Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Networks xxx (2005) xxx–xxx
Computer Networks
www.elsevier.com/locate/comnet

On performance analysis of challenge/response based authentication in wireless networks ☆

Wei Liang, Wenye Wang *

Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695-7911, United States

Received 17 September 2004; accepted 4 October 2004

Responsible Editor: G. Morabito

Abstract

The emergence of public access wireless networks enables ubiquitous Internet services, whereas inducing more challenges of security due to open mediums. As one of the most widely used security mechanisms, authentication is to provide secure communications by preventing unauthorized usage and negotiating credentials for verification. Meanwhile, it generates heavy overhead and delay to communications, further deteriorating overall system performance. Therefore, it is very important to have an in-depth understanding of the relationship between the security and quality of service (QoS) through the authentication in wireless networks. In this paper, we analyze the impact of authentication on the security and QoS quantitatively. First, a system model based on challenge/response authentication mechanism is introduced, which is wide applied in various mobile environments. Then, the concept of security levels is proposed to describe the protection of communications with regard to the nature of security, i.e., information secrecy, data integrity, and resource availability. Third, traffic and mobility patterns are taken into account for quantitative analysis of QoS. Finally, we provide numerical results to demonstrate the impact of security levels, mobility and traffic patterns on overall system performance in terms of authentication cost, delay, and call dropping probability.

Keywords: Wireless networks; Challenge/response authentication; Security association; Performance analysis

1. Introduction

With the deployment of public access wireless networks, the demand for ubiquitous Internet services is dramatically increased, whereas inducing more challenges of security due to open mediums [1]. In order to provide secure services over wireless

☆ This work was supported in part by the National Science Foundation (NSF) under award ANI-0322893 and in part by the Center for Advanced Computing and Communication (CACC) 03-06 and 04-08.

* Corresponding author. Tel.: +1 919 513 2549; fax: +1 919 515 5523.
E-mail address: wwang@ncsu.edu (W. Wang).

1389-1286/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2004.10.016

networks, security mechanisms such as authentication and encryption are deployed at the expense of quality of service (QoS) because of the implementation overhead.

As one of the most widely used security mechanisms, *authentication* is a process to identify a mobile user (MU), authorize resources to the MU, and negotiate secret credentials for protecting communications [2]. In the authentication, an MU will submit credentials like certificates and challenge/response values [3–8], which will be verified with a security association (SA), a description on keys and encryption algorithms. With the authentication, network resources are protected by only allowing legitimate users to obtain services. The information secrecy and data integrity are also guaranteed because session keys may be generated during the authentication process for data encryption and message authentication. Thus, the network security in terms of protection for network resources, information secrecy, and data integrity is affected greatly by the authentication service.

In addition, an authentication service also has significant effects on the QoS. When certificates are used for authentication, it involves with the application of public/private-key based authentication mechanism, in which more time and power are consumed due to the computation complexity of encryption and decryption of data [9]. Thus, in order to achieve efficient authentication, challenge/response authentication mechanism based on secret keys is widely used in wireless networks [10–12]. However, the credentials of the MU are encrypted and transmitted hop-by-hop for remote verification among authentication servers in challenge/response authentication. This remote transmission and encryption/decryption of credentials increase the overhead of communications, thus influence many QoS parameters such as authentication cost, delay, and call dropping probability due to extended waiting time. Therefore, the tradeoff between security service and system performance should be concerned in different scenarios, because users have different preferences on security and performance from time to time.

Furthermore, the impact of authentication on QoS parameters is far more sophisticated for different mobility and traffic patterns, since the authentication requests are generated when an MU either requests resources, or crosses boundaries of subnets with on-going communications. Thus, the authentication based on different mobility and traffic patterns may greatly impact QoS parameters such as aggregated authentication cost in a network, because the cost needs to be calculated by adding up the costs in all of the authentication requests.

In order to improve the security and efficiency during the authentication, many authentication schemes are proposed, focusing on the design of lightweight and secure authentication protocols [2,5–7,10–20]. However, none of these work provide quantitative analysis on security and system performance, simultaneously, and nor do they model the relationship between security levels and system performance analytically, although some of them evaluate the system performance for certain security policies through simulations [15,20]. Moreover, mobility and traffic patterns are not considered, which are important features in wireless networks. Therefore, new authentication solutions may not be fully adapted to mobile environments with the concerns of security, mobility and traffic patterns.

In this paper, we analyze the effect of challenge/response authentication on security and system performance quantitatively. First, we propose a system model, which is highly consistent with many wireless networks such as Mobile IP and wireless local area network (WLAN). This consistency guarantees that our analysis is applicable in realistic mobile environments. Second, we classify the security levels with regard to the nature of security, i.e., information secrecy, data integrity, and resource availability, and study the effects of authentication on QoS at different security levels. The QoS parameters that we investigate in this paper include authentication cost, delay, and call dropping probability, all of which are considered in combination with *mobility and traffic patterns*.

Our earlier work [21,22] presented a framework of performance analysis, focusing on authentication delay and call dropping probabilities. In this paper, we not only provide the analysis details

on authentication delay and call dropping probabilities, but also present analysis on authentication cost in terms of signaling cost and cryptographic load. Thus, by coupling the security and system performance, this paper provides a foundation for future design and analysis of the authentication in wireless networks, which may further cultivate the optimization of the authentication efficiency in mobile wireless networks.

The rest of our paper is organized as follows. We describe the authentication impact on security and system performance in Section 2 based on challenge/response authentication. In Section 3, a system model and corresponding metrics are defined for our analysis. We analyze these metrics at different security levels with the concern of mobility and traffic patterns in Section 4. Then, numerical results of our analysis on authentication cost, delay, and call dropping probabilities are presented in Section 5. Finally, we draw conclusions in Section 6.

2. Effect of authentication on security and QoS

In this section, we introduce challenge/response authentication and describe the effects of authentication on security and QoS with challenge/response authentication mechanism.

2.1. Overview of challenge/response authentication

The authentication in wireless networks is defined as a process in which an MU needs to send out the secret credentials for verification and negotiate credentials for communications.

In a challenge/response-based authentication, a user is identified with a shared security association (SA), which is a trust relationship with many parameters such as keys and algorithms for secure services [23], by an authentication server. During the process, the server sends a *challenge value*, a random number, to the user for encryption, and verifies the returned value, called *response value*, with decryption. In a foreign network, a visiting MU sends an authentication request to an access point (AP). The AP relays the request to a local authentication server (LAS), which only takes charge of authentication for visiting MUs from foreign networks. If the LAS has no information to verify the MU, it contacts the home authentication server (HAS) of the MU through an authentication architecture. An HAS is an authentication server to identify the MUs who subscribe the service in its network. And, an authentication architecture is composed of many authentication servers that share SAs with the LAS and HAS. If the request is an inter-domain authentication request, the HAS sends a registration request to the MU's home agent (HA), which maintains the current location of the MU, to update the MU's location.

Throughout this paper, we assume that an MU is roaming into a foreign network domain. Then, the challenge/response authentication for an MU in a foreign network domain can be categorized into three types: intra-domain handoff authentication; session authentication; and inter-domain handoff authentication, with the signaling diagrams shown in Fig. 1.

Intra-domain handoff authentication: When an MU crosses the boundary of subnets in the foreign network domain with an on-going service, an intra-domain handoff authentication is initiated. Since there is an on-going communication session between the MU and an AP, one session SA exists between the MU and the LAS in the visiting network domain. Therefore, it is unnecessary to contact the HAS of the MU for authentication. In the case shown in Fig. 1A, the LAS, which receives the authentication request from an MU, sends a challenge value to the MU. The MU encrypts the challenge value using shared SA with the LAS and replies the response value to the LAS. After decrypting the replied value and comparing it with original challenge value, the LAS can authenticate the MU when the decrypted value matches the original challenge value.

Session authentication: When an MU starts a communication session in a subnet of a foreign network, a session authentication is initiated. Since there is no on-going communication session between the MU and the AP, session SA does not exist between the MU and the LAS, and it is necessary to contact the HAS of the MU for authentication. In the case shown in Fig. 1B, when

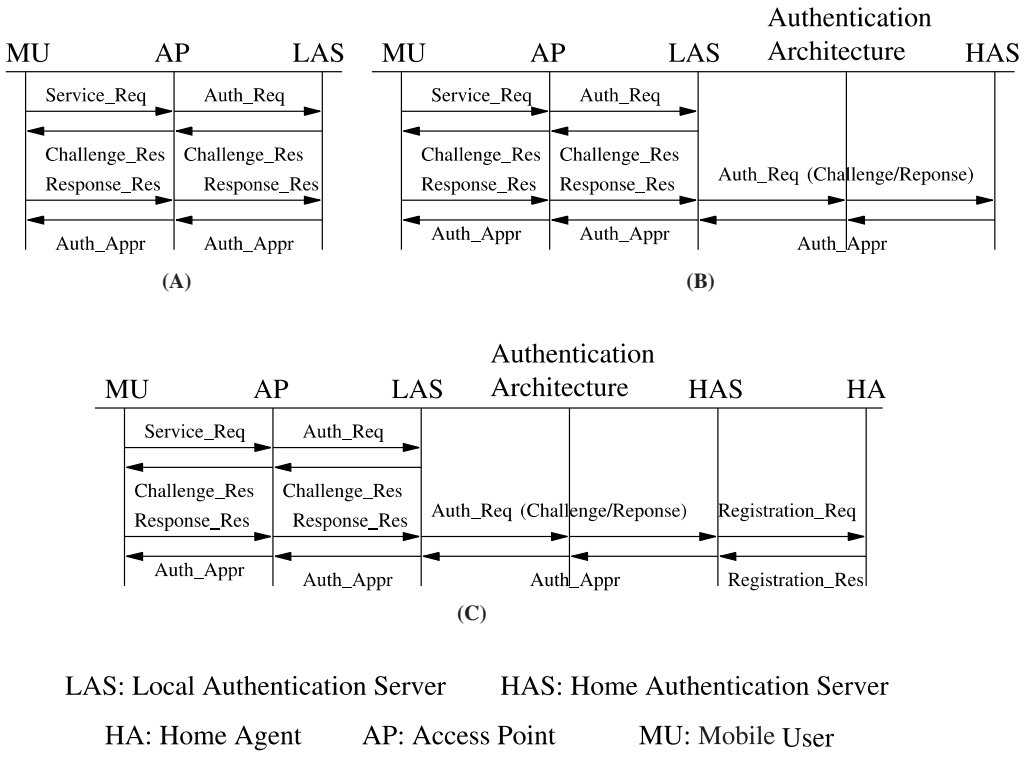


Fig. 1. Challenge/response authentication in public wireless access networks: (A) Intra-domain handoff authentication; (B) session authentication and (C) inter-domain handoff authentication.

the LAS receives the authentication request forwarded from the AP, it sends a challenge value to the MU. The MU encrypts the challenge value with the SA shared with the HAS, and replies the response value to the LAS. The LAS must forward the challenge and response values to the HAS of the MU for verification because the LAS does not share an SA with the visiting MU, and cannot decrypt the response value without the SA. After authentication at the HAS, the secret credentials such as keys to protect the communication are generated and sent to the LAS.

Inter-domain handoff authentication. When an MU is crossing the boundaries of different foreign network domains with an on-going service, an inter-domain handoff authentication occurs. Without an existed SA between the MU and the LAS, the signaling diagram shown in Fig. 1C is similar with that in the case of session authentication, except that the MU needs registration to its HA via the HAS *because we assume that the MU needs registration when it is crossing the boundaries of different network domains*.

2.2. Effect of authentication on security

Security services are to provide information secrecy, data integrity, and resource availability for users. Information secrecy means to prevent the improper disclosure of information in the communication, while data integrity is to prevent improper modification of data and resource availability is considered to preventing improper denial of services [23].

In order to provide security services in wireless networks, challenge/response-based authentication adopts several techniques to meet the requirements. First, the challenge/response authentication enables the MU to share an SA with its HAS. The SA is unique and secret to other users. Therefore,

the identification of the MU is unique, which can prevent unauthorized MUs from accessing the network resource. Thus, the resource availability for authorized users can be guaranteed. Second, new secret credentials such as session keys are generated and sent to communication partners during authentication. The distributed secret credentials are used to encrypt the data of communication and provide message authentication code for data integrity check. Therefore, the authentication mechanism plays a key role to protect the information secrecy and data integrity because new secret credentials such as session keys are generated and transferred during this period.

2.3. Effect of authentication on QoS metrics

Besides the effect on the security, authentication also influences QoS metrics, such as authentication delay, cost, call dropping probability and throughput of communications due to authentication overhead.

The authentication delay is defined as the time from when the MU sends out an authentication request to when the MU receives the authentication reply. During this authentication delay, no data for on-going service can be transmitted, which may interrupt the connection. Therefore, the call dropping probability may be increased because of the extended authentication delay.

The authentication cost is defined as the signaling cost and processing load for cryptographic techniques. In a challenge/response authentication, the challenge/response values need to be transmitted back to the HAS of the MU for verification when the LAS has no SA shared with the roaming MU. Then, the signaling messages are transmitted between different LASs. The total number of signaling messages from the LAS to the HAS of the MU can be large if the authentication distance between them is long. Furthermore, the signaling messages need to be encrypted and decrypted hop-by-hop for protection due to lack of direct trust relationship between the LAS and the HAS. These multiple encryption and decryption increase the processing load of the networks. Moreover, the mobility and traffic patterns of MUs make the authentication happen frequently in different scenarios because the authentication is initiated when an MU starts a communication session or crosses boundaries of subnets with an on-going service, which may cause an imbalance distribution of authentication cost.

Compared to the effects of authentication on delay and cost, the throughput is affected by the authentication throughout the whole communication service. The throughput is defined as the the effective data transmitted per unit time. It can be greatly decreased in authentication because of several reasons. First, the authentication delay may cause a pause for data transmission, which decreases the throughput. Second, the key size and complex algorithms used in authentication affect the processing time of authentication messages, and the attachment of message authentication code for data integrity check will affect the payload of messages.

In summary, authentication in wireless networks has great effects on both security and QoS such as authentication delay, cost, and throughput. In order to improve the security and performance of wireless networks, it is necessary to analyze the authentication effects on both security and QoS metrics by taking into account mobility and traffic patterns. To analyze these effects in combination with actual mobility and traffic patterns, we propose a system model with assumptions and definitions of performance metrics in the next section.

3. System model and metrics

In this section, we introduce a system model to analyze the impact of challenge/response authentication in wireless networks. We consider the security and system performance, in which the security is defined with regard to security levels, and the system performance are evaluated with authentication cost, delay and call dropping probabilities.

3.1. System model

We consider a generic system model for wireless networks from two aspects. One is to describe the

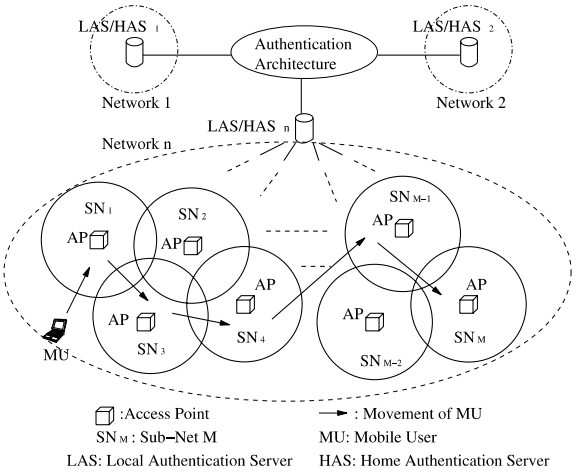


Fig. 2. System model of authentication in wireless networks.

authentication interaction between autonomous wireless networks; the other is to illustrate the authentication within a wireless network.

The system model to describe the authentication interaction between inter-connected wireless networks is shown in Fig. 2. In this model, there are a number of n autonomous wireless networks. Each network domain has an LAS and an HAS, which are central authentication servers in a network domain. However, an LAS only takes charge of authentication for visiting MUs, while an HAS is only responsible for the authentication of the MUs that subscribe services in current network domain. The trust relationships between these LASs and HASs are maintained through an *authentication architecture*, which is an infrastructure composed of many proxy authentication servers and designed to securely deliver the authentication messages between authentication servers [18]. It is assumed that the LAS and HAS are integrated together, and the authentication architecture shares an SA with the LAS/HAS of a network domain.

We further assume that a network domain is composed of M subnets of equal size, and each subnet is controlled by an access point (AP). Here, an AP is a function unit that can transmit data for MUs with established SAs, which are the secure trust relationships introduced in Section 2.1. An LAS controls the authentication in the network domain with M subnets in it, and shares SAs with M APs. The LAS also shares an SA with the authentication architecture that is trusted by and connected with the HAS of roaming users.

The generic system model in our paper is consistent with many practical wireless networks such as the authentication, authorization, and accounting (AAA) architecture in Mobile IP networks and wireless local area networks (WLAN) [18]. In order to evaluate the performance of authentication, we need to further describe specific conditions such as authentication mechanism, mobility and traffic patterns with which the impact of authentication can be evaluated clearly.

Scenario: Assume that the challenge/response authentication is implemented on the generic system model with signaling diagrams shown in Fig. 1. In this paper, we focus on the scenario that *an MU is roaming into foreign network domains*. Then, the intra-domain handoff authentication, session authentication, and inter-domain handoff authentication in foreign networks are illustrated in Fig. 1A–C, respectively.

Mobility pattern: The mobility pattern of an MU in our paper is represented by the residence time of the MU in one subnet, denoted as T_r. We assume that T_r is a random variable and the probability density function (PDF) of T_r, denoted as $f_{T_r}(t)$, is Gamma distribution with mean $1/\mu_r$ and variance V [24]. Then, the Laplace transform of $f_{T_r}(t)$, $F_r(s)$, is

$$F_r(s) = \left(\frac{\mu_r\gamma}{s+\mu_r\gamma}\right)^{\gamma}, \quad \text{where } \gamma = \frac{1}{V\mu_r^2}. \tag{1}$$

Furthermore, if the number of subnets passed by an MU is assumed to be uniformly distributed between $[1, M]$, the PDF of the residence time in a network domain, denoted as $f_{T_M}(t)$, can be expressed with a Laplace transform $F_M(s)$ as follows [25]:

$$F_M(s) = \frac{1}{M}\left(\frac{\mu_r\gamma}{s+\mu_r\gamma}\right)^{\gamma}\frac{1-\left(\frac{\mu_r\gamma}{s+\mu_r\gamma}\right)^{\gamma M}}{1-\left(\frac{\mu_r\gamma}{s+\mu_r\gamma}\right)^{\gamma}}. \tag{2}$$

Then, the mean value of residence time in this network domain, denoted as $\overline{T}_M$, can be expressed as

$$\overline{T}_M = -\frac{\partial F_M(s)}{\partial s}|_{s=0} = \frac{M+1}{2\mu_r}. \quad (3)$$

Traffic pattern: In this paper, we use call arrival rate and call duration time to indicate traffic patterns. First, we assume that the call arrival rate of an MU, which includes the incoming calls and outgoing calls, is a Poisson process with average rate λ_u, then the PDF of the call inter-arrival time, denoted as $f_{T_A}(t)$, can be determined by

$$f_{T_A}(t) = \lambda_u e^{-\lambda_u t}. \quad (4)$$

Moreover, we assume that a call duration time, denoted as T_D, has an exponential distribution with mean value $1/\eta$. Then, the PDF of call duration time, denoted as $f_{T_D}(t)$ can be written as

$$f_{T_D}(t) = \eta e^{-\eta t}. \quad (5)$$

Based on these assumptions on the mobility and traffic patterns of the MU, we evaluate the security and QoS metrics during authentication.

3.2. Performance metrics

We categorize the performance metrics into security and QoS parameters. The security parameter is represented by security levels, at which different levels of protection are provided. Meanwhile, we consider authentication cost, delay and call dropping probability as the system performance for evaluation.

3.2.1. Security levels

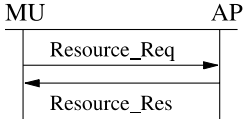
There is much quantitative analysis of QoS in networks [26,27], whereas less analysis of security exists. This gap between the QoS and security analysis demands quantization of security for the engineering research. Therefore, the concept of *security level* becomes widely used for security evaluation [28–30]. The classification of security levels in these papers is either based on the information sensitivity, or based on the key length. In the classification with the information sensitivity, if a group of users are allowed to access more sensitive data, and the data in this group is prohibited to expose to other groups, then, the security level of this group is the highest. In the classification with respect to key length, if an encryption/decryption process uses a longer key than other processes, this process is assigned a higher security level. As we can see, however, all of them do not consider the nature of security, i.e., data integrity, secrecy, and availability. Therefore, we argue that the nature of security should be involved to classify the security levels.

In our paper, the *security level* is to indicate the level of protection provided by the authentication for quantitative analysis of security. The classification of security levels is shown in Table 1 according to the security functions described in Section 2.2, i.e., protection for integrity, secrecy and resource availability. Because of different actions in challenge/response authentication, the protection of data integrity, secrecy, and availability are varied at different security levels.

- *Security level 1*: Any MUs can send data through an AP without authentication. The signaling diagram at this security level is shown in Fig 3. When an MU needs services at security level 1, it sends out a resource request to the AP. The AP checks the resources for this request. In intra-domain and session authentication with signaling diagram shown in Fig. 3, if the resource for this service is available, the

Table 1
Security level classification

Security level i	Security service			
	Integrity	Secrecy	Confidentiality	Availability protection
1	No	No	No	No
2	No	No	Low	Low
3	No	No	Medium	Medium
4	Yes	Yes	High	High

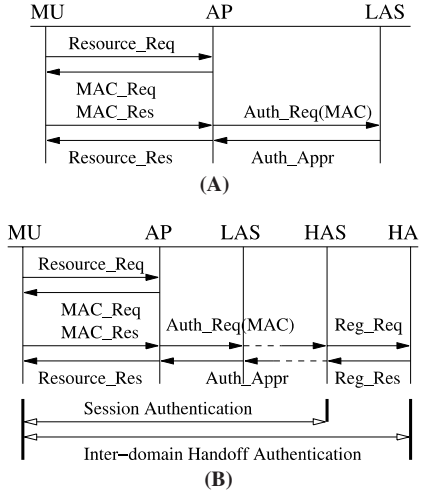


MU: Mobile User AP: Access Point

Fig. 3. Intra-domain handoff and session authentication at security level 1.

resource approval is replied to the MU to authorize the service. The signaling diagram for inter-domain handoff authentication is very similar to Fig. 1C. The difference is that when the LAS receives the request, the LAS sends registration messages to the HA and the HAS of the MU through the authentication architecture, instead of replying a challenge value to the MU. After the registration, the service is authorized to the MU. At security level 1, because no cryptographic techniques are applied, the integrity, secrecy, and resource availability cannot be protected.

- *Security level 2*: Authentication is implemented through a pre-defined list of Media access control (MAC) addresses and no keys are generated for the subsequent communication. In this case, when an MU needs resource in foreign networks, it sends a request to the local AP, which, in turn, requests the MAC address of the MU and relays the MAC address to the LAS or HAS for verification, as shown in Fig. 4. If the received MAC address matches one entry in the list, the MU is authenticated. For intra-domain handoff authentication, the LAS has the session SA of the MU, thus, the MU can be verified at the LAS. For intra-domain handoff authentication, the LAS has the session SA of the MU, thus, the MU can be verified at the LAS. For inter-domain and session authentication, the MU needs to be authenticated at the HAS because there is no SA between the MU and the LAS. In particular, registration needs to be implemented during inter-domain authentication. At security level 2, there is no protection available for data integrity and secrecy because no keys and algorithms are distributed to the MU for the communication. But the network resource is slightly protected by identifying the MAC address although the MAC address can be easily forged.
- *Security level 3*: Authentication is implemented with credentials encrypted with a shared SA, and no keys are generated for subsequent communications. In this case, the SA between the MU and its HAS is used for inter-domain handoff and session authentication with the signaling diagram shown in Fig. 5. Compared to the signaling process at security level 2, the signaling process at security level 3 is almost the same. The difference is that a pair of values, i.e., challenge/response values, are used to authenticate the MU instead of the MAC address. The challenge value is a random value generated and sent to the MU from the LAS. The MU encrypts the challenge value with corresponding SA. In the intra-domain handoff authentication, the MU encrypts the challenge value with the SA shared with the LAS during communication session, and replies the result, a response value to the LAS. The LAS can verify the challenge value by decrypting the response value with the same SA. However, in inter-domain handoff and session authentication, there is no SA between the MU and the LAS at this moment. The MU must be authenticated by the HAS. After decrypting and verifying the response value at the HAS of the MU, the authentication approval is sent back to the LAS for authorizing the resource to the MU. Specially, the registration process is required during the inter-domain authentication. At security level 3, the



LAS: Local Authentication Server MU: Mobile User
HAS: Home Authentication Server AP: Access Point
HA: Home Agent

Fig. 4. Signaling diagram at security level 2: (A) intra-domain handoff and (B) session and inter-domain handoff authentication.

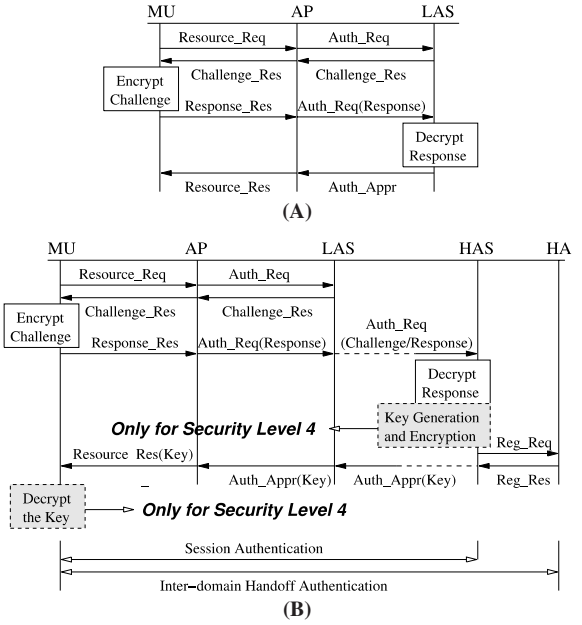


Fig. 5. Signaling diagram at security levels 3 and 4: (A) intra-domain handoff and (B) session and inter-domain handoff authentication.

network resources can be protected by only allowing the access of legitimate users. However, since the data transmission is not protected with encryption after authentication, the integrity and secrecy are not guaranteed. Furthermore, the network resource may be compromised due to the lack of data integrity and secrecy.

- *Security level 4*: Authentication is implemented with shared SA, and keys are generated for data encryption and message integrity check. The signaling diagram at this security level is shown in Fig. 5, which is similar with that at security level 3. The difference between these two security levels is that keys are generated, encrypted, and transmitted to the communication partners such as the MU, home and foreign agents involved in this session at security level 4. After the keys are decrypted by the communication entities, the keys will be used to encryption/decryption and message authentication code to protect the communication. Therefore, the integrity of data can be guaranteed by message integrity check techniques, and the secrecy is protected with data encryption. The network resource is also protected since the identification cannot be compromised due to the protection of data integrity and secrecy.

From Table 1 and description above, we can see that the higher the security level, the better security services the authentication provides. However, higher security levels are achieved by applying more complicated cryptographic techniques in the authentication process. The extra operations induce the overhead that affects the QoS metrics, such as authentication cost, delay and call dropping probability during authentication.

3.2.2. Average authentication cost

In this context, we define *authentication cost* as the sum of signaling load and processing load for cryptographic techniques during each authentication operation. And, the *average authentication cost*, $C(i)$, is defined as the sum of the authentication cost over a number of authentication requests per unit time at security level i, which can be written as

$$C(i) = \sum_{\beta=1}^{3} \lambda_\beta \left[C_\beta^{(s)}(i) + C_\beta^{(p)}(i) \right], \tag{6}$$

where β is the index of authentication type. $\beta = 1$ represents an intra-domain handoff authentication, $\beta = 2$ means a session authentication, and $\beta = 3$ is an inter-domain handoff authentication. We denote $C_\beta^{(s)}(i)$ and $C_\beta^{(p)}(i)$ as the signaling load and processing load of cryptographic techniques, respectively, for the authentication of type β at security level i. The arrival rate of requests for the authentication type β is defined as λ_β, which is related with the mobility and traffic patterns of MUs.

3.2.3. Average authentication delay

We define *authentication delay* as the time from when an MU sends out an authentication request to when the MU receives the authentication reply. The *average authentication delay*, $T(i)$, is defined as the sum of an authentication delay over a number

of authentication requests in a unit time at security level i. Then, $T(i)$ can be written as

$$T(i) = \sum_{\beta=1}^{3} \lambda_\beta T_\beta(i), \tag{7}$$

where $T_\beta(i)$ is the authentication delay per operation at security level i for authentication type β, and λ_β is the arrival rate of authentication requests with type β.

3.2.4. Average call dropping probability during authentication

When an extended waiting time for authentication is induced and greater than a threshold time, the connection will be broken [31,32]. On the other hand, even though the authentication delay is small and the MU is a valid user, an authentication failure may happen because of unknown effects on the credentials such as undetectable packet loss or damage. The data loss or damage may come from transmission error, packet drop at queues, attack of intruders and software application failure.

In order to consider the extended authentication delay and authentication failure in the definition of call dropping probability, the *call dropping probability* is defined as the probability that the service of an MU is dropped during one authentication operation because of either extended authentication delay, or an authentication failure. When an MU roams among subnets in a network domain, the *average call dropping probability*, $P(i)$, is defined as the ratio of the sum of the call dropping probability per authentication in a unit time over the number of authentication requests sent by the MU within unit time at security level i. Let $P(i)$ denote the average call dropping probability at security level i, $P(i)$ can be written as

$$P(i) = \frac{\sum_{\beta=1}^{3} \lambda_\beta [P_\beta(i) + P_e]}{\sum_{\beta=1}^{3} \lambda_\beta} \quad \text{and}$$

$$P_\beta(i) = P_{T_\beta(i)}(T_\beta(i) > T_{th}), \tag{8}$$

where T_{th} is a threshold time, $P_{T_\beta(i)}(T_\beta(i) > T_{th})$ is the probability that an authentication delay is greater than the threshold time T_{th} in authentication type β. P_e is the probability that one authentication fails due to unknown damage on the credentials of a valid MU and it is unrelated with the security level i. Since the factors that affect P_e include many unknown factors and there is no evidence on the pattern of attacks currently, we will use a mean value from experiments to represent P_e in the numeric results of our paper [33].

In summary, in order to evaluate $C(i)$, $T(i)$ and $P(i)$ in (6)–(8), we need to analyze λ_β, $C_\beta^{(s)}(i)$, $C_\beta^{(p)}(i)$, $T_\beta(i)$, and $P_\beta(i)$. Next, we derive these parameters based on the system model shown in Fig. 2, assumptions described in Section 3.1, and the definitions of the performance metrics in Section 3.2.

4. Performance analysis

In this section, we analyze the impact of authentication on security and QoS in terms of authentication cost, delay and call dropping probability. The analysis has two key aspects. First is to observe the relationship between the security levels and the QoS. We evaluate the authentication cost, delay, and call dropping probability per authentication at different security levels. Second is to obtain the arrival rates of authentication requests. After that, the average authentication cost, delay, and call dropping probability defined in (6)–(8) can be evaluated.

4.1. Performance analysis per authentication

At different security levels, the authentication has different effects on the cost, delay and call dropping probability.

4.1.1. Authentication cost per operation

The authentication cost, $C_\beta(i)$ ($\beta = 1,2,3$ and $i = 1,2,3,4$), is composed of $C_\beta^{(s)}(i)$ and $C_\beta^{(p)}(i)$ as defined in (6), which depend on the authentication type β and security level i. For convenient analysis, we define a set of cost parameters in Table 2.

Then, the transmission costs, $C_\beta^{(s)}(i)$, can be derived from the signaling diagrams in Figs. 3–5, respectively, as follows:

Table 2
Authentication cost parameters

Symbol	Description
c_s	Transmission cost on one hop
c_p	Encryption/decryption cost on one hop
c_v	Verification cost at an authentication server
c_{us}	A pair of encryption and decryption cost for a value
c_g	Key generation cost
c_{ts}	Transmission cost for a session key to other communication identities
c_{rg}	Registration cost

$$C_{\beta}^{(s)}(i) = a_{\beta,i} \cdot c_s \quad \forall \beta = 1,2,3 \text{ and } i = 1,2,3,4, \tag{9}$$

where $a_{\beta,i}$ is an element of matrix A shown in (10), β is the authentication type and represents the row of matrix A, and i is the security level and indicates the column of matrix A. $a_{\beta,i}$ shows the number of hops by which the entire authentication process passes for authentication type β at security level i, and $a_{\beta,i}$ can be obtained from observing the corresponding signaling figures.

For example, when $\beta = 3$ and $i = 4$ as shown in Fig. 5B, an MU needs to request challenge value from the LAS through an AP first. The distance that the messages traverse is 4 in this step. Then, the authentication messages need to reach the HAS of the MU since no SA exists between the MU and the LAS. The distance between the MU and its HAS is assumed to be N_h hops. Since $\beta = 3$ means inter-domain authentication, a registration process to the HA of this MU is needed, which requires the messages to traverse 2 more hops in a round-trip transmission. Thus, the total number of hops that the authentication messages traverse at round-trip transmission in the case of $\beta = 3$ and $i = 4$, i.e., $a_{3,4}$, is $4 + 2N_h + 2 = 2(N_h + 3)$. The other elements of matrix A can be derived in the same way. Thus, we obtain matrix A as

$$A = \begin{bmatrix} 2 & 6 & 8 & 8 \\ 2 & 2(N_h+1) & 2(N_h+2) & 2(N_h+2) \\ 2(N_h+1) & 2(N_h+2) & 2(N_h+3) & 2(N_h+3) \end{bmatrix}, \tag{10}$$

where β and i represent the row and column of A, respectively. N_h is the number of the hops between the MU and its HAS.

Similar to the analysis in (9), according to the signaling diagrams in Figs. 3–5, $C_{\beta}^{(p)}(i)$ can be written as

$$C_{\beta}^{(p)}(i) = \vec{b}_{\beta,i} \cdot \vec{x}_p \quad \forall \beta = 1,2,3 \text{ and } i = 1,2,3,4. \tag{11}$$

Here, $\vec{x}_p$ is a vector defined as

$$\vec{x}_p^{\mathrm{T}} = [c_p,\ c_v,\ c_{us},\ c_g,\ c_{ts},\ c_{rg}], \tag{12}$$

where all of the cost parameters are defined in Table 2. And, $\vec{b}_{\beta,i}$ are vectors determined by

$$\begin{aligned}
&\vec{b}_{1,1} = \vec{b}_{2,1} = [0,0,0,0,0,0],\\
&\vec{b}_{1,2} = [2,1,0,0,0,0],\\
&\vec{b}_{1,3} = \vec{b}_{1,4} = [4,1,1,0,0,0],\\
&\vec{b}_{2,2} = [2(N_h-1),1,0,0,0,0],\\
&\vec{b}_{2,3} = [2N_h,1,1,0,0,0],\\
&\vec{b}_{2,4} = [2N_h,1,2,1,1,0],\\
&\vec{b}_{3,1} = [0,0,0,0,0,1],\\
&\vec{b}_{3,2} = [2N_h,1,0,0,0,1],\\
&\vec{b}_{3,3} = [2(N_h+1),1,1,0,0,1],\\
&\vec{b}_{3,4} = [2(N_h+1),1,2,1,1,1].
\end{aligned} \tag{13}$$

The coefficients in front of the cost variables in $\vec{x}_p$, such as c_p, c_v, and c_{ts}, denote the number of the costs we should consider during one authentication. For example, in the case of $\beta = 3$ and $i = 4$, the authentication messages need to traverse $2(N_h + 3)$ hops as analyzed in (10). On this authentication path, no encryption/decryption exists on the hop between the MU and the AP before the arrival of authentication approval. Thus, 4 hops should be reduced from the total number of hops that authentication messages pass by when we consider the encryption/decryption cost on one hop, i.e., c_p. Thus, the coefficient for c_p is $2(N_h + 3) - 4 = 2(N_h + 1)$. In this authentication process, the challenge/response values are verified once at the HAS. Thus, the coefficient for c_v is 1. In this case, two pairs of encryption and decryption costs are needed between the MU and its HAS. One pair

is for encrypting and decrypting the challenge/response values; the other is for encrypting and decrypting the session key. Thus, the coefficient for c_{us} is 2. In addition, since one time registration is needed at the HA in this case, the coefficient for c_{rg} is 1. Because the HAS of the MU needs to generate a key for the MU, the coefficient for c_g is 1. Finally, a corresponding key needs to be transmitted to the MU's communication partners by the HAS. Thus, the coefficient of c_{ts} is 1.

Similar to the case of $\beta = 3$ and $i = 4$, the coefficients of time parameters in other cases can be determined according to the corresponding time diagram in Figs. 3–5.

4.1.2. Delay per authentication

To derive the delay for different types of authentications in different security levels, we use the same signaling diagram shown in Fig. 1. We also define a set of time parameters shown in Table 3 for convenient description.

Then, $T_\beta(i)$ can be expressed as

$$T_\beta(i) = \vec{d}_{\beta,i} \cdot \vec{x}_t \quad \forall \beta = 1,2,3 \text{ and } i = 1,2,3,4. \tag{14}$$

Here, $\vec{x}_t$ is a vector defined as

$$\vec{x}_t^{\mathrm{T}} = [T_{pr} + T_{tr}, T_{ed}, T_a, T_{sq}, T_v, T_{us}, T_g, T_{ts}, T_{rg}], \tag{15}$$

Table 3
Authentication time parameters

Symbol	Description
T_{pr}	Message propagation time on one hop
T_{tr}	Message transmission time on one hop
T_{ed}	Message encryption/decryption time on one hop
T_a	Authentication request service and waiting time at the AP
T_{sq}	Authentication request service and waiting time at the proxy authentication server
T_v	Authentication request service and waiting time at the HAS
T_{us}	A pair of encryption and decryption time for a value
T_g	Key generation time at the HAS
T_{ts}	Transmission time for the session key to the other communication identities such as HA
T_{rg}	Registration request service and waiting time at the HA

where all the time components are defined in Table 3. And, $\vec{d}_{\beta,i}$ are the vectors defined as follows:

$$\begin{aligned}
&\vec{d}_{1,1} = [2,0,1,0,0,0,0,0,0],\\
&\vec{d}_{1,2} = [6,2,3,0,1,0,0,0,0],\\
&\vec{d}_{1,3} = \vec{d}_{1,4} = [8,4,4,0,2,1,0,0,0],\\
&\vec{d}_{2,1} = [2,0,1,0,0,0,0,0,0],\\
&\vec{d}_{2,2} = [2(N_h+1),2(N_h-1),3,2(N_h-2),1,0,0,0,0],\\
&\vec{d}_{2,3} = [2(N_h+2),2N_h,4,2(N_h-2),1,1,0,0,0],\\
&\vec{d}_{2,4} = [2(N_h+2),2N_h,4,2(N_h-2),1,2,1,1,0],\\
&\vec{d}_{3,1} = [2(N_h+1),0,2,2(N_h-1),0,0,0,0,1],\\
&\vec{d}_{3,2} = [2(N_h+2),2N_h,3,2(N_h-2),2,0,0,0,1],\\
&\vec{d}_{3,3} = [2(N_h+3),2(N_h+1),4,2(N_h-2),2,1,0,0,1],\\
&\vec{d}_{3,4} = [2(N_h+3),2(N_h+1),4,2(N_h-2),2,2,1,1,1].
\end{aligned} \tag{16}$$

The coefficients in front of the time variables in $\vec{x}_t$ denote the number of time variables for each authentication. For example, in the case of $\beta = 3$ and $i = 4$ as shown in Fig. 5, the message for an MU to request challenge value from an LAS must traverse through an AP on two hops first. Then, the response value needs to be transmitted to the HAS via N_h hops since there is no shared SA between the MU and the LAS in the case of inter-domain handoff authentication. At this moment, a registration process is needed. Thus, a distance of one hop between the HAS and the HA needs to be passed by the message. Therefore, the number of hops that the round-trip signaling messages traverse in the authentication process is $2(N_h + 3)$. The coefficient in front of $T_{pr} + T_{tr}$ is $2(N_h + 3)$. On this path for authentication, there is no encryption and decryption of messages on the hop between the MU and the AP before authentication. Since the authentication message traverses this hop four times, the number of hops that we should consider the encryption/decryption during the authentication process is $2(N_h + 3) - 4 = 2(N_h + 1)$. Thus, the coefficient in front of T_{ed} is $2(N_h + 1)$.

Similarly, since the authentication process in the case of $\beta = 3$ and $i = 4$ needs to pass the AP four times, the coefficient of T_a, i.e., authentication

request service and waiting time, is 4. Because the authentication messages cross the intermediate authentication servers $2(N_h - 2)$ times, the coefficient of T_{sq}, i.e., authentication request service and waiting time at a proxy authentication server, is $2(N_h - 2)$. The authentication message also traverses the HAS twice when registration is needed. Thus, the coefficient of T_v, i.e., authentication request service and waiting time at the HAS, is 2.

Since the authentication in this case also needs one time registration at the HA, the coefficient for T_{rg}, i.e., registration request service and waiting time at an HAS, is 1. Because a key is generated at the HAS for the communication of the MU, the coefficient of T_g, i.e., key generation time at the HAS, is 1. The HAS also needs to transmit a corresponding key to the MU's communication partners, thus the coefficient of T_{ts}, i.e., transmission time for the session key to the other communication identities such as HA, is 1. In addition, two pairs of encryption and decryption time are needed between the MU and its HAS. One pair is for encrypting and decrypting the challenge/response values; the other is for encrypting and decrypting the session key. Thus, the coefficient of T_{us}, i.e., a pair of encryption and decryption time for a value, is 2.

For the other cases of authentication processes, they follow the same analysis and can be obtained from Figs. 3–5.

4.1.3. Call dropping probability

In Section 3.2.4, we consider a call is dropped during authentication if the waiting time for authentication is greater than a threshold value T_{th}, or an authentication failure happens. As defined in (8), we use a mean value from an experiment for the probability that authentication failure happens, i.e., P_e, due to the unknown distribution model of P_e. Therefore, in order to evaluate $P_\beta(i)$ ($\beta = 1, 2, 3$ and $i = 1, 2, 3, 4$), the PDF of the authentication delay shown in (14) needs to be evaluated.

In (14), we only consider the time variables, T_{sq}, T_a, T_v, and T_{rg}, as the random variables because the variance of the other time variables are small. T_{ed} and T_{us} are mainly related with the ability of computer and the message length, T_{tr} is determined by the message length and the link speed, T_{pr} is a function of the distance between two points, and T_g is directly connected with the computer ability. In reality, the computer ability, message length, link speed, and distance between two points are all fixed. Therefore, we do not consider T_{ed}, T_{tr}, T_{pr}, T_{us}, and T_g random variables in this paper. However, T_a, T_{sq}, T_v, and T_{rg} are all related with the traffic load, queue length and service time, which are varied from time to time and have big variance.

Thus, to find $P_\beta(i)$ becomes to find the PDFs of the different combinations of T_{sq}, T_a, T_v, and T_{rg} in $T_\beta(i)$. For simplification, we consider that: (1) $M/M/1$ queues are applied at APs, authentication servers, and HAs; (2) The PDFs of T_{sq}, T_a, T_v, and T_{rg} are independent identical distribution (iid). Then, the PDF of T_{sq}, T_a, T_v, and T_{rg}, i.e., $w(t)$, can be shown as [34]:

$$w(t) = (\mu_s - \lambda_s)\mathrm{e}^{-(\mu_s - \lambda_s)t}, \tag{17}$$

where μ_s and λ_s are the service and arrival rates of authentication requests, respectively. Furthermore, the PDFs of the different combinations of T_{sq}, T_a, T_v, and T_{rg} in $T_\beta(i)$, i.e., $f_{\beta,i}(t)$, can be expressed in (18), as the components of a matrix $f(t)$

$$f(t) = \begin{bmatrix} \xi\mathrm{e}^{-\xi t} & \frac{\xi(\xi t)^3\mathrm{e}^{-\xi t}}{\Gamma(4)} & \frac{\xi(\xi t)^5\mathrm{e}^{-\xi t}}{\Gamma(6)} & \frac{\xi(\xi t)^5\mathrm{e}^{-\xi t}}{\Gamma(6)} \\ \xi\mathrm{e}^{-\xi t} & \frac{\xi(\xi t)^{2N_h-1}\mathrm{e}^{-\xi t}}{\Gamma(2N_h)} & \frac{\xi(\xi t)^{2N_h}\mathrm{e}^{-\xi t}}{\Gamma(2N_h+1)} & \frac{\xi(\xi t)^{2N_h}\mathrm{e}^{-\xi t}}{\Gamma(2N_h+1)} \\ \frac{\xi(\xi t)^{2N_h}\mathrm{e}^{-\xi t}}{\Gamma(2N_h+1)} & \frac{\xi(\xi t)^{2N_h+1}\mathrm{e}^{-\xi t}}{\Gamma(2N_h+2)} & \frac{\xi(\xi t)^{2N_h+2}\mathrm{e}^{-\xi t}}{\Gamma(2N_h+3)} & \frac{\xi(\xi t)^{2N_h+2}\mathrm{e}^{-\xi t}}{\Gamma(2N_h+3)} \end{bmatrix}. \tag{18}$$

In (18), the row represents the index of authentication type β, and the column is the index of security level i. $\Gamma(x) \triangleq \int_0^\infty s^{x-1}\mathrm{e}^{-s}\mathrm{d}s$, and $\xi = \mu_s - \lambda_s$. Thus, $f_{\beta,i}(t)$ can be obtained from the combination of T_{sq}, T_a, T_v, and T_{rg} in $T_\beta(i)$. For example, in the case of $\beta = 3$ and $i = 4$, according to (14)–(16), $T_3(4)$ can be written as

$$\begin{aligned} T_3(4) = {} & 2(N_h + 3)(T_{pr} + T_{tr}) + 2(N_h + 1)T_{ed} \\ & + 4T_a + 2(N_h - 2)T_{sq} + 2T_v + 2T_{us} + T_g \\ & + T_{ts} + T_{rg}. \end{aligned} \tag{19}$$

Recall that the PDFs of T_{sq}, T_a, T_v, and T_{rg} are assumed to be iid with the same definition as shown

in (17), and the other time variables, i.e., T_{pr}, T_{tr}, T_{ed}, T_{us}, T_g, T_{ts}, are assumed to be constants. Then, $f_3(4)$ can be written as $\xi(\xi t)^{2N_h+2}e^{-\xi t}/\Gamma(2N_h+3)$. Furthermore, with these PDFs $f_\beta(i)$, $P_\beta(i)$ can be obtained in different cases.

To summarize, we have obtained authentication cost, delay, and call dropping probability for one authentication operation. However, in order to obtain the average authentication cost, delay, and call dropping probability defined in (6)–(8), we need to evaluate the arrival rates of different types of authentication requests, that is, λ_β ($\beta = 1,2,3$).

4.2. Arrival rates of authentication requests

In this paper, the authentication requests are categorized into three types: intra-domain handoff authentication, session authentication, and inter-domain handoff authentication. Thus, we analyze the arrival rates of different types of authentication requests, i.e., λ_β ($\beta = 1,2,3$), based on the mobility and traffic patterns of the MUs.

4.2.1. Arrival rate of intra-domain handoff authentication, λ_1

The intra-domain handoff authentication requests happen whenever an MU crosses the boundaries of subnets inside a network domain with an on-going service. In order to calculate the arrival rate of intra-domain handoff authentication requests, we define four events, in which calls will happen

- Y_1 is the event that an MU starts a connection before entering the network domain, enters the network domain with the on-going connection and this connection ends before the MU moves out of the network domain.
- Y_2 is the event that an MU starts a connection within current network domain and this connection ends before the MU moves out of the network domain.
- Y_3 is the event that an MU starts a connection within current network domain and this connection ends after the MU moves out of the network domain.
- Y_4 is the event that an MU starts a connection before entering the network domain, enters the network domain with the on-going connection, and the connection ends after moving out of the network domain.

Then, the arrival rate of intra-domain handoff authentication requests, λ_1, can be written as

$$\lambda_1 = \sum_{j=1}^{4} \lambda_u P_{rj}(\lceil \overline{N}_{aj} \rceil - 1), \tag{20}$$

where λ_u is the call arrival rate defined in (4), P_{rj} is the probability that event Y_j happens, $\overline{N}_{aj}$ is the average number of subnets passed by an MU in current network domain in event Y_j ($j = 1,2,3,4$). $\lceil \overline{N}_{aj} \rceil - 1$ represents the average number of intra-domain handoff authentication in event Y_j. Since the intra-domain handoff authentication only happens when an MU crosses the boundaries of subnets inside a network domain with an on-going service, the average number of intra-domain handoff authentication is equal to the average number of subnets passed by an MU minus one.

The time diagrams of these events, Y_j ($j = 1,2,3,4$), are shown in Fig. 6, where t_c^0 is the call beginning time, t_c^1 is the call ending time, t_n^0 is the time when an MU enters the network domain we are investigating, t_n^1 is the time when an MU leaves the network domain we are investigating, and t_{mr}^0 is

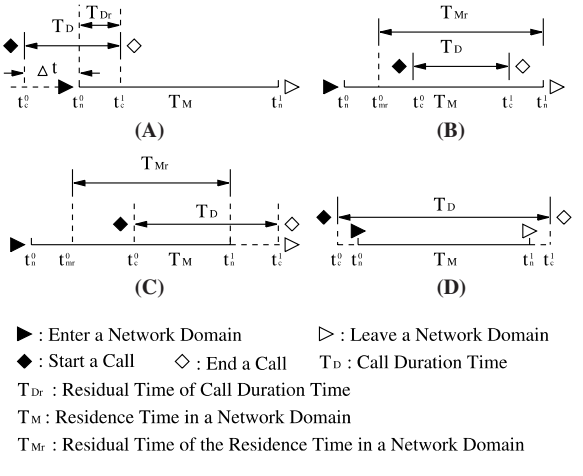


Fig. 6. Time diagrams of events: (A) Y_1; (B) Y_2; (C) Y_3 and (D) Y_4.

the beginning time of the residual time of the residence time in a network domain. Therefore, the probabilities of these events, P_{rj}, $(j = 1, 2, 3, 4)$, can be derived as follows.

According to the time diagram in Fig. 6A, and denote $\Delta t = t_n^0 - t_c^0$, we have

$$P_{r1} = \int_0^\infty P_r[I(t_c^0 + \Delta t, t_c^0) = 1] \cdot P_r(T_D > \Delta t)\mathrm{d}(\Delta t) \cdot P_r(T_{Dr} \leqslant T_M), \tag{21}$$

where $I(t_c^0 + \Delta t, t_c^0)$ is the number of calls that arrive in the time interval $[t_c^0, t_c^0 + \Delta t)$. Since we assume that the call arrival rate is a Poisson process, $P_r[I(t_c^0 + \Delta t, t_c^0) = 1]$ can be determined by

$$P_r[I(t_c^0 + \Delta t, t_c^0) = 1] = \lambda_u \Delta t \mathrm{e}^{-\lambda_u \Delta t}, \tag{22}$$

where λ_u is the average arrival rate of the calls. In (21), T_D is the call duration time with PDF defined in (5), and T_M is the residence time of an MU in the network domain with Laplace transform of PDF in (2). Thus, we have:

$$P_r(T_D > \Delta t) = \int_{\Delta t}^\infty f_{T_D}(t)\mathrm{d}t = \mathrm{e}^{-\eta \Delta t}, \tag{23}$$

where $f_{T_D}(t)$ is defined in (5), $1/\eta$ is the average call holding time and $\Delta t = t_n^0 - t_c^0$.

Furthermore, T_{Dr} is the residual time of the call duration time with the same PDF as T_D defined in (5) due to the memoryless property of exponential distribution. Since we have the Laplace transform of the PDF of T_M defined in (2), $P_r(T_{Dr} \leqslant T_M)$ can be determined by

$$P_r(T_{Dr} \leqslant T_M) = \int_0^\infty f_{X_1}(t)\mathrm{d}t, \tag{24}$$

where $X_1 \triangleq T_M - T_{Dr}$, and $f_{X_1}(t)$ can be computed from

$$f_{X_1}(t) = \mathscr{L}^{-1}\left\{\frac{(\eta + s)F_M(s)}{\eta}\right\}. \tag{25}$$

Here, $1/\eta$ is the average call holding time, $\eta/(\eta + s)$ is the Laplace transform of the PDF of T_{Dr}, and $F_M(s)$ is the Laplace transform of the PDF of T_M defined in (2).

Thus, P_{r1} can be calculated by substituting (22)–(24) into (21). Next, we need to derive P_{r2} from Fig. 6B as

$$P_{r2} = P_r(T_D < T_{Mr}) \cdot P_r(t_{mr}^0 \leqslant t_c^0 < t_{mr}^0 + T_{Mr}) = \int_0^\infty f_{X_2}(t)\mathrm{d}t \cdot \int_0^\infty \lambda_u t \mathrm{e}^{-\lambda_u t} f_{Mr}(t)\mathrm{d}t, \tag{26}$$

where $X_2 \triangleq T_{Mr} - T_D$, $f_{X_2}(t)$ and $f_{Mr}(t)$ are the PDFs of X_2 and T_{Mr}, respectively, which can be obtained by

$$f_{X_2}(t) = \mathscr{L}^{-1}\left\{F_{Mr}(s)\frac{\eta + s}{\eta}\right\}, \quad f_{Mr}(t) = \mathscr{L}^{-1}\{F_{Mr}(s)\}, \tag{27}$$

where $1/\eta$ is the average call holding time, and $F_{Mr}(s)$ is the Laplace transform of the PDF of T_{Mr}, the residual time of the residence time in a network domain. $F_{Mr}(s)$ can be obtained by

$$F_{Mr}(s) = \frac{1 - F_M(s)}{s\overline{T}_M}, \tag{28}$$

where $\overline{T}_M$ is defined in (3), and $F_M(s)$ is defined in (2).

Moreover, we can obtain P_{r3} from Fig. 6C

$$P_{r3} = P_r(T_D > T_{Mr}) \cdot P_r(t_{mr}^0 \leqslant t_c^0 < t_{mr}^0 + T_{Mr}) = \int_0^\infty f_{X_3}(t)\mathrm{d}t \cdot \int_0^\infty \lambda_u t \mathrm{e}^{-\lambda_u t} f_{Mr}(t)\mathrm{d}t, \tag{29}$$

where $X_3 \triangleq T_D - T_{Mr}$, $f_{Mr}(t)$ is the PDF of T_{Mr} defined in (27), $f_{X_3}(t)$ is the PDF of X_3, which can be obtained by

$$f_{X_3}(t) = \mathscr{L}^{-1}\left\{\frac{\eta}{(\eta + s)F_{Mr}(s)}\right\}, \tag{30}$$

where $F_{Mr}(s)$ is defined in (28), η is defined in (5).

Similar with P_{r1}, P_{r4} can be determined from Fig. 6D as follows:

$$P_{r4} = \int_0^\infty P_r[I(t_c^0 + \Delta t, t_c^0) = 1] \cdot P_r(T_D > \Delta t)\mathrm{d}(\Delta t) \cdot P_r(T_{Dr} > T_M), \tag{31}$$

where $P_r[I(t_c^0 + \Delta t, t_c^0) = 1]$ is shown in (22), $P_r(T_D > \Delta t)$ is defined in (23), and $P_r(T_{Dr} > T_M) = 1 - P_r(T_{Dr} \leqslant T_M)$, where $P_r(T_{Dr} \leqslant T_M)$ is defined in (24).

Therefore, we have calculated P_{rj} $(j = 1, 2, 3, 4)$. In order to evaluate λ_1, we will evaluate the average number of subnets passed by an MU in a network domain during one call in the events Y_j, i.e.,

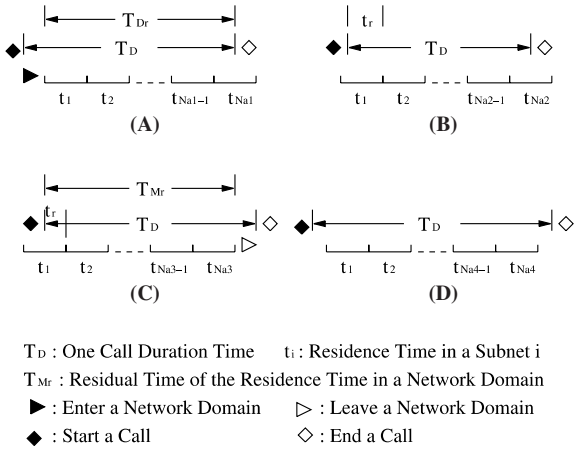


Fig. 7. Time diagram for number of subnets passed by for one call: (A) N_{a1}; (B) N_{a2}; (C) N_{a3} and (d) N_{a4}.

$\overline{N}_{aj}$ $(j = 1, 2, 3, 4)$, respectively. The time diagrams to evaluate $\overline{N}_{aj}$ are shown in Fig. 7.

In order to evaluate $\overline{N}_{a1}$ and $\overline{N}_{a2}$, we consider a theorem in [35], which says that given call holding time and subnet residence time with Gamma distribution, the average number of subnets passed by an MU within a call, denoted as $\overline{K}$, can be obtained by

$$\overline{K} = -\alpha \sum_{p \in \sigma_c} \mathrm{Res}_{s=p} \frac{1 - f^*(s)}{1 - (1 - p_f) f^*(s)} f_c^*(-s), \quad (32)$$

where $1/\alpha$ is the average residence time of an MU in a subnet, p_f is the probability that a handoff call is blocked, $f^*(s)$ is the Laplace transform of the PDF of the residence time of an MU in a subnet, $f_c^*(s)$ is the Laplace transform of the PDF of the call holding time of the MU, σ_c is the singular points of $f_c^*(-s)$, and $\mathrm{Res}_{s=p}$ denotes the residue at a singular point $s = p$.

In the events Y_1 and Y_2, the call duration time in a network domain are T_{Dr} and T_D, respectively, which are exponential distribution, one special case of Gamma distributions. Therefore, $\overline{N}_{a1}$ and $\overline{N}_{a2}$ can be calculated with (32). By assuming that $p_f = 0$, we can carry out $\overline{N}_{a1}$ and $\overline{N}_{a2}$ as

$$\overline{N}_{a1} = \overline{N}_{a2} = \frac{\mu_r}{\eta}, \quad (33)$$

where $1/\eta$ is the average call duration time of the MU and μ_r is the average residence time of the MU in a subnet in our paper.

On the other hand, note that the call duration time in events Y_3 and Y_4, i.e., T_{Mr} and T_M are not Gamma distributions, thus we cannot obtain $\overline{N}_{a3}$ and $\overline{N}_{a4}$ with (32). Therefore, we need to derive $\overline{N}_{a3}$ and $\overline{N}_{a4}$ next.

Fig. 7C illustrates the time diagram that event Y_3 happens. From Fig. 7C, the relationship between different time components can be written as follows:

$$T_{Mr} = t_r + \sum_{i=2}^{N_{a3}} t_i, \quad (34)$$

where T_{Mr} is the residual time of the residence time of an MU in a network domain. The Laplace transform of the PDF of T_{Mr} is shown in (28). t_r is the residual time of the residence time of an MU in a subnet. The Laplace transform of the PDF of t_r, denoted as $F_{tr}(s)$, is

$$F_{tr} = \mu_r \frac{1 - F_r(s)}{s}, \quad (35)$$

where $1/\mu_r$ is the average residence time of an MU in a subnet, $F_r(s)$ is the Laplace transform of the PDF of the residence time of an MU in a subnet defined in (1). In (34), t_i is the residence time of an MU in subnet i, which is assumed to be Gamma distribution with Laplace transform of PDF defined in (1), and N_{a3} is the random number of the subnets passed by an MU in the current network domain for event Y_3.

Based on the relationship in (34), we can obtain:

$$F_{Mr}(s) = F_{t_r}(s) G_{N_{a3}-1}(z)|_{z=F_r(s)}, \quad (36)$$

where $F_{Mr}(s)$ is defined in (28), $Ft_r(s)$ is defined in (35), $G_{N_{a3}} - 1(z)$ is the generating function of the PDF of $N_{a3}-1$. Then, $\overline{N}_{a3}$ can be obtained by

$$\overline{N}_{a3} = \frac{\partial G_{N_{a3}-1}(z)}{\partial z}|_{z=1} + 1 = \frac{2M^2 - M - 1}{12\overline{T}_M \mu_r} + \frac{(M+1)}{4}\frac{(\gamma+1)}{\gamma} + 1, \quad (37)$$

where $\overline{T}_M$ is defined in (3), M is the number of subnets in current network domain, $1/\mu_r$ is the average residence time that an MU stays in a subnet, and γ is defined in (1).

According to Fig. 7D, $\overline{N}_{a4}$ is equal to the average number of subnets that an MU passes when

the MU is roaming inside the network domain. Recall that the number of subnets that the MU passes by in a network domain, N_{sn}, is uniformly distributed between 1 and M, i.e.,

$$P(N_{sn} = m) = \frac{1}{M}, \quad m = 1, 2, \cdots, M. \tag{38}$$

Here, N_{sn} is the number of subnets that an MU passes by in a network domain, M is the total number of subnets in current network domain. Therefore, we have

$$\overline{N}_{a4} = \overline{N}_{sn} = \sum_{j=1}^{M} \frac{j}{M} = \frac{M+1}{2}. \tag{39}$$

Now we have obtained all $\overline{N}_{aj}$ at event Y_j, $j = 1, 2, 3, 4$. Since we have calculated P_{rj}, $j = 1, 2, 3, 4$, in (21), (26), (29) and (31), respectively, we can evaluate λ_1 by substituting the values of P_{rj} and N_{aj}, $j = 1, 2, 3, 4$, into (20). Next, in order to obtain $C(i)$, $T(i)$ and $P(i)$ defined in (6)–(8), we need to evaluate λ_2 and λ_3.

4.2.2. Arrival rate of session authentication, λ_2

After an MU has moved into a network domain, a session authentication is initiated whenever a call arrives. Therefore, the arrival rate of session authentication requests for one MU, e.g. λ_2, is equal to the call arrival rate of an MU,

$$\lambda_2 = \lambda_u, \tag{40}$$

where λ_u is assumed to be the call arrival rate in (4).

4.2.3. Arrival rate of inter-domain handoff authentication, λ_3

The inter-domain handoff authentication requests happen when an MU enters the network domain with an on-going service. Therefore, the arrival rate of inter-domain handoff authentication requests, λ_3, can be obtained by

$$\lambda_3 = \lambda_u(P_{r1} + P_{r4}), \tag{41}$$

where λ_u is the call arrival rate assumed in (4), P_{r1} and P_{r4} are the probabilities that events Y_1 and Y_4 occur, which are defined in Section 4.2.1 and evaluated in (21) and (31), respectively.

Thus, we have obtained the arrival rates of authentication requests in the cases of intra-domain handoff authentication, session authentication, and inter-domain handoff authentication. Since two key aspects, i.e., the relationship between the security and system performance, and the relationship between the QoS metrics and traffic load, have been evaluated, the impact of authentication on security and the system performance can be observed clearly through $C(i)$, $T(i)$, and $P(i)$ in (6)–(8).

5. Numerical results

In this section, we evaluate the effects of mobility and traffic patterns on authentication cost, $C(i)$, delay, $T(i)$, and call dropping probability, $P(i)$, at different security levels.

5.1. Assumptions and parameters

The numerical results are presented based on the assumptions introduced in Sections 3 and 4.1.3. Of the assumptions in Section 3, we consider an MU roaming within a foreign network shown in Fig. 2. The mobility pattern of the MU is represented with the residence time in a subnet of the network domain, which is assumed to be Gamma distribution with the mean value $1/\mu_r$. The traffic patterns of an MU are represented by all arrival rate and call duration time. The call arrival rate is assumed to be Poisson process with mean value $1/\lambda_u$, and the call duration time is assumed to be exponential distribution with mean value $1/\eta$.

In Section 4.1.3, we further assume that $M/M/1$ queues are used at APs, authentication servers such as LAS and HAS, and HAs with service rate μ_s and arrival rate of authentication requests λ_s. Let $\xi = \mu_s - \lambda_s$. According to (17), the service and waiting time at an AP, authentication server, and HA, e.g., T_a, T_{sq}, and T_v, become random variables with identical exponential distribution with mean value of $1/\xi$. The parameters to evaluate the authentication cost and delay are shown in Table 4.

There are many ways to determine the values for the authentication costs. For example, the authentication cost for signaling can be measured with the number of messages, and the authentication cost for encryption can be measured with the

Table 4
Parameters for evaluation on QoS metrics

Parameters for authentication cost					
c_s	c_p	c_v	c_g	c_{ts}	N_h
10	1	20	1	110	10
Parameters for authentication delay					
T_{th}	T_{pr}	T_{tr}	T_{ed}	T_g	M
3 s	40 μs	20 ms	2 ms	2 ms	120
Parameters for random variables					
λ_u	η	γ	μ_r	ξ	
0.1 min^{-1}	0.3 min^{-1}	225	1/15 min^{-1}	15 s^{-1}	

number of CPU cycles. However, the most important problem here is how to make them consistent, i.e., the values of the costs can be compared with each other in the same scale. To solve this problem, we assume that the encryption/decryption cost on one hop, c_p, and the key generation cost, c_g, are normalized to a cost unit because they are the lightest load compared to other costs and they have the similar operation in cryptography techniques [36,37]. The values of other costs are determined by comparing to c_p and c_g with the time to finish the operation, i.e., we use the ratio of processing time to represent the authentication cost instead of the actual processing time. The reason is that the time needed to finish an operation represents the load of the server to complete it. However, we do not use the processing time to represent the cost directly because we do not want to confuse the authentication cost with the authentication delay and the authentication cost can be evaluated with many other ways.

When the maximum authentication message size is 4096 bytes [3], the transmission delay is about 20 milliseconds with the assumption of 2 Mbps link capacity [36]. The values of T_{ed} and T_g are obtained from existing research [36,38]. By assuming one network domain is about 100 km^2 with radius 6 km, the value of the propagation time, T_{pr}, can be determined by the distance between two LASs as shown in Table 4.

5.2. Effects of mobility pattern at different security levels

The effects of mobility pattern on the authentication cost, delay, and call dropping probability are shown in Figs. 8–10. In these figures, we illustrate the relationships between the residence time of an MU in a subnet, authentication cost, delay, and call dropping probability, respectively.

In Fig. 8, authentication costs at different security levels decrease with the increase of the residence time of an MU in a subnet because the longer an MU stays in the subnets, the less the intra-domain handoff authentication requests. And, if the residence time of an MU approaches to infinity, the authentication cost will be stable on the session authentication cost because only session authentication exists in this case. Moreover, we

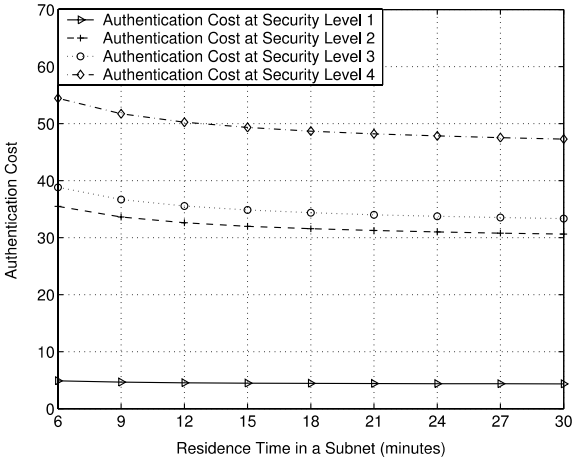


Fig. 8. Authentication cost vs. residence time in a subnet.

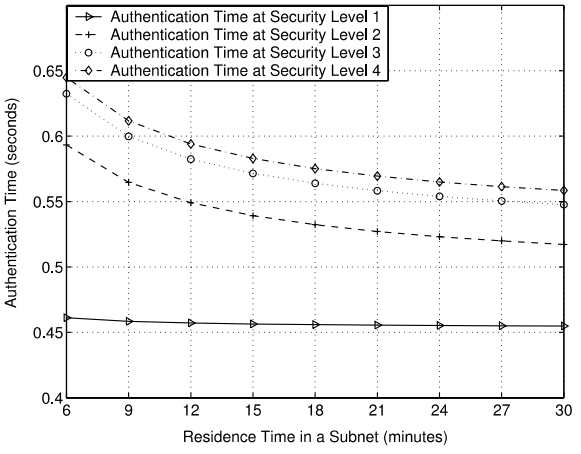


Fig. 9. Authentication time vs. residence time in a subnet.

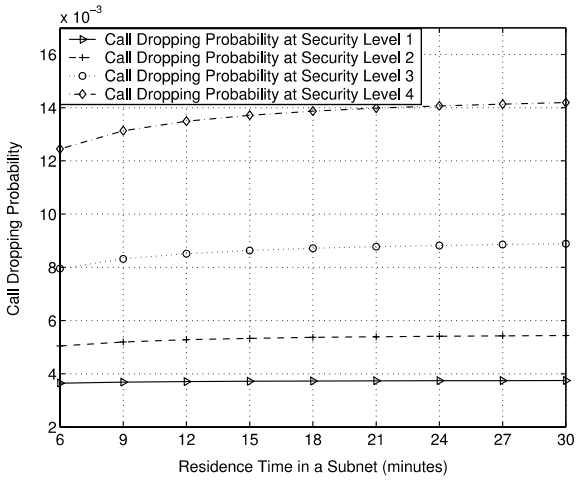


Fig. 10. Call dropping probability vs. residence time in a subnet.

can see that the security levels have different effects on the cost at the same residence time in a subnet. The higher the security level, the more the authentication cost because higher security levels impose more operations to provide secure services. For example, if we degrade the security level from 4 to 3, the authentication cost can be reduced up to 32%.

Fig. 9 reveals the effect of residence time on the authentication delay. As we can see, authentication delay decreases with the increase of the residence time of an MU in a subnet. Similar with the authentication cost, this trend is due to the decrease in the intra-domain handoff authentication requests. And, the higher security levels cause more authentication delay because of more operations needed for more secure services. The improvement of authentication delay by changing security levels from 4 to 3 is around 0.1 s, which is around 18.2% of the authentication delay at security level 3 when the residence time of an MU in a subnet is 27 min.

The effect of call dropping probability in authentication is shown in Fig. 10. The call dropping probability increases with the increase of the residence time of an MU in a subnet. When the residence time of an MU in a subnet increases, the arrival rate of intra-domain handoff authentication requests will decrease. Then, the session authentication requests become the major part of authentication requests. Note that the call dropping probability for session authentication is much higher than that in intra-domain handoff authentication due to the longer authentication delay caused by remote authentication. The call dropping probability will approximate that in session authentication if the residence time of an MU approaches infinity. In other words, the upper bound of the call dropping probability can be achieved when the authentication requests are all *session* authentication requests. Similar with the cost and delay, call dropping probability is greatly affected by the security levels. When the security level is leveraged from 3 to 4, call dropping probability increases about 0.45%, which is about 50% more than the call dropping probability at security level 3 when the residence time of an MU is 27 min.

5.3. Effect of traffic load at different security levels

The effects of traffic pattern on the authentication cost, delay, and call dropping probability at different security levels are demonstrated in Figs. 11–13.

Figs. 11 and 12 show that the authentication cost and delay increase with the call arrival rate of an MU. As shown in (6) and (7), the authentication cost and delay are proportional to the call arrival rate λ_u since variables λ_β ($\beta = 1, 2, 3$) are

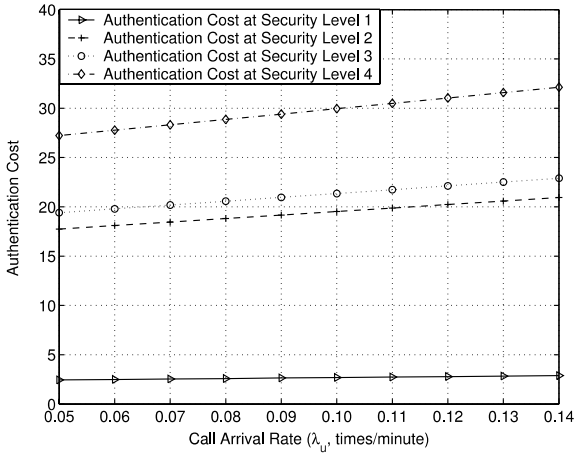


Fig. 11. Authentication cost vs. call arrival rate.

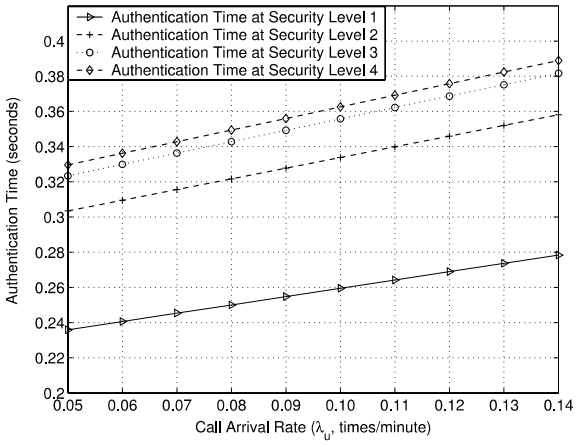


Fig. 12. Authentication time vs. call arrival rate.

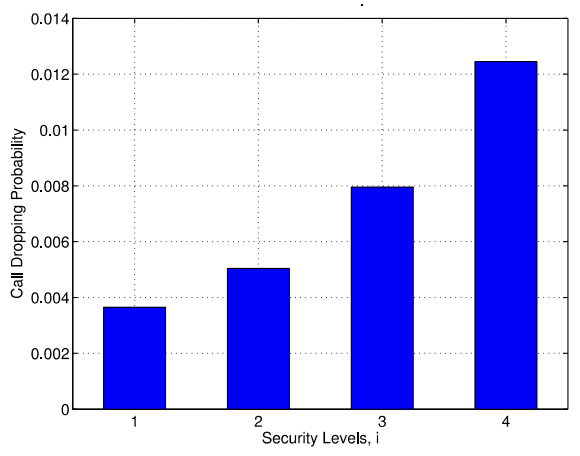


Fig. 13. Call dropping probability vs. security levels.

proportional to λ_u. Moreover, a higher security level causes more cost and delay than a lower one. For example, if the security level is changed from 1 to 2, the authentication will be about 740% more cost and 29% more time than those at security level 1.

As for the call dropping probability at different call arrival rates, the call arrival rate of an MU does not affect the call dropping probability. As we can see in (8), $P(i)$ for security level i is average call dropping probability computed in the cases of intra-domain handoff authentication, session authentication, and inter-domain handoff authentication. As shown in (20), (40) and (41), λ_β ($\beta = 1, 2, 3$) are all proportional to λ_u. Thus, λ_u disappears in (8), which is $P(i)$'s definition equation. Then, once the PDF of the call duration time and the mobility patterns of the MU are known, i.e., η, μ_r, and γ are fixed, the call dropping probability of the MU is a constant at different call arrival rates shown in Fig. 13. However, the call dropping probability is different at different security levels. As we can see in Fig. 13, the call dropping probability at security level 4 is about 56% more than that at security level 3.

6. Conclusion

In this paper, we conducted a performance analysis of authentication with respect to security and quality of service (QoS) in combination with mobility and traffic patterns because it is extremely important to deliver secure and efficient services in wireless networks. We proposed a system model that is consistent with realistic mobile environments, and analyzed authentication cost, delay, and call dropping probability at different security levels based on the system model and challenge/response authentication mechanism. Therefore, by coupling the security and QoS parameters in mobile environments, this paper presents a solid ground for an in-depth understanding of authentication impact, and demonstrates a framework for the future design of efficient authentication schemes in wireless networks.

References

[1] A. Arumugam, A. Doufexi, A. Nix, P. Fletcher, An investigation of the coexistence of 802.11g WLAN and high data rate bluetooth enabled consumer electronic devices in indoor home and office environments, IEEE Transactions on Consumer Electronics 49 (3) (2003) 587–596.

[2] L. Salgarelli, M. Buddhikot, J. Garay, S. Patel, S. Miller, The evolution of wireless LANs and PANs—efficient authentication and key distribution in wireless IP networks, IEEE Wireless Communications 10 (6) (2003) 52–61.

[3] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, J. Arkko, Diameter Base Protocol, draft-ietf-aaa-diameter-17.txt, December 2002.

[4] S. Jacobs, Mobile IP Public Key Based Authentication, draft-jacobs-mobileip-pki-auth-02.txt, March 1999.
[5] C. Perkins, P. Calhoun, Mobile IPv4 Challenge/Response Extensions, RFC3012, November 2000.
[6] M. Xu, S. Upadhyaya, Secure communication in PCS, in: Vehcular Technology Conference, 2001, VTC 2001, IEEE, 2001, pp. 2193–2197.
[7] B. Lee, T. Kim, S. Kang, Ticket-based authentication and payment protocol for mobile telecommunications systems, in: Proceedings of the International Symposium on Dependable Computing 2001, 2001, pp. 218–221.
[8] IEEE 802.11 Working Group. http://grouper.ieee.org/groups/802/11/index.html.
[9] V. Gupta, S. Gupta, S. Chang, Performance analysis of elliptic curve cryptography for SSL, in: WiSe'02-ACM Workshop on Wireless Security, September 2002.
[10] H. Kim, H. Afifi, Improving mobile authentication with new AAA protocols, in: Proceedings of the IEEE International Conference on Communications, vol. 1, 2003, pp. 497–501.
[11] W. Simpson, PPP challenge handshake authentication protocol (CHAP), RFC1334, August 1996.
[12] S. Shieh, F. Ho, Y. Huang, An efficient authentication protocol for mobile networks, Journal of Information Science and Engineering 15 (1999) 505–520.
[13] W. Liang, W. Wang, A cost-aware control scheme for efficient authentication in wireless networks, in: Proceedings of IEEE PIMRC'04, December 2004.
[14] B. Aboba, D. Simon, PPP EAP TLS Authentication Protocol, RFC2716, October 1999.
[15] P.G. Argyroudis, R. Verma, H. Tewari, D. O'Mahony, Performance analysis of cryptographic protocols in handheld devices, Technical Report TCD-CS-2003-46, University of Dublin, November 2003.
[16] L. Blunk, J. Vollbrecht, PPP extensible authentication protocol, RFC2284, March 1998.
[17] L. Dell'Uomo, E. Scarrone, The mobility management and authentication/authorization mechanisms in mobile networks beyond 3G, in: 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 1, September 2001, pp. c44–c48.
[18] S. Glass, T. Hiller, S. Jacobs, C. Perkins, Mobile IP authentication, authorization and accounting requirements, RFC2977, October 2000.
[19] Available from <http://standards.ieee.org/getieee802/download/802.1X-2001.pdf>.
[20] G. Kambourakis, A. Rouskas, S. Gritzalis, Performance evaluation of public key based authentication in future mobile communication systems, EURASIP Journal on Wireless Communications and Networking 1 (1) (2004) 184–197.
[21] W. Liang, W. Wang, An analytical study on the impact of authentication in wireless local area networks, in: Proceedings of IEEE ICCCN'04, October 2004.
[22] W. Liang, W. Wang, A quantitative study of authentication and QoS in Wireless IP Networks, in: Proceedings of IEEE INFOCOM'05, March 2005.
[23] W. Stallings, Network Security Essentials, Applications and Standards, Prentice-Hall, Upper Saddle River, NJ, 2000.
[24] J. Ho, I. Akyildiz, Mobile user location update and paging under delay constraints, Wireless Networks 1 (4) (1995) 413–425.
[25] W. Wang, I. Akyildiz, Intersystem location update and paging schemes for multitier wireless networks, in: Proceedings of ACM/IEEE MobiCom'2000, August 2000, pp. 99–109.
[26] J. Chen, K. Liu, Joint source-channel multistream coding and optical network adapter design for video over IP, IEEE Transactions on Multimedia 4 (1) (2002) 3–22.
[27] C. Chien, M. Srivastava, R. Jain, P. Lettieri, V. Aggarwal, R. Sternowski, Adaptive radio for multimedia wireless links, IEEE Transactions on Selected Areas in Communications 17 (5) (1999) 793–813.
[28] E. Bertino, S. Jajodia, L. Mancini, I. Ray, Advanced transaction processing in multilevel secure file stores, IEEE Transactions on Knowledge and Data Engineering 10 (1) (1998) 120–135.
[29] D. Rosenthal, F. Fung, A Test for Non-disclosure in security level translations, in: Proceedings of the 1999 IEEE Symposium on Security and Privacy, May 1999, pp. 196–206.
[30] S. Sutikno, A. Surya, An architecture of $F(2^{2N})$ multiplier for elliptic curves cryptosystem, in: Proceedings of ISCAS 2000 on Circuits and Systems, vol. 1, Geneva, Switzerland, May 2000, pp. 196–206.
[31] F. Hu, N. Sharma, Priority-determined multiclass handoff scheme with guaranteed mobile QoS in wireless multimedia networks, IEEE Transactions on Vehicular Technology 53 (1) (2004) 118–135.
[32] W. Wang, I. Akyildiz, A new signaling protocol for intersystem roaming in next-generation wireless systems, IEEE Journal on Selected Areas in Communications 19 (10) (2001) 2040–2052.
[33] Available from <http://www.paganini.net/ask/paper/node4.html>.
[34] D. Gross, C. Harris, Fundamentals of Queueing Theory, Wiley, New York, 1974.
[35] Y. Fang, I. Chlamtac, Y. Lin, Channel occupancy times and handoff rate for mobile computing and PCS networks, IEEE Transactions on Computer 47 (6) (1998) 679–692.
[36] A. Hess, G. Schafer, Performance evaluation of AAA/mobile IP authentication. Available from <http://www-tkn.ee.tu-berlin.de/publications/papers/pgts2002.pdf>, 2002.
[37] P. Calhoun, T. Johansson, C.E. Perkins, draft-ietf-aaa-diameter-mobileip-13.txt, IETF AAA Working Group, October 2002.
[38] K. Michael, C. Robert, D. Gary, Harris 2G encryption engine performance measurements, Harris Corporation, RF Communications Division Technical Report, RFCD MD04, October 2004.

Wei Liang (SM'04) received the B.S. degree from the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 1998. Then, he went to Institute of Electronics, Chinese Academy of Sciences, Beijing, China, and received M.S. degree there in 2001. He is currently a Ph.D. candidate at North Carolina State University, Raleigh, NC. Since 2001, he has been a teaching assistant in the Department of Electrical and Computer Engineering, North Carolina State University. Then, he took part in the networking group in 2002. His research interests include mobile and secure computing, authentication in wireless networks, quality of service in mobile networks, mobility management, modeling and performance analysis of wireless information networks.

Wenye Wang (M'98/ACM'99) received the B.S. and M.S. degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 1986 and 1991, respectively. She also received the M.S.E.E. and Ph.D degree from Georgia Institute of Technology, Atlanta, Georgia in 1999 and 2002, respectively.

She is now an Assistant Professor with the Department of Electrical and Computer Engineering, North Carolina State University. Her research interests are in mobile and secure computing, quality-of-service (QoS) sensitive networking protocols in single- and multi-hop networks. She has served on program committees for IEEE INFOCOM, ICC, ICCCN in 2004. She has been a member of the Association for Computing Machinery since 2002.