Seamless VoWLAN Handoff Management based on Estimation of AP Queue Length and Frame Retries

Muhammad Niswar, Eigo Horiuchi, Shigeru Kashihara,
†Kazuya Tsukamoto, Youki Kadobayashi, Suguru Yamaguchi
Nara Institute of Science and Technology, †Kyushu Institute of Technology, JAPAN
Email:{niswar-m, eigo-h, shigeru, youki-k, suguru}@is.naist.jp, †tsukamoto@cse.kyutech.ac.jp

***Abstract*—Switching a communication path from one Access Point (AP) to another in inter-domain WLANs is a critical challenge for delay-sensitive applications such as Voice over IP (VoIP) because communication quality during handoff (HO) is more likely to be deteriorated. To maintain VoIP quality during HO, we need to solve many problems. In particular, in bi-directional communication such as VoIP, an AP becomes a bottleneck with the increase of VoIP calls. As a result, packets queued in the AP buffer may experience a large queuing delay or packet losses due to buffer overflow, thereby causing the degradation of VoIP quality for the Mobile Nodes (MNs) side. To avoid this degradation, MNs need to appropriately and autonomously execute HO in response to the change in wireless network condition, i.e., the deterioration of wireless link quality and the congestion state at the AP. In this paper, we propose an HO decision strategy estimating AP queue length at an MN and exploiting frame retries to maintain VoIP quality during HO.**

I. Introduction

Wireless LAN (WLAN, IEEE802.11a/b/g) has been the dominant wireless network and is extensively deployed today. Meanwhile, there is a huge demand for Voice over IP (VoIP) service over WLANs. However, delivering VoIP over WLANs (VoWLANs) has many challenges because VoIP is a delay and packet loss sensitive application. In some metropolitan areas, WLANs (WiFi hotspots) have already provided Internet connectivity to many mobile nodes (MNs) everywhere. In such an environment, the MNs are likely to traverse several WLANs with different IP subnets during a VoIP call because the coverage of individual WLAN is relatively small. Consequently, VoWLAN quality could be drastically degraded due to the severe wireless network condition caused by the increase of the number of the MN and the movement of MN. Therefore, to maintain VoWLANs quality, MNs need to appropriately and autonomously execute handoffs (HOs) in response to the wireless network condition.

In a mobile environment, typically, two main factors degrade VoWLAN quality: (1) degradation of wireless link quality and (2) congestion at an access point (AP). First, as an MN freely moves across WLANs, the communication quality degrades due to the fluctuation of wireless link condition. Second, as VoIP is a bi-directional communication, an AP becomes a bottleneck with the increase of VoIP calls. That is, VoIP packets to MNs are liable to experience a large queuing delay or packet loss due to buffer overflow in the AP buffer because each MN and AP has almost the same priority level of frame transmission by following the CSMA/CA scheme. In addition, in multi-rate WLANs, although a rate adaptation function changes transmission rate in response to wireless link condition, low transmission rate occupies more wireless resources than a high transmission rate. Thus, compared with a high transmission rate, a low transmission rate tends to cause a congestion at an AP. Therefore, to maintain VoWLAN quality, we need to develop an HO strategy considering these two factors in WLANs.

So far, many researchers have studied HO strategies. Although most of them focus on the mechanism to switch wireless networks, they do not sufficiently study an HO strategy considering both wireless network condition and characteristics of an application. In a bi-directional real-time communication such as VoIP, packets queued in the AP buffer experience a queuing delay or packet loss, thereby resulting in degradation of VoIP quality for MN. However, a common AP, which has already spread, does not have a mechanism to report the congestion state to MNs. Thus, MNs needs to estimate the occurrence of the congestion at the AP for avoiding degradation of VoIP quality.

In this paper, first, we study a method to estimate AP queue length at an MN side to detect the congestion in a WLAN. Then, we propose a new HO strategy method considering wireless network conditions, i.e., the deterioration of wireless link condition and congestion at the AP. Finally, we show the effectiveness of our proposed method through simulation experiments.

II. Related Work

Many HO decision strategies have been studied for various layers of the protocol stack where network and transport layers are most widely studied. Mobile IP [**?**] is a network layer scheme utilizing and relying on network infrastructures. However, an HO process in Mobile IP takes a significant time period including the period for acquisition of the IP address in a new WLAN and binding update to a Home Agent and a Corresponding Node (CN). On the transport layer, mobile Stream Control Transmission Protocol (mSCTP) [**?**], which is a mobility extension of SCTP, has been proposed. Although mSCTP supports multi-homing and dynamic address reconfiguration for mobility, the issue of the HO decision is not discussed in details. The authors in [**?**] proposed an SCTP based HO scheme for VoIP using a Mean Opinion Score (MOS [**?**]) as an HO decision metric. This HO mechanism

Fig. 1: Simulation Model 1

Fig. 2: Relationship between RTS Retry and MOS over Distances

Fig. 3: Relationship among # of MNs, AP queue length and MOS

TABLE I: Simulation Parameters

VoIP Codec	G.711
WLAN Standard	IEEE802.11g
Supported Data Rate	6 9 12 18 24 36 48 54Mbps
Fading Model	Nakagami Ricean K = 4.84
SIFS	16μs
Slot Time	9μs
CWmin, CWmax	15, 1023

employs a probe message called heartbeat to estimate a Round Trip Time (RTT) and calculates MOS value based on the RTT. However, since upper layer (above layer 3) information such as packet loss, RTT, and MOS indicate end-to-end communication quality, the information is varied due to both the wireless and wired network [**?**]. Therefore, the existing study could execute unnecessary HOs due to some factors in the wired network, such as temporal congestion (not in the wireless network). That is, in a mobile environment, MNs need to promptly and reliably detect wireless link condition by exploiting the lower layer (below layer 2) information. Furthermore, our practical experiments in [**?**] proved that the number of frame retries on the MAC layer has the potential to detect the wireless link degradation during movement because packets over wireless inevitably experience frame retries before being treated as packet loss.

Ref. [**?**] proposed an HO mechanism employing the number of frame retries as an HO decision metric through analytical study. This method, however, only considers the frame retransmission caused by the collision with frames transmitted from other MNs in a non-interference environment. On the other hand, previously, we proposed an HO strategy method considering the number of frame retries on the MAC layer [**?**] [**?**] [**?**]. This strategy employs single-path and multi-path transmission modes to execute soft-HO between two WLANs with different IP subnets. Although our previous method can detect the degradation of wireless link condition due to both movement of MN and radio interference, it cannot detect congestion at a targeted AP. This is because our previous method detects wireless link condition based on only frame retries without considering congestion at the AP. Therefore, in our previous method, an MN could execute an HO to a congested AP, and then VoIP quality would be degraded.

III. HO DECISION METRICS

We discuss HO decision metrics that can precisely indicate wireless network condition. Many HO technologies employ Received Signal Strength (RSS) on PHY layer as an HO decision metric. However, according to our practical experiments [**?**], RSS is very difficult for an MN to properly detect deterioration in communication quality because RSS fluctuates abruptly due to distance and interfering objects. It also cannot detect the degradation due to radio interference. Furthermore, in [**?**], we showed that the information of the MAC layer, frame retry, has a potential to serve as a significant metric. In this section, we describe two HO metrics employed in our new proposed method.

A. Frame Retries

In the IEEE802.11 standard, a sender confirms a successful transmission by receiving an ACK frame in response to the transmitted data frame. When a data or ACK frame is lost, the sender retransmits the same data frame until achieving a successful transmission or reaching a predetermined retry limit. If Request-to-Send/Clear-to-Sent (RTS/CTS) is applied, a retry limit of four is applied, otherwise, a retry limit of seven is applied. When frame retries reach the retry limit, the sender treats the data frame as a lost packet. That is, we can detect the occurrence of packet loss in advance by utilizing frame retries. Moreover, unlike the RSS, frame retries can promptly and reliably detect the wireless link degradation due to not only reduction of RSS but also radio interference and collisions [**?**]. Therefore, frame retries allows an MN to detect wireless link condition properly.

In [**?**], we employed data frame retry as an HO decision metric in WLANs with a fixed transmission rate (11 Mb/s). However, in a real environment, almost all WLANs employ a multi-rate function which can change the transmission rate according to wireless link condition. If the transmission rate is dropped through the multi-rate function, more robust modulation type is used and thus data frame retries are decreased. Thus, an MN cannot properly detect the degradation of wireless link quality only from data frame retries in multi-rate WLANs. We then consider RTS frame as an alternative metric of data frame retries. As RTS frame is always transmitted at the lowest rate (6 Mb/s), an MN can appropriately detect the change of wireless link quality. To show the effectiveness, we investigate the behavior of RTS retry ratio when an MN moves away from an AP through a simulation experiments.

Fig. **??**a and Table **??** show a simulation model and parameters, respectively. Note that we employ MOS [**?**] to assess the VoIP quality where MOS$\geq$3.6 indicates an adequate VoIP call quality. We also employ RTS retry ratio instead of the number of RTS retries. The RTS retry ratio is calculated as follow:

Fig. 4: AP Queuing Delay and RTT between AP and MN

Fig. 5: Relationship among AP queue length, RTT, and MOS

Fig. 6: Switching to Single/Multi-Path Transmission

Fig. 7: Switching to Single-Path Transmission

$$RTS\ Retry\ Ratio = \frac{Number\ of\ RTS\ Frame\ Retries}{Total\ Transmitted\ Frames} \tag{1}$$

Note that the number of RTS frame retries and the total transmitted frames are sampled every 100 ms.

Fig. **??** shows a relationship between the MOS and RTS retry ratio as a function of distance between the AP and the MN. We can see that the MOS is degraded with the increase in the RTS retry ratio when the MN moves away from the AP. Since the RTS retry ratio is drastically varied due to the fluctuation of wireless link quality, we employ a least-squares method to grasp their trend and estimate the best fit of the occurrences of RTS retry ratio over the distance shown as a straight line. The straight line shows that an RTS retry ratio of 0.6 indicates the starting point of VoIP quality degradation. Therefore, we set the RTS retry ratio of 0.6 as one of the thresholds to execute the HO in this study.

B. AP Queue Length

With the increase of VoIP calls in a WLAN, packets queued in an AP buffer are increased as well. When AP queue length increases, each of the packets queued in the AP buffer experiences a large queuing delay or packet loss due to buffer overflow. Consequently, the queuing delay and the packet loss severely affect VoIP quality of MNs.

Unfortunately, the IEEE802.11 (a/b/g/n) standard does not provide a mechanism to inform MNs of AP queue length. Therefore, to maintain VoIP quality, an MN needs to detect the congestion of the AP from an MN side. We then investigate the relationship between the number of MNs (VoIP calls) and AP queue length through simulation experiments using Qualnet 4.0.1 [**?**]. Fig. **??**b and Table **??** show a simulation model and parameters, respectively. In the simulation scenario, MNs are randomly located in a WLAN. Fig. **??** shows the relationship among the number of MNs, AP queue length, and MOS. From Fig. **??**, we can see that VoIP quality of MNs (MN MOS value) degrades with the increase of AP queue length. On the other hand, at the CN side (CN MOS value), VoIP quality is kept in adequate quality even if the number of VoIP calls increases. That is, a bottleneck of AP seriously affects only flows from AP to MNs.

From Fig. **??**, we found the significance of the AP queue length. However, how can MNs detect AP queue length without modifying an AP? Therefore, we propose a method to estimate AP queue length based on RTT between MN and AP. As illustrated in Fig. **??**, MN periodically sends a probe packet (ICMP message) to an AP and then calculates RTT between the MN and the AP. The RTT increases in response to the increase of AP queuing delay because a probe response packet to MN experiences queuing delay in the AP buffer. Therefore, the RTT can be used to derive information about AP queuing delay. We then investigate the relationship between AP queue length and the RTT between MN and AP through simulation experiment using the simulation model in Fig. **??**b. From Fig. **??**, we can see that the RTT increases with the increase of AP queue length. The graph also shows that the RTT should be kept under 200 ms to satisfy adequate VoIP quality. Therefore, in our proposed method, we employ RTT between MN and AP to estimate AP queue length and set the RTT threshold (RTT_thr) of 200 ms to maintain the adequate VoIP quality.

IV. Proposed HO Strategy

As described in Sec. III, we employ both RTS frame retry ratio and AP queue length as HO decision metrics. To adapt to multi-rate and congested WLANs, we then propose an HO strategy method based on reference [**?**]. In [**?**], an MN has two WLAN interfaces (IFs), and an HO Manager (HM) implemented on transport layer controls HO based on HO decision metrics.

A. Single-Path and Multi-Path Transmission

Our proposed HO method employs multi-homing similar to [**?**]. The HM properly switches between single-path and multi-path transmission modes in response to wireless network condition. Single-path transmission mode means that an MN communicates with a CN using only one IF. Multi-path transmission, on the other hand, means that an MN sends duplicated packets to a CN through two IFs to support soft-HO.

Fig. **??** shows an algorithm of switching to single/multi-path transmission when an MN is located in an overlap area of two APs. An MN associated with two APs (AP1 and AP2) transmits a probe packet at every 500 ms intervals to estimate AP queue length of each AP. If both RTT values between MN and AP1/AP2 are below an RTT threshold (RTT_thr: 200 ms), an MN detects that both APs are not congested. Then, the MN investigates RTS frame retry ratio of the current active IF since it also affects wireless link condition. If the

Fig. 10: Calculating RTT from captured probe packet and obtaining the right to send probe packets

Fig. 8: HO based on RTS Frame Retry Ratio

Fig. 9: HO based on Transmission Rate

RTS frame retry ratio reaches a retry ratio threshold of single-path (R_Sthr: 0.6), the HM switches to multi-path mode to investigate both wireless link condition of these two IFs as well as supporting soft-HO. On the other hand, if the RTT of AP1 reaches RTT_thr, i.e., AP1 is congested, and an MN switches to the AP2 directly without switching to multi-path mode because multi-path mode may cause a serious congestion in WLANs. If both measured RTTs reach RTT_thr, an MN then investigates the wireless link condition by using the RTS frame retry ratio of the current active IF.

In a multi-path transmission, to maintain VoIP quality, an MN sends duplicate data packets through two WLAN IFs, hence, the MN needs to switch back to single-path transmission as soon to prevent unnecessary network overload. As shown in Fig. **??**, an algorithm of switching to single-path transmission works as follows. First, an MN measures RTTs of both APs. If either of the RTTs is below the RTT_thr, the MN switches to an IF with a smaller RTT. If both RTTs are below the RTT_thr, the MN then compares the RTS frame retry ratio of both IFs. Fig. **??** shows an algorithm for the comparison of the RTS frame retry ratio obtained from both IFs. If both RTS frame retry ratios of the IFs are equal, the MN continues multi-path mode. On the other hand, if either of the frame retries is below the retry threshold of multi-path (R_Mthr: 0.4), the MN switches to single-path mode through the IF with a small retry ratio.

B. Deal with Ping-Pong Effect

If all MNs send probe packets to measure the RTT between MN and AP as proposed in Sec. IV-A, the MNs may unfortunately detect congestion of the serving AP (e.g., AP1) at nearly the same time. Then, all MNs may switch the communication to a neighbor AP (e.g., AP2) and leave the AP1 simultaneously. As a result, neighbor AP2's queue length is drastically increased, and then, all MNs detect the congestion at the AP2 and switch back to the AP1 again. This phenomena is typically called ping-pong effect and leads to degradation of VoIP quality due to fluctuation of both APs queue length.

To avoid the ping-pong effect, we extend the strategy proposed in Sec. IV-A. In the extension method, all MNs first examine their own current transmission rate before executing HO. Fig. **??** shows an algorithm of HO based on transmission rate. A WLAN provides a multi-rate function that can change the transmission rate dynamically based on wireless link condition. As mentioned earlier, since an MN with lower transmission rate occupies more wireless resources, the MN is liable to lead to congestion of an AP. Moreover, as MNs with the lowest transmission rate typically are far away from the connected AP, that is, near the edge of its coverage, they have to execute handover as soon as possible to maintain their communication quality. Therefore, in the proposed scheme, MNs with the lowest transmission rate (6 Mb/s) first execute HO. Then, if the AP queue length is still high even after $Time\_thr$ ($CurrTime - LastTime$) of 2 seconds expires, MNs with the next lowest transmission rate (12 Mb/s) starts to execute HOs. Note that an MN does not need to know the transmission rate of other MNs because we assume that every MN employs this algorithm to deal with the issue of synchronization of all MNs' transmission rates.

C. Elimination of Redundant Probe Packets

If every MN measures RTT by using probe packets according to the method proposed in Sec. IV-B, these probe packets may aggravate congestion in a WLAN. To eliminate the redundant probe packets, we also extend the strategy of section IV-B, in which one representative MN sends a probe packet to the AP and all MNs including the representative MN measure RTT by capturing the probe and probe ACK packets. This method works as follows (see Fig. **??**).

Each MN first monitors all packets over a wireless link before sending a probe packet. If it finds a probe packet sent by another MN, it cancels sending a probe packet and measures RTT by using the probe packet sent by another MN. As each MN captures the header of all received packets, it can identify whether a captured packet is a probe packet or not by observing the frame length of the ICMP message (64 bytes).

Furthermore, an MN can also identify whether a probe packet is for request (ICMP Request) or for reply (ICMP Response) by observing the MAC address of the probe packet because all MNs connected to an AP can identify the MAC address of the AP. Therefore, if the *destination MAC address* of the captured packet is that of the AP, each MN can judge the packet as a *probe request* packet transmitted from another MN. On the other hand, if the *source MAC address* is an AP's one, then each MN judges the packet as a *probe reply* packet transmitted from the AP.

Fig. 11: Simulation Model 2

In Fig. **??**, $probeReq\_Time$ and $probeReply\_Time$ are the receiving time of the probe request transmitted from another MN and the probe reply transmitted from the AP, respectively. As every MN can identify whether a captured packet is a probe request or probe reply, it can calculate the RTT ($probeReq\_Time - probeReply\_Time$) properly. This method can eliminate the redundant probe packets because only one representative MN sends probe packets and all MNs measure the RTT by capturing existing probe packets over a wireless link.

If an MN that sends probe packets leaves a WLAN, one of the remaining MNs needs to start sending a probe packet in order to measure RTT. Here, we describe how an MN obtains the right to send probe packets. First, all MNs always examine the diffrence between the last receiving time of a probe packet ($ProbeLastTime$) and the current time ($CurrTime$). If the difference is greater than $Wait\_Interval$ time (twice probe packet sending interval: $500ms \times 2$), first, MNs with the lowest transmission rate in a WLAN try to send a probe packet. This is because a probe packet sent at the lowest transmission rate can be captured by almost all MNs in a WLAN due to the use of more robust modulation that inherently has a large transmission range. Note that the timing to send a probe packet among MNs with the same transmission rate is determined by $WaitingTime$, which is the random waiting time. That is, an MN with the smallest $WaitingTime$ can send a probe packet as the representative MN. Then, if other MNs with the lowest transmission rate captures a probe packet sent by the representative MN, they cancel sending a probe packet.

V. Simulation Experiment

In this section, we show the effectiveness of our proposed HO strategy through simulation experiments using Qualnet 4.0.1 [**?**]. We also employ our previous HO strategy [**?**] as a comparative method. In our study, we use MOS to assess the VoIP quality.

(a) Comparative Method (b) Proposed Method

Fig. 12: Variation of AP queue length and MOS

Fig. **??** and Table **??** show a simulation model and parameters, respectively. In the simulation scenario, 15 MNs are randomly located in a wireless area and move randomly at the speed of 1 m/s in two APs' coverage area. We then evaluate MOS value of MN and AP queue length. Fig. **??** shows results for comparative and proposed methods, respectively. From Fig. **??**, in the comparative method, we can see that the average of AP queue length is extremely high and MOS of MN does not satisfy the adequate VoIP quality (MOS$\geq$3.6) at all, i.e., Avg. MOS of 1.86. On the other hand, in Fig. **??**, our new proposed method can almost maintain adequate VoIP quality (Avg. 3.60) during the simulation time though including some degradations. Therefore, since our new proposed method can promptly and reliably detect not only the increase of AP queue length but also degradation of wireless link condition, MNs autonomously and properly execute HO in response to the change in the wireless network condition even under the congested WLAN environment.

VI. Conclusion

VoWLANs have many challenges because VoIP is a delay and packet loss sensitive application. In a congested WLAN, VoIP packets routed to MNs often experience a large queuing delay and buffer overflowed packet loss at the AP buffer. As a result, as VoIP quality toward MN degrades, an MN and a CN cannot continue conversation. To maintain VoIP quality during HO, we proposed an MN-centric HO decision strategy estimating AP queue length to detect the congestion at the AP and exploiting RTS frame retry of MN to detect the deterioration of wireless communication quality due to the movement of the MN. We first found that AP queue length has a potential to serve as an HO decision metric. However, since an MN cannot directly obtain AP queue length from an AP, we employed a probe packet mechanism in order to estimate AP queue length at the MN side. Furthermore, only the one representative MN exchanged probe packets with the AP for eliminating the redundant packets as much as possible. Simulation results showed that our proposed HO strategy can autonomously and promptly detect the wireless network condition in WLAN, i.e., wireless link condition and congestion state at APs, thereby maintaining adequate VoIP quality during HOs even under a congested WLAN environment.

Acknowledgment

This work was supported by the Kinki Mobile Radio Center Inc., the Telecommunications Advancement Foundation, and the Japan Society for the Promotion of Science, Grant-in-Aid for Young Scientists (B).

References

[1] C. Perkins (Ed.), "IP Mobility Support for IPv4," IETF RFC3344, Aug. 2002.

[2] S. J. Koh, et al., "Mobile SCTP for Transport Layer Mobility," draft-reigel-sjkoh-sctp-mobility-04.txt, Internet draft, IETF, Jun. 2004.

[3] John Fitzpatrick et al., "An Approach to Transport Layer Handover of VoIP over WLAN," Proc. of IEEE CCNC, Jan. 2006.

[4] K. Tsukamoto, et al., "Experimental Evaluation of Decision Criteria for WLAN handover: Signal Strength and Frame Retransmission," IEICE Trans. on Communications, Vol. E90-B, No. 12, pp. 3579-3590, Dec. 2007.

[5] H.Velayos and G.Karlsson, "Techniques to reduce the IEEE802.11b handoff time," Proc. of IEEE ICC, Vol. 7, pp. 3844-3848, Jun. 2004.

[6] S. Kashihara and Y. Oie, “Handover Management based on the number of data frame retransmissions for VoWLAN,” Elsevier Computer Communications, Vol. 30, no. 17, pp. 3257-3269, Nov. 2007.

[7] S. Kashihara, et al., “Service-oriented mobility management architecture for seamless handover in ubiquitous networks,” IEEE Wireless Communications, Vol. 14, No. 2, pp. 28-34, Apr. 2007.

[8] Y. Taenaka, et al., “Design and Implementation of Cross-layer Architecture for Seamless VoIP Handover,” Proc. of IEEE MHWMN, Oct. 2007.

[9] ITU-T:”G.107”, http://www.itu.int/rec/T-REC-G.107/en.

[10] Scalable Network Technologies, http://www.scalable-networks.com/