will be initiated, which is sent to the HA via the FA$_{\text{old}}$. When the HA receives the pre-handoff request message, it sets up an alternative channel by sending the data through both FAs in the WLAN and the WWAN simultaneously. In the handoff step, the MS simply disconnects the link to the FA$_{\text{old}}$. Note that all the above signaling mechanisms can be easily performed through the media information and command services provided in the 802.21 standard.

The following is a list of parameters in the decision process of the proposed scheme.

1) T_{set} is the time duration for the MS to complete the link setup to the FA$_{\text{new}}$.
2) T_0 is the time duration for the MS to have both links to FA$_{\text{old}}$ and FA$_{\text{new}}$, which is referred to as the multi-channel connection.
3) T_{th} is the threshold time duration.
4) t_{ipre} is the time instant that the MS starts to trigger the setup of new link to the FA$_{\text{new}}$.
5) t_{fset} is the time instant that the new link to the FA$_{\text{new}}$ is ready.
6) t_{ho} is the time instant that the link to the FA$_{\text{new}}$ is cut, where the handoff procedure completes.
7) t_{fmc} is the time instant that the double links condition is complete.
8) $\widehat{Po}(t)$ is the estimated position of the MS at the time instant t.

9) s_i represents the position state i. A detailed illustration is in Fig. 5, where each numbered sub-area represents a position state. Because the handoff procedure will be implemented by those MSs with weaker RSS, we only need to consider those region which are the neighborhood of the WLAN/WWAN boundary.
10) β is the threshold on the confidence for a successful handoff action when considering whether a pre-handoff request should be initiated or not.

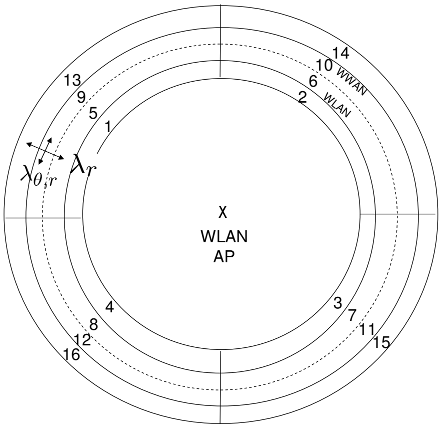

Fig. 5. The position states of the WLAN and WWAN region.

In this study, the time instant to issue a pre-handoff request, t_{ipre}, is the time instant at which the AP finds the following two conditions satisfied:

$$\begin{cases} \widehat{SNR}(t_{ipre}) \leq \alpha, \\ \sum\limits_{s_j \in \mathbf{A}} Pr\Big(\widehat{Po}(t_{ipre}+T_{th}) = s_j \Big| \widehat{Po}(t_{ipre}) = s_i\Big) \geq \beta, \end{cases} \tag{1}$$

where $\widehat{SNR}(t_{ipre})$ is the estimated SNR at the time instant t_{ipre} and α is some SNR threshold value [2]. The symbol $\mathbf{A}$ is the set of position states s_i at the boundary of a network domain, which are also referred to as the *absorbing states* with respect to the ones in the area of the original network domain. Both quantities β and T_{th} are design parameters that can be manipulated to meet a given performance requirement in terms of handoff delay and multi-tunnel time. The number of slots corresponding to the duration of T_{th} is N_{th}. The *absorbing probability* from the position state s_i to the position state s_j at n-th step is denoted as $p_{i,j}^{abs}(n)$. This probability can be evaluated according to the Eq. (38) derived in the Appendix.

In Fig. 6, the flowcharts for the proposed two-step vertical handoff scheme at the MS and the AP sides are illustrated, respectively, where the WLAN is the FA$_{\text{old}}$. In this case, the AP continues tracing each MS with a SINR lower than a threshold denoted as $alpha$ (or the MSs close to the boundary of the WLAN coverage) and analyzing the corresponding RSS in every small time interval (e.g., 500 ms). As shown in Fig. 6(a), the MS first checks if the SINR is qualified. If not, an immediate handoff is required, where the MS tries to set up a link to the WWAN. If the signal from the AP is all right, it checks if the MS is in the multi-tunnel stage. If not, it checks if the AP has notified it for pre-handoff. If the MS has been notified, the MS will enter the multi-tunnel stage after the new link is set (i.e., T_{set}). Once the MS enters the multi-tunnel stage, the MS has to activate T_{timer} with an initial value of T_0. When counting down to zero, the MS checks if the SINR from the WLAN is qualified. If not, it switches to the WWAN by cutting the old link, where a successful handoff action is done; otherwise, it goes back to the WLAN by cutting the new link, where an undo action is performed. In both cases, the AP will be notified with the updated status of the MS.

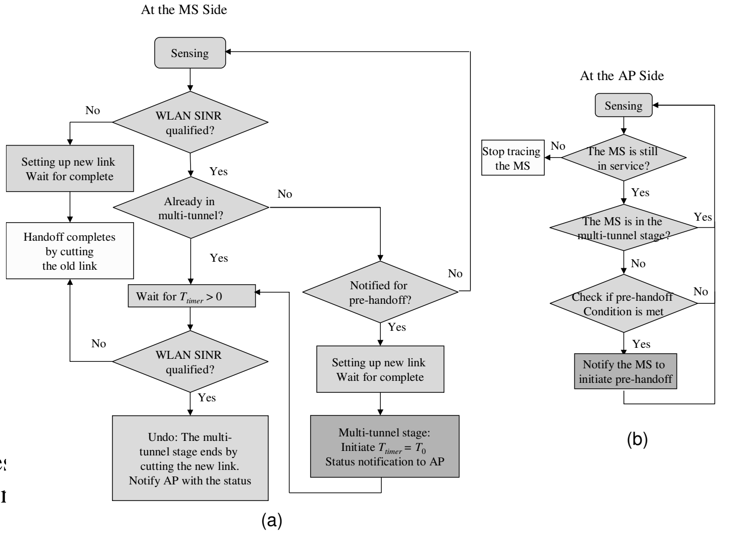
At the MS Side
Sensing
WLAN SINR qualified?
No
Yes
Setting up new link Wait for complete
Handoff completes by cutting the old link
Already in multi-tunnel?
No
Yes
Wait for $T_{timer} > 0$
WLAN SINR qualified?
No
Yes
Undo: The multi-tunnel stage ends by cutting the new link. Notify AP with the status
Notified for pre-handoff?
No
Yes
Setting up new link Wait for complete
Multi-tunnel stage: Initiate $T_{timer} = T_0$ Status notification to AP
(a)
At the AP Side
Sensing
Stop tracing the MS
No
The MS is still in service?
Yes
The MS is in the multi-tunnel stage?
Yes
No
Check if pre-handoff Condition is met
No
Yes
Notify the MS to initiate pre-handoff
(b)

Fig. 6. The flowcharts for cooperatively performing vertical handoff between AP and MS: (a) at the MS side; (b) at the AP side.

At the AP, it periodically check Eq. (1) for all the MSs which are not in a multi-tunnel stage to see if any MS meets the pre-handoff condition. If yes, the AP will notify the MS to perform pre-handoff. After the MS completes setting up the new link, the AP will be notified and stop tracing the MS except that the MS performs an undo action. It is clear that the frequency of undo events increases when the pre-handoff condition is getting looser and the multi-tunnel time becomes longer, which can be determined by manipulating the parameters of β and T_{th}.

As an off-line effort, the AP has to perform the proposed parameter training mechanism by recording and analyzing the historical RSS of MSs that have been serviced through the AP (which will be introduced in the next section). With the results of the parameter training, the AP is aware of the position state transition probabilities, which equivalently form the whole picture of user mobility patterns in the hotspot. With the knowledge of user mobility patterns, the AP can trace the RSS of each MS close to the boundary of the WLAN coverage and help the MS to make a proper pre-handoff decision. Note that in a hotspot such as a coffee shop, train station, and hotel, the user mobility patterns could be strongly restricted by the indoor geographic arrangement in the hotspot. For example, the geographic setup in a hotspot such as pathways, aisles, and entrances/exits of an office floor can be taken as fixed by which

[2] The pre-handoff algorithm will be initiated if the received signal strength from the AP is below than some SNR value.

the MSs' movement is constrained. In such a circumstance, the knowledge of user mobility patterns is expected to serve as a power tool for estimating the timing of pre-handoff for each MS.

In the on-line implementation supported by IEEE 802.21 standard, the AP can automatically identify the peer MSs at the beginning through the MIHF information services. The MIHF remote command services can help to notify the accessing MSs to start setting up a link to the WWAN. With the MIHF event services, the MSs subject to undo can notify the AP to revoke its multi-tunnel status such that the AP will restart tracing the MS. Due to the small coverage of an AP, the number of accessing MSs could be quite limited, where scalability should not be an issue.

In summary, the proposed two-step vertical handoff scheme imposes intelligence on the AP instead of the accessing MSs, which is one of the features that distinguish it from all the previously reported counterparts. Such a design can gain advantages in terms of power saving and design complexity reduction of the MSs, which is more than essential in case the MSs are going to be positioned as low-priced and battery-powered devices. With the help of the IEEE 802.21 standard, the signaling, network state exploration, and information exchange, can be performed with great interoperability and vendor-independency.

III. Mobility and Parameter Training Model

In this section, the mobility prediction and parameter training model is presented, where a discrete Markov chain is developed to measure and estimate the MS position states. With our approach, a transition matrix can be obtained through a parameter training process such that the user movement can be estimated by a series of matrix operation. By taking the advantage of the fixed geographic setup in a hotspot, the proposed method is expected to yield high accuracy in presence of user mobility diversity.

Fig. 5 shows the whole region in which the MS is observed, where the AP of the WLAN is at the center, and the largest circle with the dashed-dotted line represents the boundary within which the MS can sense the signal from the AP. The whole region is then partitioned into small sections by two geometric objects: one is a group of circles with a common center at the AP, and the other is a group of diameters which equally divide the total radian 2π. Each section under this partition represents a position state. The region outside of the dash circle is the WWAN region while the one inside the dash circle is called the WLAN region. The absorbing states of the WLAN region (with respect to the WWAN) are those position states at the boundary of the WLAN region, indexed with 9, 10, 11, 12 in Fig. 5. Similarly, the absorbing states of the WWAN region (with respect to the WLAN) are those states at the boundary of the WWAN region, which are indexed with 5, 6, 7, 8. The value of λ_r represents the transition probability between any pair of position states in the radiant direction when the index of the circle is r. The value of $\lambda_{\theta,r}$ represents the transition probability between any pair of position states in the tangent direction of a circle when the index of the circle is r.

To determine λ_r, a state space of $(2m+1)$ states to represent all the possible movements in the radiant direction per time slot is constructed, i.e., $(-m \times d, -(m-1) \times d, \cdots, -d, 0, d, \cdots, (m-1) \times d, m \times d)$, where d (in units of meter) denotes the distance between two adjacent states, as shown in part (A) of Fig. 7.

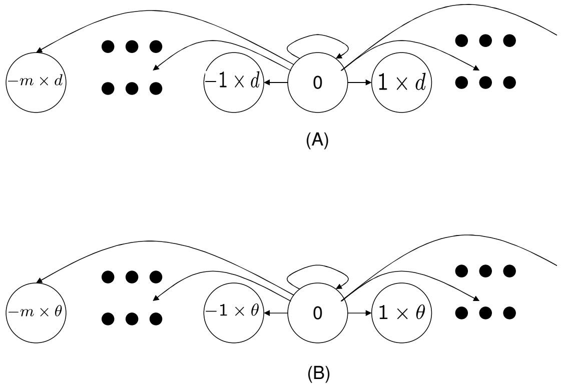

Fig. 7. The Markov state diagram for the mobility training. We only plot those arrows from 0 state for simpler illustration.

Algorithm 1: Parameter p_{ij}^r determination

```
Result: $p_{ij}^r$;
Input: A MS is launched to move according to the
  mobility model in the area shown in Fig. 5 by totally
  $N$ time slots;
States generation: The radiation distance made by the
  MS in each time slot are rounded up to the nearest
  state. Thus, we derive a series of states that the MS
  has experienced in the $N$ time slots;
$N_i$ determination: Count the number of transitions
  from state $i \times d$ among the series of states we
  derived, which is denoted as $N_i$;
$S_{ij}$ determination: Count the number of transitions
  from state $i$ to $j$ among the series of states we
  derived, which is denoted as $S_{ij}$;
$p_{ij}^r$ evaluation: The probability $p_{ij}^r$ is determined by
  $\frac{S_{ij}}{N_i}$.
```

Each state $i \times d$ represents that the MS is moving away from the AP by the speed of $\frac{i \times d}{\sigma}$ meters/sec in the current time slot where σ is the time duration per time slot, while each state $-i \times d$ represents that the MS is moving toward the AP by the speed of $\frac{i \times d}{\sigma}$ meters/sec. Let p_{ij}^r represent the transition probability that the radiation distance made by the MS at the current time slot is $j \times d$ meters given that the radiation distance made in the previous time slot is $i \times d$ meters. Similarly, we also partition the total radian of a circle (2π) into $2m$ states; i.e., $(-m \times \theta_r, -(m-1) \times \theta_r, \cdots, -\theta_r, 0, \theta_r, \cdots, (m-1) \times \theta_r, m \times \theta_r)$ to represent all possible radian variation per time slot, as shown in the part (B) of Fig. 7. Each state $i \times \theta_r$ represents that the MS is moving counter-clockwise by the speed of $\frac{i \times r\theta_r}{\sigma}$ meters/sec in the current time slot, while each state $-i \times \theta$ represents that the MS is moving clockwise by the speed of $\frac{i \times r\theta_r}{\sigma}$ meters/sec. Let $p_{ij}^{\theta_r}$ represent the transition probability that the distance variation over the

tangent direction of circle with radius r
the current time slot is $j \times r\theta_r$ meters gi
variation over the tangent direction made
slot is $i \times r\theta_r$ meters. Our approach
presented by Algorithm 1.

The derivation for $p_{ij}^{\theta_r}$ will be similar
p_{ij}^{r}, which jointly form the transition pr
are used to characterize the mobility patt
In the study, N is taken as 10^9 to obtai
of the user mobility patterns. With the tra
the Markov chain shown in Fig. 7 can be
After solving the Markov chain (part (A
transition probabilities p_{ij}^{r}, the stationar
MS to stay at the position state $k \times d$
(denoted as $p_{MT,k}^{r}$) can be obtained. Ther

that the MS transitions to state r becomes

Similarly, the stationary probability for th
$k \times \theta_r$ for $-m \leq k \leq m$ (denoted as

obtained, by which we can have: $\lambda_{\theta,r} =$

Suppose we have n_θ partitions of tota
RL partitions of the radius of the
$(RW - RL)$ partitions of the radiant dire
region, where n_θ, RL and RW are three
$RW > RL$. Each state shown in Fig. 8 i
components: the first component is the in
in the radiant partition and the second co
of sectors in the partition of the total radian (2π). Let three states exist in part (A) of Fig. 7, and the transition probability of two adjacent position states for the inward and outward radiation directions be the same. Thus, just one value, λ_R, is needed to represent all transition probabilities in the radiation direction. Similarly, let only three states exist in part (B) of Fig. 7, and the transition probabilities of counter-clockwise and clockwise direction be the same. Thus, a single quantity $\lambda_{\theta,i}$ can be used to represent the transition probabilities for the tangent direction when the MS locates at the i-th circle region. With the abovementioned approach, λ_R and $\lambda_{\theta,i}$ can be obtained to form a two-dimensional Markov chain, where the mobility model taken by the MS can be formulated as shown in Fig. 8.

After obtaining the transition probabilities between each pair of position states, the AP can derive the probability that the MS moves from the current position state (denoted as s_i) to some absorbing state s_j at the n-th step. This probability is also referred to as the *absorbing probability* of state s_i denoted as $p_{i,j}^{abs}(n)$, which is derived in the the Appendix. With the absorbing probability at any position state s_i, the AP can examine Eq. (1) by using the formula of Eq. (38) in the Appendix.

There are two possible ways to perform the condition examination of pre-handoff each MS. The first method is to compute parameters $c^{(u)}, x_i^{(u)}, y_j^{(u)}$ and t_u for $1 \leq u, i, j \leq \rho$ [3] off-line based on the knowledge of $p_{i,j}^{abs}(n)$ given that the total number of position states is ρ. In this case, the AP has to store

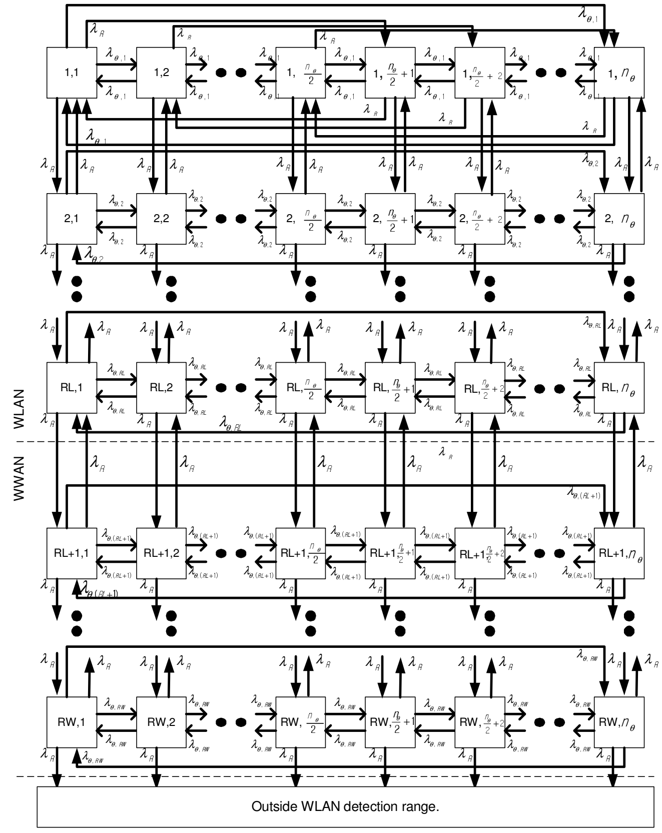
1,1
1,2
1, $\frac{n_\theta}{2}$
1, $\frac{n_\theta}{2}+1$
1, $\frac{n_\theta}{2}+2$
1, n_θ
2,1
2,2
2, $\frac{n_\theta}{2}$
2, $\frac{n_\theta}{2}+1$
2, $\frac{n_\theta}{2}+2$
2, n_θ
RL,1
RL,2
RL, $\frac{n_\theta}{2}$
RL, $\frac{n_\theta}{2}+1$
RL, $\frac{n_\theta}{2}+2$
RL, n_θ
RL+1,1
RL+1,2
RL+1, $\frac{n_\theta}{2}$
RL+1, $\frac{n_\theta}{2}+1$
RL+1, $\frac{n_\theta}{2}+2$
RL+1, n_θ
RW,1
RW,2
RW, $\frac{n_\theta}{2}$
RW, $\frac{n_\theta}{2}+1$
RW, $\frac{n_\theta}{2}+2$
RW, n_θ
λ_R
$\lambda_{\theta,1}$
$\lambda_{\theta,2}$
$\lambda_{\theta,RL}$
$\lambda_{\theta,(RL+1)}$
$\lambda_{\theta,RW}$
WLAN
WWAN
Outside WLAN detection range.

Fig. 8. The Markov state diagram of position states.

these $2(\rho + \rho^2)$ parameters, which corresponds to the user mobility patterns in the hotspot. When tracing and examining a MS in the hotspot, the AP requires $|A|\rho(2N_{th} + \frac{(N_{th}+1)N_{th}}{2})$ arithmetical operations for checking Eq. (1) in real-time, where the value of $|A|$ represents the number of position states in the absorbing region. The second method is to simply keep all the n-th transition probabilities between each pair of position states for $1 \leq n \leq N_{th}$, instead of doing any real-time calculation for deriving $p_{i,j}^{abs}(n)$. With this approach, the AP requires a memory to store $N_{th}\rho^2$ parameters, while the hardware is required to take $|A|\rho(2N_{th} - 1)$ arithmetical operations in real-time to check Eq. (1). In comparison of these two approaches, the first method obviously requires less memory space at the AP compared with that required in the second method at the expense of consuming more realtime computation resources. In practical, the choice of the first or second method should depend on the hardware/software supports at the AP and the number of position states defined in the WLAN.

IV. Performance Analysis

In this section, the following system performance measures are of interest and will be evaluated: the average vertical handoff delay, the average multi-tunnel duration, the probability of undo events, and the probability of missed detection events.

[3] There parameters are used in the evaluation of the absorbing probability as shown in the Appendix

A. Vertical Handoff Delay

The vertical handoff delay τ_{delay} comes from the fact that the MS fails to finish the setup of the new link to the FA_{new} before the MS losses the signal from the FA_{old}. The following analysis will focus on the case where the MS moves from the WLAN region to the WWAN region since a non-zero handoff delay could be introduced.

The probability that τ_{delay} equals to n time slots is given by:

$$Pr(\tau_{delay} = n|s_i) \;=\; \begin{cases} \sum\limits_{j \in \mathbf{S}^{abs}} p^{abs}_{i,j}(N_{set} - n) & \\ \qquad\qquad \text{if } 1 \leq n < N_{set}\ , & \\ (1 - \sum\limits_{k=1}^{N_{set-1}} \sum\limits_{j \in \mathbf{S}^{abs}} p^{abs}_{i,j}(N_{set} - k)) & \\ \qquad\qquad \text{if } n = 0, & \\ 0 \qquad\quad \text{if } n \geq N_{set}, & \end{cases}$$

where N_{set} is the number of time slots required to set up the new link, and $\mathbf{S}^{abs}$ is the set of those absorbing states in the WLAN. Let π be the set of those initial position states which satisfy Eq. (1) under a specific values of β and T_{th}. If we assume that the initial position of a MS is uniformly distributed on the area of wireless network region, the average τ_{delay} with respect to the initial position states, denoted as $avg.\tau_{delay}$, could be expressed as

$$avg.\tau_{delay}(T_{th}, \beta) \;=\; \sum_{i \in \pi} \frac{A_i \mathbb{E}[\tau_{delay}|s_i]}{\sum\limits_{k \in \pi} A_k}, \qquad (2)$$

where A_i is the area of the position state s_i, $\mathbb{E}[\tau_{delay}|s_i]$ represents the average τ_{delay} for the MS initiating its pre-handoff procedure at the position state s_i and is defined as

$$\mathbb{E}[\tau_{delay}|s_i] \;=\; \sum_{n=0}^{N_{set}} n Pr(\tau_{delay} = n|s_i). \qquad (3)$$

B. Multi-Tunnel Redundant Time

In addition to τ_{delay}, the multi-tunnel redundant time denoted as τ_{mcr} is also the performance measure of interest, which reflects the redundancy induced in the vertical handoff process. Let N_{mc} be the number of time slots consumed for setting up the new link. If the MS begins to initiate its pre-handoff procedure at the position state s_i, the probability that τ_{mcr} equals to n time slots where $0 \leq n \leq N_{mc}$ is given as:

$$Pr(\tau_{mcr} = n|s_i) \;=\; \sum_{j \in \mathbf{S}^{abs}} p^{abs}_{i,j}(N_{set} + N_{mc} - n), \quad \text{if } 0 \leq n \leq N_{mc}. \qquad (4)$$

Similarly, the average τ_{mcr} with respect to the initial position states, denoted as $avg.\tau_{mcr}$, could be expressed as

$$avg.\tau_{mcr}(T_{pre}, \beta) \;=\; \sum_{i \in \pi} \frac{A_i \mathbb{E}[\tau_{mcr}|s_i]}{\sum\limits_{k \in \pi} A_k}, \qquad (5)$$

where A_i is the area of the position state s_i and $\mathbb{E}[\tau_{mcr}|s_i]$ which represents the average τ_{mcr} for the MS initiating its pre-handoff procedure at the position state s_i is defined as

$$\mathbb{E}[\tau_{mcr}|s_i] \;=\; \sum_{n=0}^{N_{mc}} n Pr(\tau_{mcr} = n|s_i). \qquad (6)$$

C. Undo Probability

An undo event happens when the MS still stays in the same region of the state where the MS initiates its pre-handoff procedure after the time duration $T_{set} + T_0$. For a given initial position state, say s_i, the probability of an undo event, denoted as $P_{ud,i}$, can be expressed as

$$P_{ud,i} \;=\; \sum_{j \in \mathbf{S}^{ini}} p_{i,j}(N_{pre}), \qquad (7)$$

where $\mathbf{S}^{ini}$ represents the set of position states in the same network region as that of the initial position state, and N_{pre} is the number of time slots in the time duration of $T_{set} + T_0$. Let π be the set of the initial position states which satisfy the condition Eq. (1) under specific values of β and T_{th}, and the MS be uniformly distributed on the area of the same network region. The average undo probability with respect to the initial position states can be expressed as

$$P_{ud}(T_{th}, \beta) \;=\; \sum_{i \in \pi} \frac{A_i P_{ud,i}}{\sum\limits_{k \in \pi} A_k}, \qquad (8)$$

where A_i is the area of the position state s_i.

D. Missed Detection Probability

The event of missed detection happens when the MS moves to a different network region before the MS initiates its next examination of the pre-handoff criterion. For a given initial state, s_i, the probability of missed detection, denoted as $P_{md,i}$, can be written as

$$P_{md,i} \;=\; \sum_{k=1}^{N_{est}-1} \sum_{j \in \mathbf{S}^{abs}} p^{abs}_{i,j}(k), \qquad (9)$$

where N_{est} is the number of time slots between two consecutive examinations of the pre-handoff criterion made by the MS, and $\mathbf{S}^{abs}$ is the set of absorbing states with respect to the original network region. Thus, the average missed detection probability with respect to the initial position states can be expressed as:

$$P_{md}(T_{th}, \beta) \;=\; \sum_{i \in \pi} \frac{A_i P_{md,i}}{\sum\limits_{k \in \pi} A_k}, \qquad (10)$$

where π denotes the set of initial position states which satisfy the condition Eq. (1) under specific values of β and T_{th}.

V. Effects of Estimation Error

In this section, the effect of imperfect position estimation due to the estimation error of SINR will be investigated. It is well known that the received signal power decays with some exponent,α, in the path distance, d. The value of α depends on the channel and varies from 2 to 4 in general [17]. The formula characterizing the relationship between the transmitted signal power, Ω_t, and the received signal power, Ω_r, can be written as:

$$\Omega_r \;=\; \frac{C_{\text{ant}} \Omega_t}{d^{\alpha}}, \qquad (11)$$

where C_{ant} is a constant which depends on parameters of the transmitting and the receiving antenna.

Let the initial distance between the MS and the center of WLAN region be in a uniform distribution, and the maximum distance that can be sensed by the MS be denoted as R. Thus, the probability density function of the random variable for the sensed SINR [4], denoted as Υ, can be derived as follows :

$$\begin{aligned} F_{\Upsilon}(\rho) &= Pr(\Upsilon < \rho) \\ &= Pr(\frac{C_{\text{ant}}\Omega_t}{N_0 d^{\alpha}} < \rho) \\ &= Pr(\Big[\frac{C_{\text{ant}}\Omega_t}{\rho N_0}\Big]^{\frac{1}{\alpha}} < d). \end{aligned} \tag{12}$$

By taking the derivative of ρ at both sides of Eq. (12), we obtain the p.d.f. of Υ as

$$f_{\Upsilon}(\rho) = \frac{\Big[\frac{C_{\text{ant}}\Omega_t}{N_0\rho}\Big]^{\frac{1}{\alpha}}}{R\alpha\rho} \tag{13}$$

The estimator for the received SINR is based on Maximum Likelihood (ML) estimation, which was introduced by Gagliardi and Thomas [18]. The estimation can be performed to derive the p.d.f. of the estimated SINR ($\widehat{\rho}$) for M-ary PSK modulation over an AWGN channel. By setting N_{est} as the number of collected samples used in each received SINR estimation, we have the following p.d.f. for $\widehat{\rho}$ given that the true SINR value is ρ

$$f(\widehat{\rho}|\rho) = e^{-N_{est}\rho}\sum_{i=0}^{\infty}\frac{(N_{est}\rho)^i\Gamma(N_{est}+i)(\widehat{\rho})^{i-\frac{1}{2}}}{i!\Gamma(\frac{1+2i}{2})\Gamma(\frac{2N_{est}-1}{2})(1+\widehat{\rho})^{N_{est}+i}}, \tag{14}$$

where Γ denotes a Gamma function.

By Bayes rule, we have the p.d.f. of ρ given $\widehat{\rho}$ from Eq. (12) and Eq. (14), which yields:

$$\begin{aligned} f(\rho|\widehat{\rho}) &= \frac{f(\widehat{\rho}|\rho)f_{\Upsilon}(\rho)}{f(\widehat{\rho})} \\ &= \frac{f(\widehat{\rho}|\rho)f_{\Upsilon}(\rho)}{\int\limits_{\frac{C_{\text{ant}}\Omega_t}{N_0 R^{\alpha}}}^{\frac{C_{\text{ant}}\Omega_t}{N_0}} f(\widehat{\rho}|\rho)f_{\Upsilon}(\rho)d\rho}. \end{aligned} \tag{15}$$

From Eq. (11) and Eq. (15), we obtain the p.d.f. of r, denoted as $f_R(r|\widehat{\rho})$, which represents the distance from the MS to the AP given $\widehat{\rho}$. After some manipulation, $f_R(r|\widehat{\rho})$ can be expressed as:

$$f_R(r|\widehat{\rho}) = \frac{\alpha C_{\text{ant}}\Omega_t}{N_0 r^{\alpha+1}} f(\frac{C_{\text{ant}}\Omega_t}{N_0 r^{\alpha}}|\widehat{\rho}). \tag{16}$$

Let the estimation error be only introduced by the SINR estimation. Thus, we only need to consider the estimation error for the position states with different radiant distances from the AP. The probability that the true position state of the MS is s_j given the estimated position state at s_i can be evaluated from Eq. (16), which yields:

$$Pr(s_j|\widehat{s_i}) = \int_{\frac{C_{\text{ant}}\Omega_t}{N_0 r_{i,l}^{\alpha}}}^{\frac{C_{\text{ant}}\Omega_t}{N_0 r_{i,s}^{\alpha}}} \Big[\int_{r_{j,s}}^{r_{j,l}} f_R(r|\widehat{\rho})dr\Big] f(\widehat{\rho})d\widehat{\rho}, \tag{17}$$

where $r_{i,s}$ is the inner radius of the position state s_i, and $r_{i,l}$ is the outer radius of the position state s_i. A similar reason applies to the derivation of $r_{j,s}$ and $r_{j,l}$ at the position state s_j.

With Eq. (17), we should modify our previous performance criteria, such as the distribution of the vertical handoff delay and the distribution of the multi-tunnel duration, by averaging over all the possible true position states. Thus, the distribution of the vertical handoff delay given that the estimated position state is $\widehat{s_i}$, can be expressed as:

$$Pr(\tau_{delay} = n|\widehat{s_i}) = \sum_{j\in\mathbf{B}} Pr(\tau_{delay} = n|s_j)Pr(s_j|\widehat{s_i}), \tag{18}$$

where $\mathbf{B}$ is the network region that position state $\widehat{s_i}$ belongs to. Similarly, the distribution of the multi-tunnel duration given the estimated position state is $\widehat{s_i}$ can be expressed as:

$$Pr(T_0 = n|\widehat{s_i}) = \sum_{j\in\mathbf{B}} Pr(T_0 = n|s_j)Pr(s_j|\widehat{s_i}). \tag{19}$$

The undo probability given the estimated position state is $\widehat{s_i}$ will thus be modified as:

$$\widehat{P_{ud,i}} = \sum_{j\in\mathbf{B}} P_{ud,j}Pr(s_j|\widehat{s_i}), \tag{20}$$

and the missed detection probability given the estimated position state is $\widehat{s_i}$ becomes:

$$\widehat{P_{md,i}} = \sum_{j\in\mathbf{B}} P_{md,j}Pr(s_j|\widehat{s_i}). \tag{21}$$

VI. Numerical Results

In this section, the numerical and simulation results are obtained for the MS with $RL = 3$ and $n_{\theta} = 4$. The performance measures are evaluated according to the handoff procedure for the MS moving from the WLAN region to the WWAN region. All the figures are plotted with the perfect estimation of the position state except for Fig. 9. In Fig. 10, the average vertical handoff delay $avg.\tau_{delay}$ is plotted with respect to T_{th} under different values of λ_r by specifying $T_{set} = 10$ time slots. The other parameters used in the figure are $\lambda_{\theta,1} = \lambda_{\theta,2} = \lambda_{\theta,3} = 0.02$. With a larger value of T_{th} (e.g., > 50 slots), a larger value of λ_r is found, where a larger average vertical handoff delay will be introduced because the MS has a higher chance to leave the original network region when the value of λ_r becomes larger. However, with a smaller T_{th} (< 50 slots), we find that a larger λ_r will introduce a smaller average vertical handoff delay because the MS has a higher chance to trigger pre-handoff. By comparing with the conventional MIP or HAWAII, the two-step handoff scheme can reduce the average vertical handoff delay by 80% when $T_{th} > 50$ time slots.

In Fig. 11, the average vertical handoff delay $avg.\tau_{delay}$ is plotted with respect to T_{th} for different values of β by

[4] Because all the MS share the same channel, it would be more appropriate to consider the interference from other users and background noise.

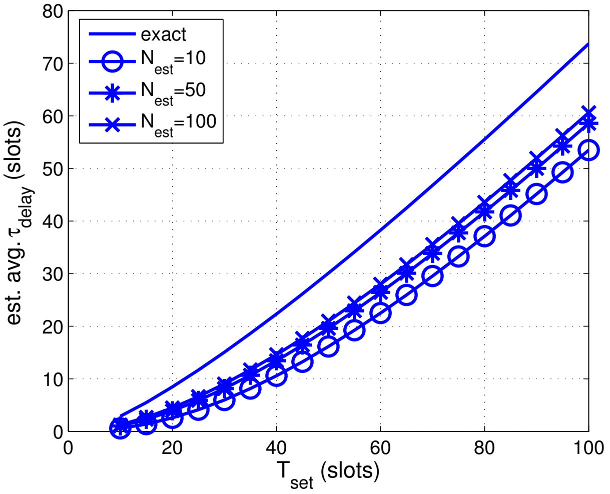

Fig. 9. Average vertical handoff delay for different estimation precision.

specifying $T_{set} = 10$ time slots and $\lambda_r = 0.1$. The other parameters are the same as the parameters used in Fig. 10. For the effect of different β, we find that with a larger value of β, a larger T_{th} is required to achieve a lower average handoff delay because the MS needs to trigger the pre-handoff scheme earlier as the value of β becomes larger. It is also observed that the simulation and analytic results match with each other quite well in Fig. 10 and Fig. 11.

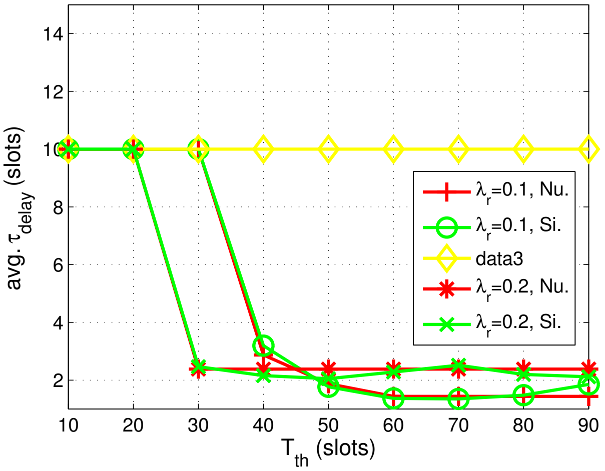

Fig. 10. Average vertical handoff delay for different λ_r.

In Fig. 12, we plot the average undo probability with respect to T_{th} for different λ_r with $\lambda_{\theta,1} = 0.02$, $\lambda_{\theta,2} = 0.02$ and $\lambda_{\theta,3} = 0.02$. With a larger value of λ_r, the MS has a higher chance to leave the original network region, which yields a smaller undo probability. The undo probability can also be reduced when T_{th} is increased since the MS has a lower chance to stay in the original network.

In Fig. 13, the average redundant time duration of multichannel condition, denoted as $avg.\tau_{mcr}$, for different initial position states is plotted with the T_{mc} by setting $T_{set} = 10$ slots. For the effect of λ_r, we observe that the average redundant time duration of multichannel condition decreases when the value of λ_r decreases. This phenomenon can be explained that the MS has less chance to leave the region of the WLAN earlier if the value of λ_r is small and it results in lower average redundant time duration of multichannel condition. For same value of λ_r, the average redundant time duration of multichannel condition increases with the increase of the distance between the center of WLAN region (AP) and the MS since the MS is easier to leave the WLAN region as the distance between the AP and the MS becomes larger.

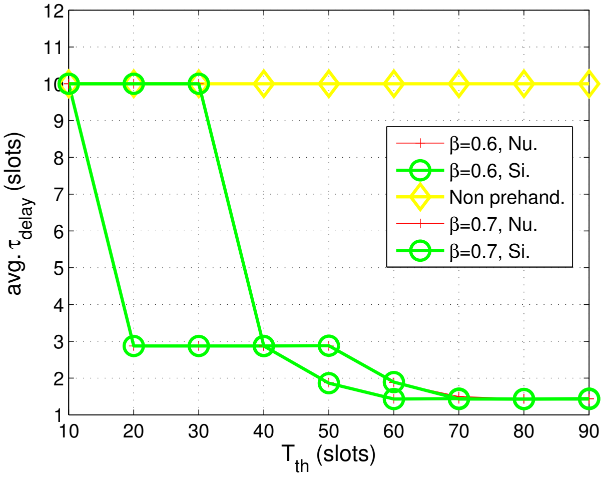

Fig. 11. Average vertical handoff delay for different threshold values β.

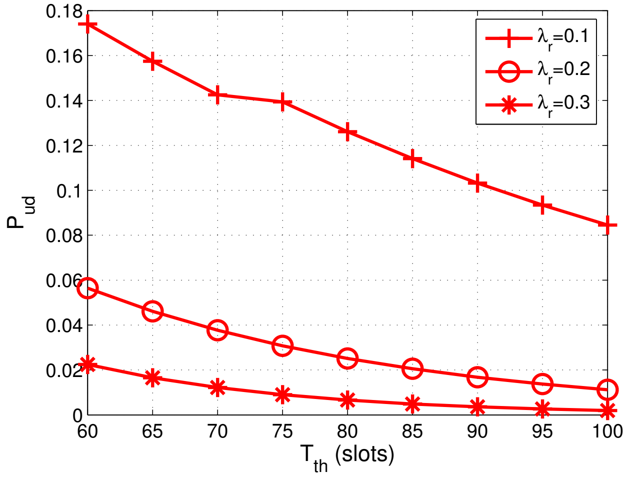

Fig. 12. Undo probabilities for different thresholds.

In Fig. 9, the average vertical handoff delay for the MS at the state position 5 (middle tire) is plotted with the following parameters: $\lambda_r = 0.2$, $\lambda_{\theta,1} = 0.1$, $\lambda_{\theta,2} = 0.05$ and $\lambda_{\theta,3} = 0.01$. It is observed that the difference between the estimated vertical handoff delay decreases at the expense of taking a larger number of samples (N_{est}) for SINR estimation increases, which obviously results in a higher overhead and hardware complexity in the implementation.

We have deployed the experimental topology illustrated in Fig. 2. Each mobile device is equipped with two network

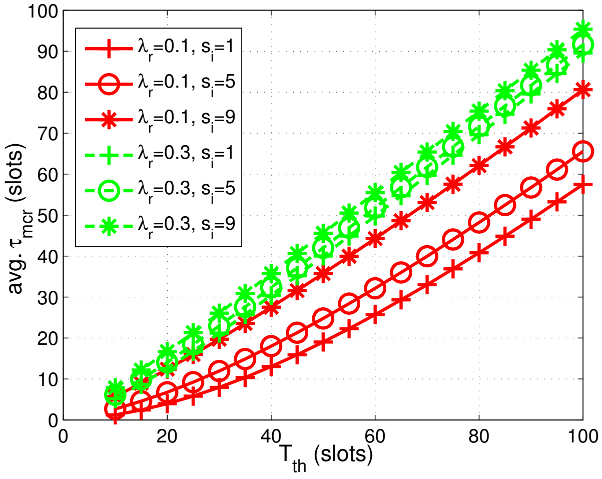

Fig. 13. The time duration of multi-tunnel for different initial positions.

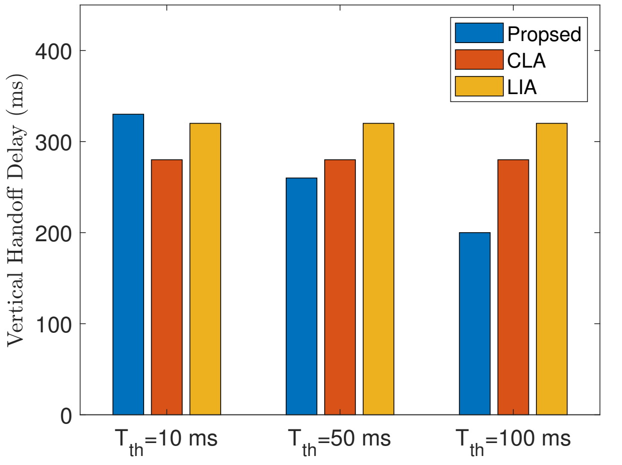

Fig. 14. The vertical handoff delay comparison with other existing methods.

interfaces WiFi and LTE. The simulation parameter values are: access bandwidth 20MBps, access link delay 10ms, and packet loss probability is 2%. We compare the proposed two-steps vertical handoff scheme (denoted as "Proposed") with CLA method [15] and LIA method [14] in Fig. 14. When the value T_{Th} is getting larger, we can observe that the vertical delay for the proposed method becomes shortest compared to current existing handoff methods.

VII. Conclusions

This paper has proposed and evaluated a two-steps vertical handoff scheme based on a suite of novel mobility prediction and parameter training processes. The proposed scheme is characterized by a cooperative decision making process between the AP and the MS through the IEEE 802.21 standard, where the AP performs the parameter training and mobility prediction and makes the pre-handoff decision for the accessing MSs. We have justified the proposed framework and demonstrated detailed procedures in the implementation of the proposed approach, which was further evaluated through four performance measures of each vertical handoff event, including the average handoff delay, average multi-tunnel time duration, undo probability and missed detection probability. In addition, the effect of the estimation error for the position state of the MS was studied by using a received SINR estimator based on the maximum likelihood (ML) estimation. We observed that the proposed approach can achieve various performance objectives and multiple classes of service by manipulating the proposed system parameters: the confidence (denoted as β) for the MS to move to the boundary of the two network domains after a specific duration (denoted as T_{th}). The simulation and analytic results on the average vertical handoff delay and multi-tunnel time match with each other very well.

In this appendix, we will derive explicit expressions for the n-step transition probability of a finite state Markov chain with the state number ρ and transition probability matrix $[p_{jk}]$ (Chap 16 of [19] v.1).

By taking the z-transform for the sequence $\{p^0_{j,k}, p^1_{j,k}, p^2_{j,k}, \cdots\}$, we obtain

$$P_{j,k}(z) \;=\; \sum_{n=0}^{\infty} p^n_{j,k} z^n. \qquad (22)$$

Multiplying $zp_{i,j}$ at both sides of Eq. (22) and adding over $j \in \{1, 2, \cdots, \rho\}$, we have a set of linear equations for each k. They are

$$P_{i,k}(z) - z\sum_{j=1}^{\rho} p_{i,j} P_{j,k}(z) \;=\; p_{i,k}. \qquad (23)$$

for $1 \le i \le \rho$. The solutions for $P_{j,k}(z)$ are rational functions of z with the common denominator $D(z)$, the determinant of coefficients of the system equations. Let $\{t_1, t_2, \cdots, t_\rho\}$ be the eigenvalues of the transition probability matrix P. We assume that these roots are distinct and not equal to zero since most interesting cases will be covered by this assumption. After solving the above linear system equations Eq. (23) and doing partial fraction decomposition, we can express $P_{j,k}(z)$ as

$$P_{j,k}(z) \;=\; \frac{b^{(1)}_{j,k}}{1 - zt_1} + \frac{b^{(2)}_{j,k}}{1 - zt_2} + \cdots + \frac{b^{(\rho)}_{j,k}}{1 - zt_\rho}. \quad (24)$$

By taking the inverse Z-transform of Eq. (24), we obtain $p^{(n)}_{j,k}$ as

$$p^n_{j,k} \;=\; b^{(1)}_{j,k} t^n_1 + b^{(2)}_{j,k} t^n_2 + \cdots + b^{(\rho)}_{j,k} t^n_\rho. \qquad (25)$$

The next step is to determine the quantities $b^{(1)}_{j,k}, b^{(2)}_{j,k}, \cdots, b^{(\rho)}_{j,k}$. Note that the quantity of $p^{(n+1)}_{i,k}$ can be evaluated in two different ways. The first way is to change the subscript (n) of Eq. (25) to $(n+1)$. The second is to use the property that $p^{(n+1)}_{i,k} = \sum_{j=1}^{\rho} p_{i,j} p^{(n)}_{j,k}$. Then we have the following identity

$$\Big(\sum_{j=1}^{\rho} p_{i,j} b^{(1)}_{j,k} - t_1 b^{(1)}_{i,k}\Big) t^n_1 + \Big(\sum_{j=1}^{\rho} p_{i,j} b^{(2)}_{j,k} - t_2 b^{(2)}_{i,k}\Big) t^n_2 + \cdots + \Big(\sum_{j=1}^{\rho} p_{i,j} b^{(\rho)}_{j,k} - t_\rho b^{(\rho)}_{i,k}\Big) t^n_\rho = 0 \quad (26)$$

For all the combination of indexes i, k, the coefficients of $t_1^n, t_2^n, \cdots, t_\rho^n$ need to be zero for the equation Eq. (26) satisfied for any n. Therefore, we have the following ρ equations

$$\sum_{j=1}^{\rho} p_{i,j} b_{j,k}^{(v)} - t_v b_{i,k}^{(v)} = 0, \tag{27}$$

for $1 \leq v \leq \rho$. On the other hand, if we multiply $p_{k,r}$ of equation Eq. (25) and sum up over index k, we also obtain the following ρ equations by similar method used in deriving Eq. (26). They are

$$\sum_{k=1}^{\rho} b_{j,k}^{(v)} p_{k,r} - t_v b_{j,r}^{(v)} = 0, \tag{28}$$

for $1 \leq v \leq \rho$. We also can write Eq. (27) and Eq. (28) as matrix multiplication form as $Pb^{(v)} = t_v b^{(v)}$ and $b^{(v)} P = t_v b^{(v)}$, respectively. Let $x_i^v = c_i^v b_{i,k}^{(v)}$ for some fixed k, where c_i^v is a constant independent of k. Then the k-th column of $b^{(v)}$ represents a solution of the following ρ linear equations

$$\sum_{j=1}^{\rho} p_{i,j} x_j^v = t_v x_i^v, \tag{29}$$

and, similarly, the j-th row of $b^{(v)}$ represents a solution of the following ρ linear equations

$$\sum_{k=1}^{\rho} y_k^v p_{k,r} = t_v y_r^v, \tag{30}$$

where $y_r^v = c_r^v b_{j,r}^{(v)}$ by some constant c_r^v. By solving Eq. (29) and Eq. (30) for each t_v, we have $(x_1^{(v)}, x_2^{(v)}, \cdots, x_\rho^{(v)})$ and $(y_1^{(v)}, y_2^{(v)}, \cdots, y_\rho^{(v)})$. Then $b_{j,k}^v$ can be determined as

$$b_{j,k}^v = c^{(v)} x_j^v y_k^v, \tag{31}$$

where $c^{(v)}$ is a constant independent of j, k. To evaluate the value of $c^{(v)}$, we use the following relation for $1 \leq v \leq \rho$

$$\sum_{j=1}^{\rho} p_{i,j}^{(n)} x_j^v = t_v^n x_i^v. \tag{32}$$

Equation Eq. (32) can be proved by mathematical induction of n. By using the expression of Eq. (25) for $p_{i,j}^n$ and Eq. (31), we have

$$\begin{aligned} t_u^n x_i^v = \; & t_1^n c^{(1)} x_i^{(1)} \sum_{k=1}^{\rho} y_k^{(1)} x_k^{(u)} + t_2^n c^{(2)} x_i^{(2)} \sum_{k=1}^{\rho} y_k^{(2)} x_k^{(u)} + \\ & \cdots + t_\rho^n c^{(\rho)} x_i^{(\rho)} \sum_{k=1}^{\rho} y_k^{(\rho)} x_k^{(u)} \end{aligned} \tag{33}$$

for any $u \in \{1, 2, \cdots, \rho\}$. Since Eq. (33) holds for all n, the coefficient of t_u^n must be zero. Therefore, we have

$$c^{(u)} \sum_{k=1}^{\rho} x_k^u y_k^u = 1, \tag{34}$$

for $1 \leq u \leq \rho$. Finally, $p_{j,k}^n$ can be expressed as

$$p_{j,k}^n = \sum_{u=1}^{\rho} c^{(u)} x_j^{(u)} y_k^{(u)} t_u^n, \tag{35}$$

where $x_j^{(u)}, y_k^{(u)}$ and $c^{(u)}$ are evaluated from Eq. (29), Eq. (30) and Eq. (34) respectively.

A closed state set is defined as a set of states in which the system cannot transfer out from this set whenever any one state of this set is reached. We use SC_j to represent a state in the closed state set with index j. The remaining states except those states in closed state sets are called transitive states. For a finite state Markov chain with k closed state sets, we always can partitioned the transition matrix as

$$\mathbf{P} = \begin{vmatrix} E_1 & 0 & 0 & \ldots & 0 \\ 0 & E_2 & 0 & \cdots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots \\ 0 & \cdots & 0 & E_k & 0 \\ A_1 & A_2 & \cdots & A_k & T \end{vmatrix}, \tag{36}$$

where $E_1, E_2, \cdots, E_k$ are transition matrices for those k closed state sets. The matrix T is the transition matrix of transitive states and $A_1, A_2, \cdots, A_k$ are matrices from transitive states to closed states. The n-step transition matrix P^n becomes

$$\mathbf{P^n} = \begin{vmatrix} E_1^n & 0 & 0 & \ldots. & 0 \\ 0 & E_2^n & 0 & \cdots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots \\ 0 & \cdots & 0 & E_k^n & 0 \\ A_{1,n} & A_{2,n} & \cdots & A_{k,n} & T^n \end{vmatrix}, \tag{37}$$

The quantities of elements in matrices $A_{1,n}, A_{2,n}, \cdots, A_{k,n}$ are related to absorption probabilities. If the system starts from a transitive state with index i, we want to obtain the probability of reaching closed state with index j at step n which is denoted as $p_{i,j}^{abs}(n)$. This probability can be evaluated as

$$p_{i,j}^{abs}(n) = \sum_{k=1}^{N_T} p_{i,k}^{n-1} p_{k,j}, \tag{38}$$

where N_T is the total number of transitive states. Note that the probability $p_{i,k}^{n-1}$ can be evaluated according to Eq. (35) with the transition probability matrix T.

REFERENCES

[1] Q. Zhang, C. Guo, Z. Guo, and W. Zhu, “Efficient mobility management for vertical handoff between WWAN and WLAN,” *IEEE Commun. Mag.*, vol. 41, no. 11, pp. 102–108, Nov. 2003.

[2] M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, S. Miller, and L. Salgarelli, “Integration of 802.11 and third-generation wireless data networks,” in *Proc. INFCOM*, Mar. 2003, pp. 503–512.

[3] C. Guo, Z. Guo, Q. Zhang, and W. Zhu, “A seamless and proactive end-to-end mobility solution for roaming across heterogeneous wireless networks,” *IEEE J. Select. Areas Commun.*, vol. 22, no. 5, pp. 834–848, June 2003.

[4] *Draft IEEE Standard for Local and Metropolitan Area Networks: Media Independent Handover Services: IEEE P802.21/D03.00*, Std., Dec. 2006.

[5] C. E. Perkins, “Ip mobility support,” IETF RFC 2002, Oct. 1996.

[6] C. E. Perkins, *Mobile IP: Design Principles and Practices*, 1st ed. Prentice Hall PTR, 1998.

[7] J. Schiller, Ed., *Mobile Communications*. Addision Wesely, 2003.

[8] I.-W. Wu, W.-S. Chen, H.-E. Liao, and F. F. Young, “A Seamless Handoff Approach of Mobile IP Protocol for Mobile Wireless Data Networks,” *IEEE Trans. Consumer Electr.*, vol. 48, no. 2, pp. 335–344, May 2002.

[9] S.-C. Lo, G. Lee, W.-T. Chen, and J.-C. Liu, “Architecture for Mobility and QoS Support in all-IP Wireless Networks,” *IEEE J. Select. Areas Commun.*, vol. 22, no. 4, pp. 691–705, May 2004.

[10] *IETF Working Group: Low Latency Handoffs in Mobile IPv4*, http://www.ietf.org/internet-drafts/draft-ietf-mobileip-lowlatency-handoffs-v4-11.txt.

[11] M. H. Ye, Y. Liu, and H. M. Zhang, "The mobile ip handoff between hybrid networks," in *Proc. IEEE Intn'l Symposium on Personal, Indoor, and Mobile Radio Communications*, Sept. 2002, pp. 265–269.

[12] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S. Y. Wang, and T. L. Porta, "HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, pp. 396–410, June 2002.

[13] A. Campbell, J. Gomez, S. Kim, A. Valko, C. Y. Wan, and Z. Turanyi, "Design, implementation and evaluation of cellular ip" ieee personal communications," *IEEE Trans. Wireless Communications*, vol. 7, pp. 42–49, Aug. 2000.

[14] L. Li, K. Xu, T. Li, K. Zheng, C. Peng, D. Wang, X. Wang, M. Shen, and R. Mijumbi, "A measurement study on multi-path TCP with multiple cellular carriers on high speed rails," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 161–175.

[15] H. Sinky, B. Hamdaoui, and M. Guizani, "Proactive multipath TCP for seamless handoff in heterogeneous wireless access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4754–4764, 2016.

[16] K. Gao, C. Xu, P. Zhang, J. Qin, L. Zhong, and G.-M. Muntean, "GCH-MV: Game-enhanced compensation handover scheme for multipath TCP in 6G software defined vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 142–16 154, 2020.

[17] T. S. Rappaport, *Wireless Communications – Principles and Practice*, 2nd ed. Prentice Hall PTR, 2002.

[18] R. Gagliardi and C. Thomas, "PCM data reliability monitoring through estimation of signal-to-noise ratio," *IEEE Commun. Mag.*, vol. 16, no. 3, pp. 479–486, June 1968.

[19] W. Feller, Ed., *An Introduction of Probability Theory and Its Applications*. John Wiley & Sons, 1957.

Shih Yu Chang (Senior Member, IEEE) received a B. S. E. E. degree from National Taiwan University, Taiwan, in 1998, and Ph. D. degrees in electrical engineering and computer engineering from University of Michigan, Ann Arbor, in 2006. From August 2006 to February 2016, he was the faculty in the Department of Computer Engineering, National Tsing Hua University, Hsinchu, Taiwan. From July to August 2007, Dr. Wu had been a visiting assistant professor at Television and Networks Transmission Group, Communications Research Centre, Ottawa, Canada. From June 2018, he began to provide lectures about machine learning, data science, and AI in San Jose State University, San Jose, CA, USA. Besides academic position, Dr. Chang also works as an AI technical lead focusing on applying machine learning techniques to automate office works.

Dr. Chang has published more than 80 peer-refereed technical journals and conference articles in electrical and computer engineering. His research interests include the areas of wireless networks, wireless communications and signal processing. He currently serves as the technical committee, symposium chair, track chair, or the reviewer in networking, signal processing, communications, and computers.

Pin-Han Ho (Fellow, IEEE) is currently a Full Professor with the Department of Electrical and Computer Engineering, University of Waterloo. He is the author/coauthor of over 400 refereed technical papers and several book chapters. He is the coauthor of two books on the Internet and optical network survivability. His current research interests include wired and wireless communication networks involving transmission techniques, mobile system design and optimization, and Internet of things (IoT) operation and deployment. He is a Professional Engineer Ontario (PEO).