Call Setup Latency Analysis in SIP-Based Voice over WLANs

Sangheon Pack, *Member, IEEE*, and Hojin Lee, *Student Member, IEEE*

***Abstract*— In this letter, we consider session initiation protocol (SIP)-based voice over wireless local area networks (VoWLANs). We derive the analytical expression for the average call setup latency, and we observe the impact of the number of contending mobile nodes in a WLAN. Extensive simulation results are given to validate the analytical results.**

***Index Terms*— Voice over WLANs, SIP, call setup latency.**

I. Introduction

RECENTLY, IEEE 802.11 wireless local area network (WLAN) systems [1] are widely deployed in hot spot areas such as convention centers, airports, campus, etc. Meanwhile, a lot of efforts have been conducted with the focus of quality-of-service (QoS) support in real-time applications over WLANs. For instance, IEEE 802.11e has been standardized and numerous schemes (e.g., scheduling, admission control, etc.) have been proposed in the literature. As a sequence, it is expected that voice over IP (VoIP) in WLANs, i.e., voice over WLANs (VoWLANs), will be a promising application in the near future. However, due to the fundamental limitations of distributed channel access, QoS provisioning for VoWLANs is still a challenging issue. QoS performance for VoWLANs can be characterized at the bit-level (i.e., bit error rate (BER)), packet-level (i.e., packet loss rate), and call-level (i.e., call dropping/blocking probability). In this work, we focus on a call-level performance metric, *call setup latency*, which is defined as the elapsed time until a call request is correctly processed and the call is established. The call setup latency is significantly affected by the signaling protocols used (i.e., H. 323 or Session Initiation Protocol (SIP) [2]) and network conditions.

So far, several studies for analyzing the call setup latency in wireless networks have been reported. Das *et al.* [3] analyze the call setup latency in a H.323-based VoIP system. On the other hand, Fathi *et al.* [4] consider a SIP-based VoIP system and evaluate the call setup latency in wireless fading channels. Curcio *et al.* [5] evaluate the SIP call setup latency using a 3G network emulator. All of these works consider cellular systems where a dedicated channel is assigned to a mobile user. On the other hand, Banerjee *et al.* [6] introduce a SIP call setup latency in WLANs. However, they do not consider the effect of the channel access method employed in WLANs, which is a key factor to determine the call setup latency. Hesselman *et al.* [7] present experimental results for call setup latency in WLAN testbeds; however, no analytical model is developed.

Manuscript received July 23, 2007. The associate editor coordinating the review of this letter and approving it for publication was F. Granelli. This work was supported in part by the IT R&D program of MIC/IITA [2007-F-038-01, Fundamental Technologies for the Future Internet], and in part by the second stage of the BK 21 project.

S. Pack is with the School of Electrical Engineering, Korea University, Seoul, Korea (e-mail: shpack@korea.ac.kr).

H. Lee is with the School of Computer Science and Engineering, Seoul National University, Seoul, Korea (e-mail: lumiere@mmlab.snu.ac.kr).

Digital Object Identifier 10.1109/LCOMM.2008.071230.

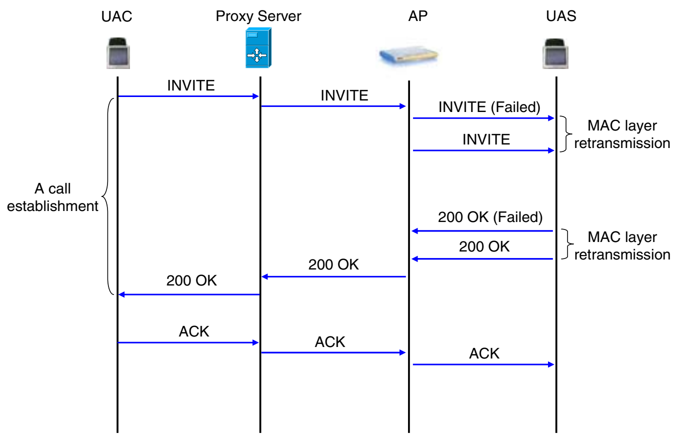


Fig. 1. VoIP call setup procedure.

In this letter, we propose an analytical model for the call setup latency in a SIP-based VoWLAN system. The analytical model considers a distributed channel access method (i.e., distributed coordination function (DCF)) in VoWLANs and demonstrates the impact of the channel access method on the call setup latency. We validate the analytical results through extensive simulations. The remainder of this letter is organized as follows. In Section II, a SIP-based VoIP system and WLAN channel model is described. The call setup latency is analyzed in Section III and simulation results are given in Section IV followed by the concluding remarks in Section V.

II. System Description

Figure 1 shows a call setup procedure in a SIP-based VoWLAN system. To establish a VoIP call, a user agent client (UAC) first sends an INVITE request message to a local SIP server. The local SIP server can be a proxy or redirect server: a proxy server relays the received SIP message to another SIP server or a user agent; on the other hand, a redirect server responds with the SIP message with location information. Throughout this letter, our description is based on the proxy server. After routing over SIP servers, the INVITE message arrives at a user agent server (UAS) via an access point (AP) in WLANs, and the UAS then responds with a status code[1]. For instance, if the INVITE message is successfully processed by a UAS, the UAS sends a response message with a status code of 200. After receiving the 200 OK response, the UAC sends

[1]Only final responses have impact on the session setup latency [4], and therefore no provisional responses are not considered in this letter.

an ACK request message to notify that the 200 OK response is corrected received. Since the ACK message is used only for confirmation, we define the call setup latency as the interval from the time instant the INVITE message is sent to the time instant the 200 OK message is received.

Since the delay in wired links, t_{wired}, is fairly stable, it is assumed that the call setup procedure is delayed only due to wireless link delay. For WLAN, N saturated MNs (i.e., each MN always has a packet to send) are assumed to share the WLAN channel and the transmission failures are only due to collisions. We consider the IEEE 802.11 distributed coordination function (DCF) [1] with a basic access mode for media access control because request-to-send (RTS)/clear-to-send (CTS) is not quite effective in infrastructured WLANs and it is disabled in most products available in the current market.

We assume that user datagram protocol (UDP) is used as a transport protocol for SIP messages. Since UDP does not support reliable transmissions, the UAC performs end-to-end (E2E) retransmissions based on an exponential backoff algorithm [2]. The initial backoff timer T_{Init} is typically set to 500 msec. After the UAC sends an INVITE message, the UAC retransmits the INVITE message at most for 32 seconds or until it receives a response. As a result, the UAC can retransmit an INVITE message at most 6 times, that is, the number of total transmissions is 7. Similarly, the UAS retransmits a 200 OK message at most for 32 seconds or until it receives an ACK message. However, unlike the UAC, the UAS keeps the maximum backoff timer at 4 seconds. Therefore, retransmissions of a 200 OK message occur at most 10 times, i.e., at 0.5, 1.5, 3.5, 7.5, 11.5, 15.5, 19.5, 23.5, 27.5, and 31.5 seconds after the transmission of the first 200 OK message. Independently from E2E retransmissions, MAC layer retransmissions are performed in IEEE 802.11 WLANs. That is, when a packet[2] transmission fails due to collision, the packet is retransmitted until it is successfully delivered or up to m times.

III. VoIP Call Setup Latency Analysis

Let ε be the probability that a SIP message is lost over a WLAN link after m MAC layer retransmissions. Since a packet transmission at each backoff stage is independent, $\varepsilon = p^{m+1}$ where p is the collision probability for a packet transmission. p can be obtained from [9] by an iterative method. Let R be the total number of E2E transmissions. Then, R is 7 and 11 for the INVITE and 200 OK messages, respectively. By [1], the number of backoff slots in the jth stage is uniformly selected in $[0, W_j - 1]$ where $W_j = 2^j W_0$ and W_0 is the minimum contention window size. Therefore, the probability that the number of chosen backoff slots is k at the jth stage is $1/(W_j - 1)$.

Let $\theta_X(i, j, k)$ be the probability that a transmission of a SIP message X is successful at the ith E2E transmission and the jth backoff, and the chosen backoff slot length is k ($1 \leq i \leq R,\ \ 0 \leq j \leq m,\ \ 1 \leq k \leq W_j - 1$). Then, $\theta_X(i, j, k)$ can be obtained from

$$\theta_X(i, j, k) = \frac{1}{1 - \varepsilon^R} \frac{\varepsilon^{i-1} p^j (1 - p)}{W_j - 1}, \tag{1}$$

where $1/(1 - \varepsilon^R)$ is the normalized factor that is required because we consider only successful transmissions.

Let $l_X(i, j, k)$ be the transmission time when the counters for E2E transmission and MAC layer transmission for a transmission of a SIP message X are i and j, respectively, and the randomly chosen contention window size is k ($1 \leq i \leq R,\ \ 0 \leq j \leq m,\ \ 1 \leq k \leq W_j - 1$). Then, $l_X(i, j, k)$ is given by (2), where $Tr(n)$ is the value of the E2E retransmission timer when a packet is successfully transmitted at the nth transmission. $Tr(n)$ for the INVITE and 200 OK messages are given by $2^{n-2} T_{Init}$ and $\min\{2^{n-2} T_{Init}, 4\}$, respectively, where T_{Init} is the initial E2E timeout value. In (2), $E[slot]$ is the average slot length defined as the time interval between two consecutive backoff counter decrements in [9], and it can be computed as

$$E[slot] = (1 - P_{tr})\sigma + P_{tr} P_S T_S + P_{tr}(1 - P_S) T_C,$$

where $P_{tr} = 1 - (1 - \tau)^n$ and $P_S = \frac{n\tau(1-\tau)^{n-1}}{1-(1-\tau)^n}$. T_S and T_C are the time durations that the channel is sensed busy during a successful frame transmission and a collision, respectively. T_S and T_C are given by

$$T_S = DIFS + H + P + \delta + SIFS + ACK + \delta \quad \text{and}$$
$$T_C = DIFS + H + P + SIFS + ACK,$$

where δ is the propagation delay. $DIFS$ and $SIFS$ represent DCF inter frame space and small inter frame space, respectively. H, P, ACK are the transmission time for the header, payload, and ACK frame, respectively.

Then, the average transmission time of a SIP message X is

$$L(X) = \sum_{i=1}^{R} \sum_{j=0}^{m} \sum_{k=1}^{W_j - 1} \theta_X(i, j, k) \cdot l_X(i, j, k). \tag{3}$$

The average call setup latency can be represented as

$$S = L(INVITE) + L(200OK), \tag{4}$$

where $L(INVITE)$ and $L(200OK)$ are delivery latency for the INVITE and 200 OK messages, respectively.

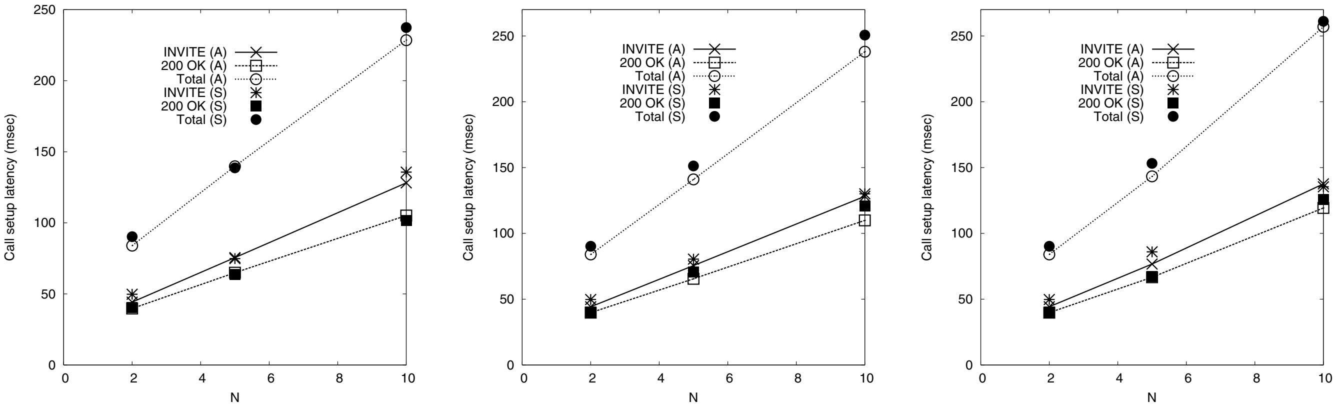
IV. Simulation Results

To validate the analytical model, we have carried out simulations using *ns-2* simulator [10]. In simulations, the wired link delay is fixed at 20 msec. The retransmission limit is set to 3 and the data rate is 1 Mbps. Note that $m = 3$ is less than the value in the IEEE 802.11 specification. This is because we focus on time-sensitive VoIP applications and thus a larger m is not appropriate due to a long end-to-end delay and delay jitter even though it can reduce the packet loss rate. The lengths (including UDP and IP headers) of INVITE and 200 OK messages are 788 and 488 bytes, respectively.

Figure 2 shows the call setup latency as a function of N under different T_{Init}. It can be seen that the call setup latency increases with the increase in N. Also, the latency

[2]Since the maximum protocol data unit in IEEE 802.11 MAC layer is sufficiently large, we do not consider link layer fragmentation and therefore the term *packet* is used for both a protocol data unit (PDU) both in the data link layer and in the transport layer.

$$l_X(i,j,k) = \begin{cases} \sum_{n=2}^{i} Tr(n) + t_{wired} + \sum_{m=0}^{j-1} \frac{W_m - 1}{2} E[slot] + kE[slot], & i > 1, \quad j \geq 1 \\ \sum_{n=2}^{i} Tr(n) + t_{wired} + \frac{W_0 - 1}{2} E[slot] + kE[slot], & i > 1, \quad j = 0 \\ t_{wired} + \sum_{m=0}^{j-1} \frac{W_m - 1}{2} E[slot] + kE[slot], & i = 1, \quad j \geq 1 \\ t_{wired} + \frac{W_0 - 1}{2} E[slot] + kE[slot], & i = 1, \quad j = 0 \end{cases} . \quad (2)$$



(a) T_{Init} = 500 msec (b) T_{Init} = 1000 msec (c) T_{Init} = 2000 msec

Fig. 2. Call setup latency vs. N (A: Analysis, S: Simulation).

for the INVITE message is larger than that for the 200 OK message because the INVITE message is longer than the 200 OK message. From Figures 2, it can be found that some discrepancy between simulation and analytical results especially when there are many contending mobile nodes (i.e., N is large). This discrepancy can be explained by two reasons: 1) the AP queueing delay is not considered in our model. However, the AP queueing delay cannot be neglected when N is large; 2) unnecessary retransmissions can occur for a large N if a packet is retransmitted before the original packet is completely served at the AP queue due to long channel contention time. Since the unnecessary retransmission is not taken account in our model, larger discrepancy can be observed in the situation. Actually, the AP queueing delay can be mitigated by increasing the initial E2E retransmission timer, T_{Init}. Consequently, as shown in Figure 2, the discrepancy between simulation and analytical results decreases as T_{Init} increases. It can be also seen that the average call setup latency also increases with the increase of T_{Init}. Therefore, an adaptive setting for T_{Init} depending on network conditions needs to be devised to optimize the call setup latency.

V. Conclusion

In this letter, we have analyzed the call setup latency in a SIP-based VoWLAN system. Extensive simulations have been carried out to validate the analytical model. Simulation and analytical results demonstrate that the call setup latency is sensitive to the number of mobile nodes in a WLAN, and T_{Init} should be carefully chosen to avoid unnecessary end-to-end retransmissions while minimizing the call setup latency. In our future work, we will exploit the call setup blocking probability, which is defined as the probability the call setup latency exceeds an acceptable bound, and extend the analytical model for wireless mesh networks with multi-hop wireless links.

References

[1] IEEE 802.11b WG, Part 11, "Wireless LAN medium access control (MAC) and physical layer (PHY) specification: high-speed physical layer extension in the 2.4 GHz Band," IEEE, Sept. 1999.

[2] J. Rosenberg *et al.* "SIP: session initiation protocol," *IETF RFC 3261*, June 2002.

[3] S. Das, E. Lee, K. Basu, and S. Sen, "Performance optimization of VoIP calls over wireless links using H.323 protocol," *IEEE Trans. Computers*, vol. 52, no. 6, pp. 742-752, June 2003.

[4] H. Fathi, S. Chakraborty, and R. Prasad, "On SIP session setup delay for VoIP services over correlated fading channels," *IEEE Trans. Veh. Technol.*, vol. 55, no. 1, pp. 286-295, Jan. 2006.

[5] I. Curcio and M. Lundan, "SIP call setup delay in 3G networks," in *Proc. IEEE ISCC 2002*, pp. 835- 840, July 2002.

[6] N. Banerjee, W. Wu, K. Basu, and S. Das, "Analysis of SIP-based mobility management in 4G wireless networks," *Computer Commun.*, vol. 27, no. 8, pp. 697-707, 2004.

[7] C. Hesselman, H. Eertink, I. Widya, and E. Huizer, "Measurements of SIP signaling over 802.11b links," in *Proc. ACM WMASH 2005*, pp. 74-83, Sept. 2005.

[8] T. Eyers and H. Schulzrinne, "Predicting Internet telephony call setup delay," in *Proc. IP Telephony Workshop 2000*, pp. 107-126, Apr. 2000.

[9] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement," in *Proc. IEEE INFOCOM 2002*, pp. 599- 607, June 2002.

[10] ns-2, "The Network Simulator ns-2", http://www.isi.edu/nsnam/ns, 2005.