Proposals of Handoff Procedures and Queue Management Methods for Improving Communication Efficiency in Mobile Internet Environments

Doctoral Program in Engineering
University of Tsukuba

2003, March

OHZAHATA, Satoshi

Abstract

This thesis mainly discusses the micro mobility problem in Mobile IP, QoS for mobile Internet environments and TCP in wireless environments. To overcome micro-mobility problem for Mobile IP, a seamless handoff method is proposed. The proposed method has simple mobility prediction control for the mobile node and the registration procedures are previously executed until the mobile node completely changes the connection point to the Internet. The seamless handoff method achieves complete elimination of the communication disruption with changing the connection point. The mobility prediction control is also extended to secure and high quality communications. These methods are also realized by the previous authentication or resource reservation for the next connectable base station. An adaptive handoff algorithm is also proposed for improving network resource utilization. The control is realized by improving typical "threshold-hysteresis algorithm." Finally, a queue management method is proposed for improving TCP performance in wireless environments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Internet is originally developed as data network. With growth of the network, some quality of service (QoS) architectures are being prepared for delivering multimedia traffic, like video and voice. Moreover, computers have become small and been highly increased on their performance, and then many multimedia network services are easily available in the mobile Internet environments. Since the original Internet protocol (IP)[1] assumes that all network nodes are fixed at their locations and does not provide any mobility schemes in the Internet.

Mobile IP[2] is developed to extend IP to the mobile environments. In Mobile IP, the mobile node (MN) can communicate even if it changes its connection point to the Internet. However Mobile IP serves macro mobility in the Internet, there are many important problems for high quality communications in mobile Internet environments. This thesis mainly discusses two subjects, i.e., handoff procedures in Mobile IP and queue management method for TCP in wireless environments, to solve the problems of communication efficiency.

Mobile IP provides macro mobility but does not consider micro mobility. This causes disrupted communications such as real-time audio and video streams during a handoff procedure. To improve the problem, this thesis introduces four improved handoff methods for Mobile IP. The seamless handoff method is proposed to provide the micro mobility and no disrupted communications in the procedures. In this method, the handoff procedure can be completed before the MN changes the current base station (BS), which is a connection point to the Internet, to a new one for eliminating the disrupted communications. This simple mobility predictive scheme is the basis of the following two improved handoff procedures.

When the handoff procedures should include the authentications among the communication entities, the handoff latency becomes longer and causes the large communication disruption. The fast authentication method provides the authentication and key distribution mechanism with extending the seamless handoff method. In this method, the authentication of MNs is also previously completed before the MNs change the current BS and thus the handoff latency can completely be ignored.

The quality of the communications depends not only on the handoff procedure but also the admission control and resource reservation method. Then, an adaptive resource reservation method and an admission control for mobile Internet environments are proposed. In order to serve the reservation method, the registration procedures of the seamless handoff method is extended. For high quality communications, the QoS negotiation is also required between a user and networks before the user starts communications or handoff of communications. Diffserv-based QoS architecture[3] is also proposed for Mobile IP to implement the QoS negotiation.

In the previously proposed handoff algorithms, "threshold-with-hysteresis" is widely used to improve the handoff decision, however, such algorithms are considered only for a specific user. Finally, an improved adaptive handoff algorithm is presented for effectively sharing the network resources for all users. To realize the handoff decision based for MNs, each agent advertisement from the BS includes information of the congestion status at the BS. By the proposed algorithm, the bandwidth of the wireless link is effectively traded for the users among the BSs.

The second problem is that the congestion control protocol of TCP[4],[5] is not suitable for mobile or wireless Internet environments. In such environments, packet losses easily occur in the wireless link and cause unrequired congestion control of TCP. To solve the problem, an improved queue management method is designed to help a local retransmission mechanism of the MAC layer for the wireless link. From the control, any packet loss in the wireless link seems to be completely eliminated and all packets are delivered in the order. No modification is required to TCP itself.

To evaluate these proposed methods, the computer simulation experiments are executed on the network simulator ns[6].

The rest of this thesis is constructed as follows. Chapter 2 briefly introduces Mobile IP. Chapter 3 proposes the seamless handoff method. Based on the seamless handoff method, Chapter 4 provides the fast authentication methods and Chapter 5 presents the adaptive resource reservation. Chapter 6 proposes the adaptive handoff method with extending "threshold-with-hysteresis" algorithm. Chapter 7 describes the queue management method for improving TCP performance in wireless environments. Finally, this thesis is concluded in Chapter 8.

Chapter 2

Mobile IP

This chapter explains Mobile IP and discusses how Mobile IP provides macro mobility in the Internet and occurs its problem for micro mobility.

2.1 Introduction of Mobile IP

Internet protocol (IP) is originally developed to integrate different kinds of media access control networks (data link layer), and thus is placed on the network layer to bridge the connections between these networks (Fig. 2.1).

IP works well when the communication entities do not change their connection point to the Internet. Recently, however, small and powerful portable computers can be easily available and have wireless network devices to communicate through the Internet. When such a computer changes the connection point to the Internet, its IP address is also modified. This means that IP can not provide communications in mobile environments by itself because IP address is used for routing packets in the Internet.

In Fig. 2.2, a mobile node (MN), whose IP address is 130.158.85.10 and subnetwork address is 130.158.85.0/24, stays at its home network. When a server transmits a packet

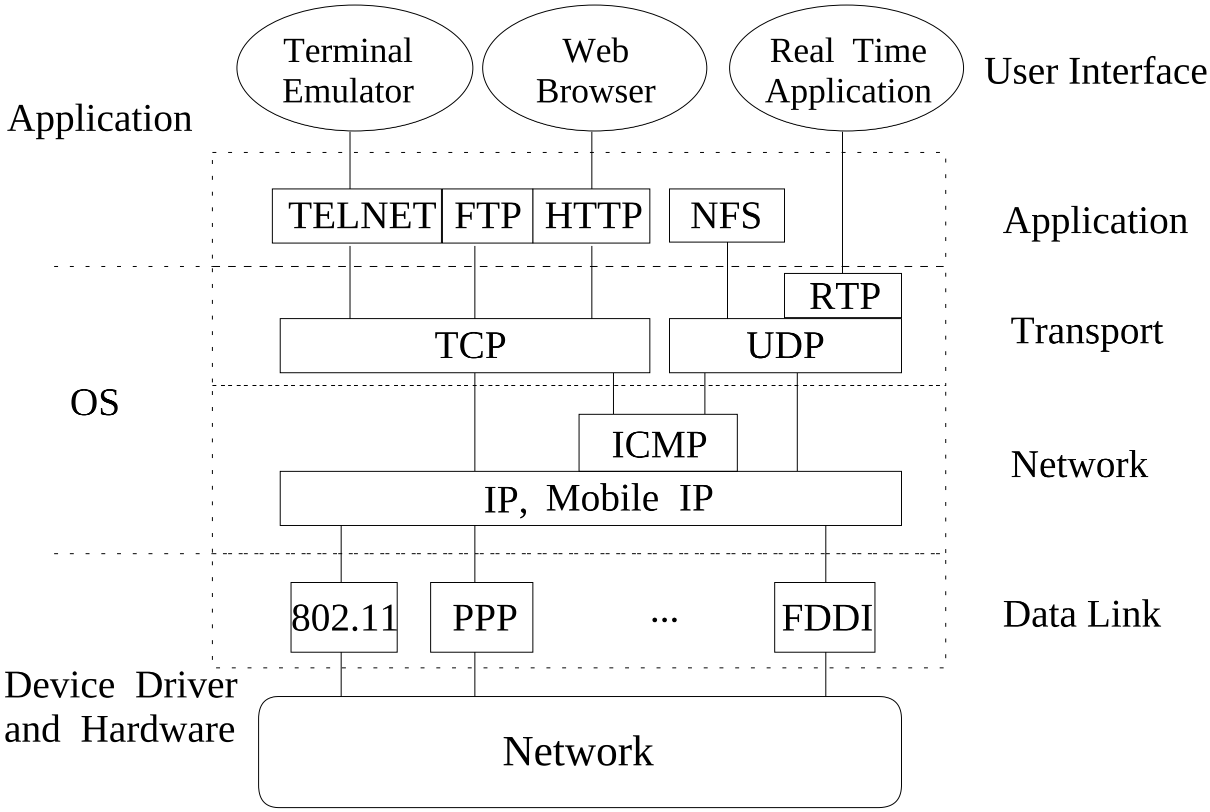

Figure 2.1: Reference model.

to the MN, the packet is routed by using the home subnetwork address and MN's IP address. When the MN changes its connection position to a foreign network whose subnetwork address is 130.69.96.0/24, the MN's IP address must also be modified the one prefixed by the subnetwork address. If the modification of the address is occurred during the communication, the communication is disrupted since the IP address is used as identification for the higher layer to keep the session of the communications.

To provide mobility to the Internet, Mobile IP is introduced. Next subsection shows how Mobile IP provides mobility functions, which is called macro mobility.

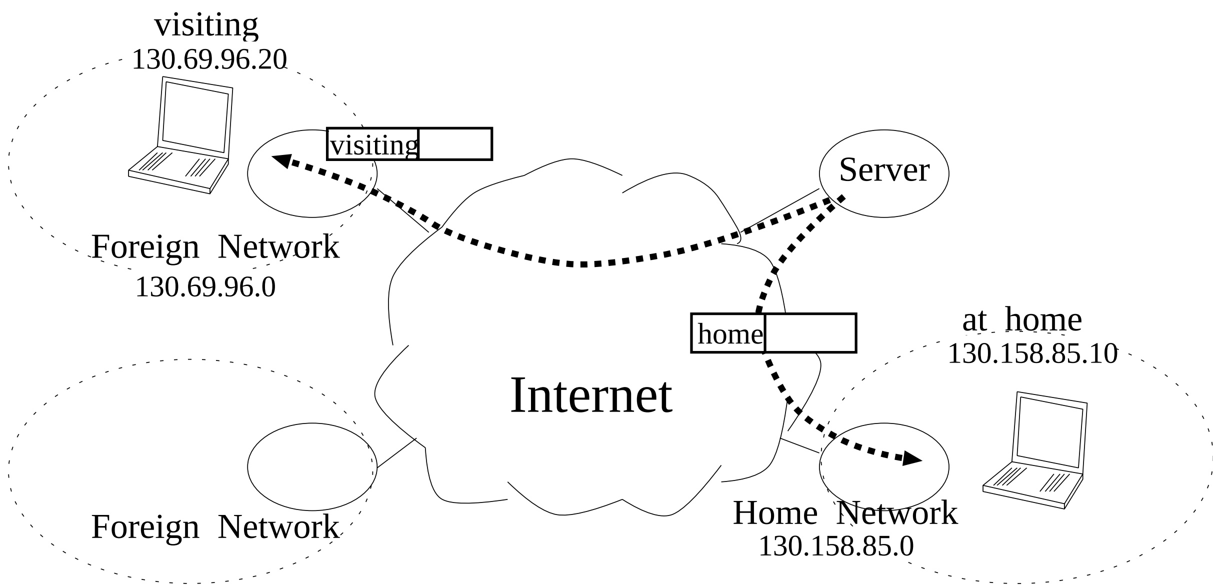

Figure 2.2: IP routing.

2.2 Macro Mobility

Mobile IP works as a mobility and route management protocol in the Internet. Under Mobile IP, MN can change its connection point to the Internet without modification of its home IP address. For this architecture, two kinds of agents, i.e., a home agent (HA) and a foreign agent (FA) are prepared (Fig. 2.3).

The MN has the home network (HN) and is mainly connected to the network. The HN has HA (or HAs) which always manages all MNs of the HN. When packets are transmitted from a sender to the MN in the HN, the MN is accessed by its home IP address and delivered by ordinary IP communication.

Suppose that the MN changes the connection point from the HN to another network which is called a foreign network (FN). The FN has an FA which manages the MN in cooperation with the HA. In the communications, the HA intercepts every packet for the MN, and then encapsulates and forwards the packet to the FA (the care of address). The FA decapsulates the forwarded packets and transmits the original packets to the MN.

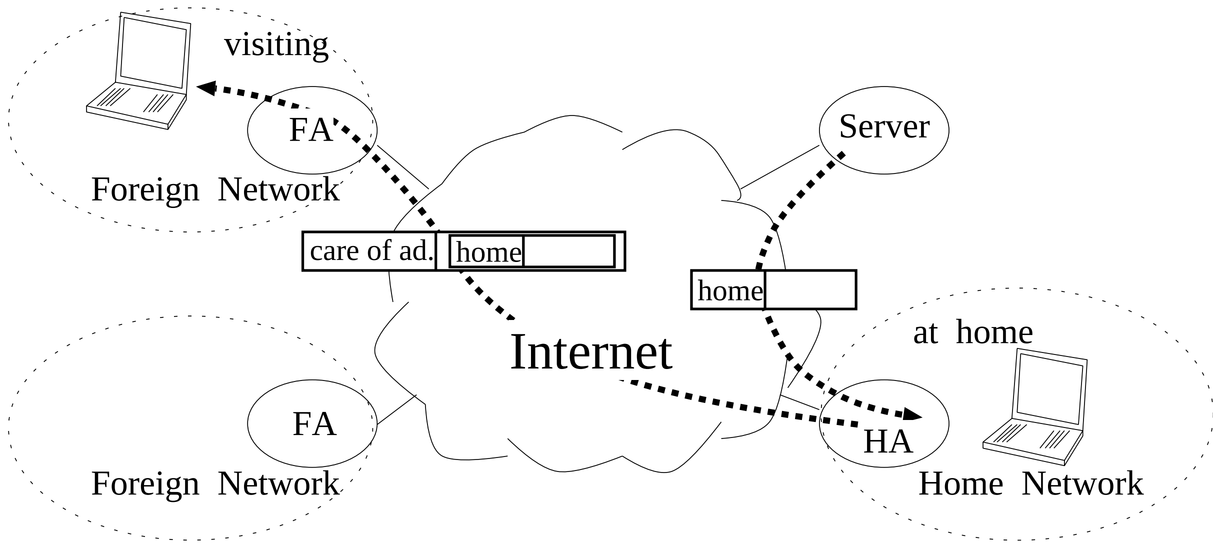

Figure 2.3: Mobile IP routing.

From above the procedures, each MN can transparently communicate via any FNs.

Mobile IP provides macro mobility in the Internet, however, the procedure still causes the handoff latency, because Mobile IP does not consider micro mobility and the route switching control also depends on the MAC media. The next section discusses the handoff procedures in Mobile IP.

2.3 Handoff Procedures in Mobile IP

Fig. 2.4 shows handoff procedures when MN changes its connection point from FA1 to FA2. (1) BS (HA or FA) periodically broadcasts an agent advertisement in the cell. Each agent advertisement has information for the MNs to connect with the BS (care of address, etc). (2) When MN accepts an agent advertisement from FA2, the MN sends the registration request to FA2. (3) FA2 registers the MN and forwards the request to HA. (4), (5) The HA updates the registration and sends a registration reply to the MN via FA2. From the above procedures, the registration is completed and the MN can start

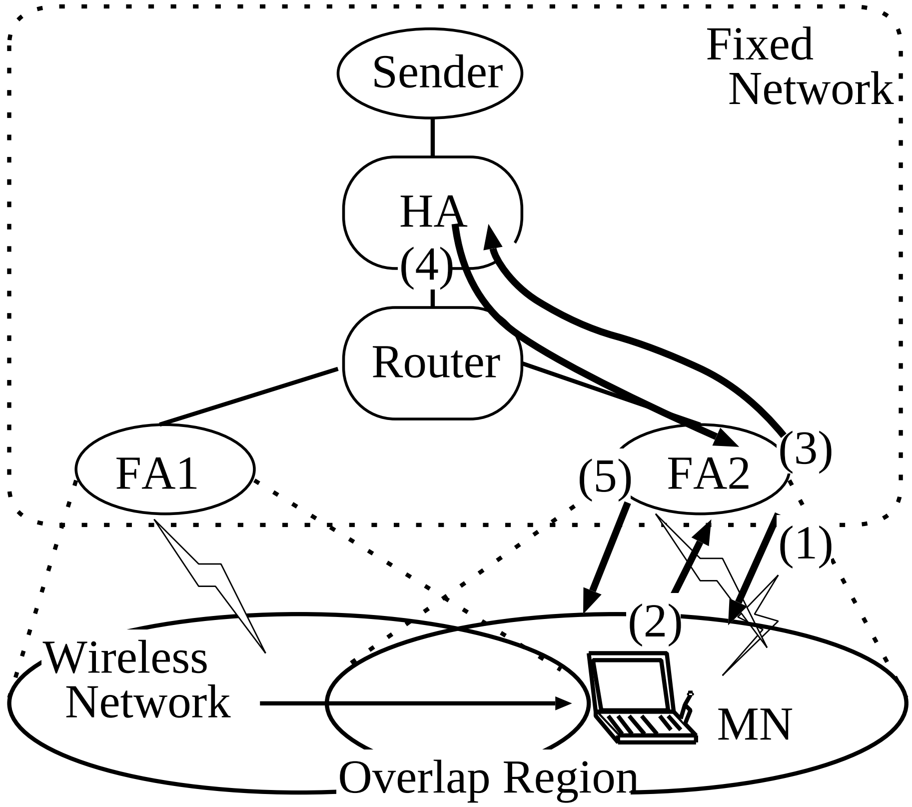

Figure 2.4: Handoff procedures.

communication via FA2, but the procedure also produces disrupted communication in the procedures.

Fig. 2.5 depicts detail of the handoff procedures on communications. This figure shows that MN can not send a registration request to FA2 until the MN has completely released from FA1. Although the MAC layer in the MN can immediately recognize when it is released, the information is not notified to the IP layer (Mobile IP), since the communication between layers is not permitted for independence of each layer. Thus, the Mobile IP must wait for at least three periods of receiving agent advertisements (at least 3 seconds) to confirm the release before the registration request is sent to FA2.

In addition, the registration procedure takes at least a round trip time between the MN and the HA. After the above procedures are completed, the MN can restart the

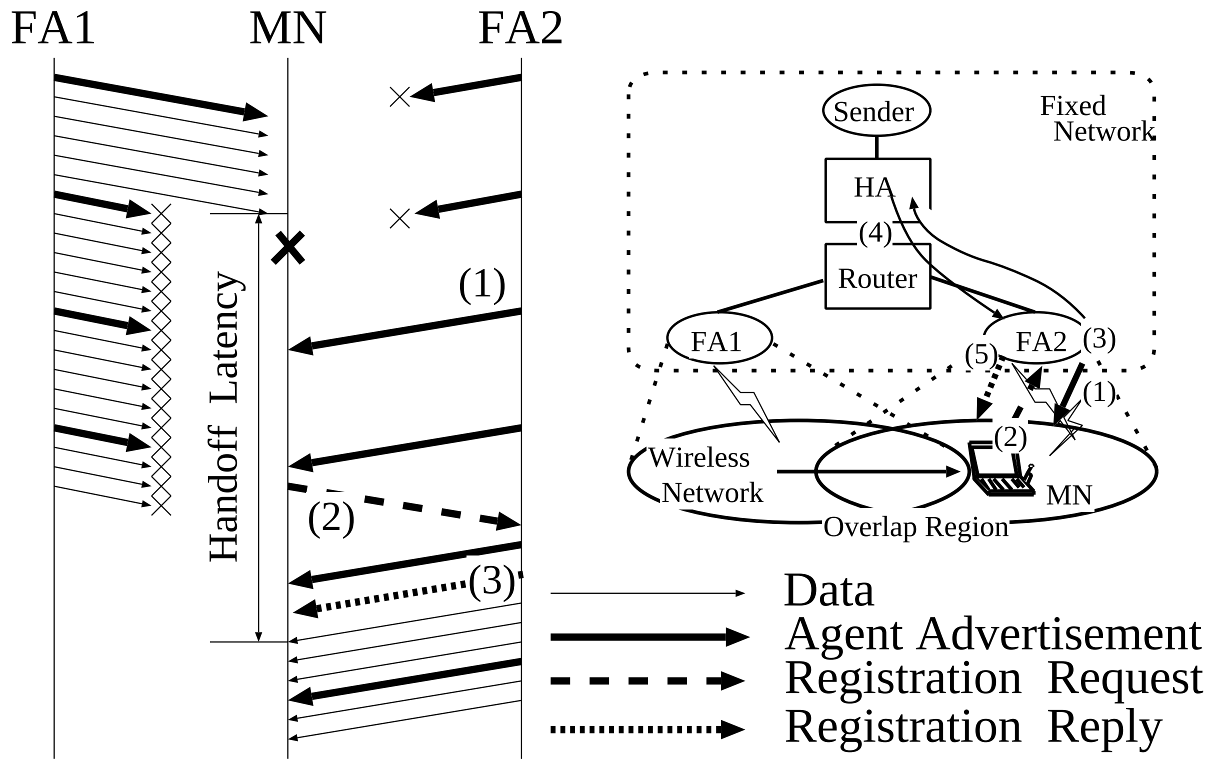

Figure 2.5: Handoff procedures on communications.

communication via FA2.

This handoff latency becomes further long when the procedures have authentications of the communication entities and resource reservation for the communications.

2.4 Conclusion for Chapter 2

Chapter 2 shows how Mobile IP provides macro mobility in the Internet. A packet is encapsulated between the HA and the FA, which is called tunneling, to keep the session of the communications even if the MN changes the communication point to the Internet. However, Mobile IP has a problem that it does not provide micro mobility and the communication must be disconnected each time it changes the connection point to the Internet. These are caused by inconsistence of handoff between layers and handoff procedures in Mobile IP. The next chapter proposes a seamless handoff method to serve no disrupted communications in the handoff procedures.

Chapter 3

Seamless Handoff Method for Mobile IP

This chapter proposes a seamless handoff method to introduce micro mobility in Mobile IP for multimedia communications. In the method, the MN can deal with two suitable communication routes by information from the MAC media. The control serves seamless communications even if the MN frequently changes its current BS.

3.1 Introduction of Micro Mobility

To reduce handoff latency, several micro mobility solutions are proposed. In [7] and [8], the several FAs are prepared for each MN to be able to accept data correctly from one of the FAs. The FAs are pre-defined by using multicast at any time. From the handoff procedures, although the MN can not receive the data from the previous FA, the MN can accept the data as soon as the MN connects to the one of the other FAs which are also buffered the data. However, since HA must send the same data to all the FAs, the above methods waste bandwidth in the fixed networks between the HA and each FA, because it is quite difficult to predict which FA will be selected by the MN. Further more, the FAs must have large data buffer to support many MNs.

The hierarchical BSs[9][10] are proposed to shorten the delay of registration procedures when the MN moves within the same domain or Mobile IPv6[11]. These methods also save extra packets arisen from the registration procedures.

However, the periodically broadcasted agent advertisement from the BS also causes handoff latency because the MN should know the information of the BS at any time in order to shorten the handoff latency. Therefore, in the above methods, the MN watches the link condition or arrival intervals of packets to recognize disconnecting the link, and then sends a registration request when the first time of an agent advertisement from a new cell is received. This method is called a fast handoff method which is discussed in the next section.

The all above methods succeed in shortening a handoff latency or reducing packets losses, however, the handoff latency can not be completely eliminated. One of the important reasons for occurring the handoff latency from the handoff procedures is that Mobile IP allows that the MN, FA and HA deal with only one care of address for the MN. Indeed, the MN must disconnect the current connection before it establishes a new one.

3.2 Fast Handoff Method

For reducing the handoff latency when the MN must have only one care of address and one network interface, it is an important factor how to shorten the procedures. In [7], [8], [9] and [10], the information is notified from the MAC layer to the Mobile IP agent (IP layer) when the MAC media changes the connectable point. This method is generally called a fast handoff method and the procedure is shown in Fig. 3.1.

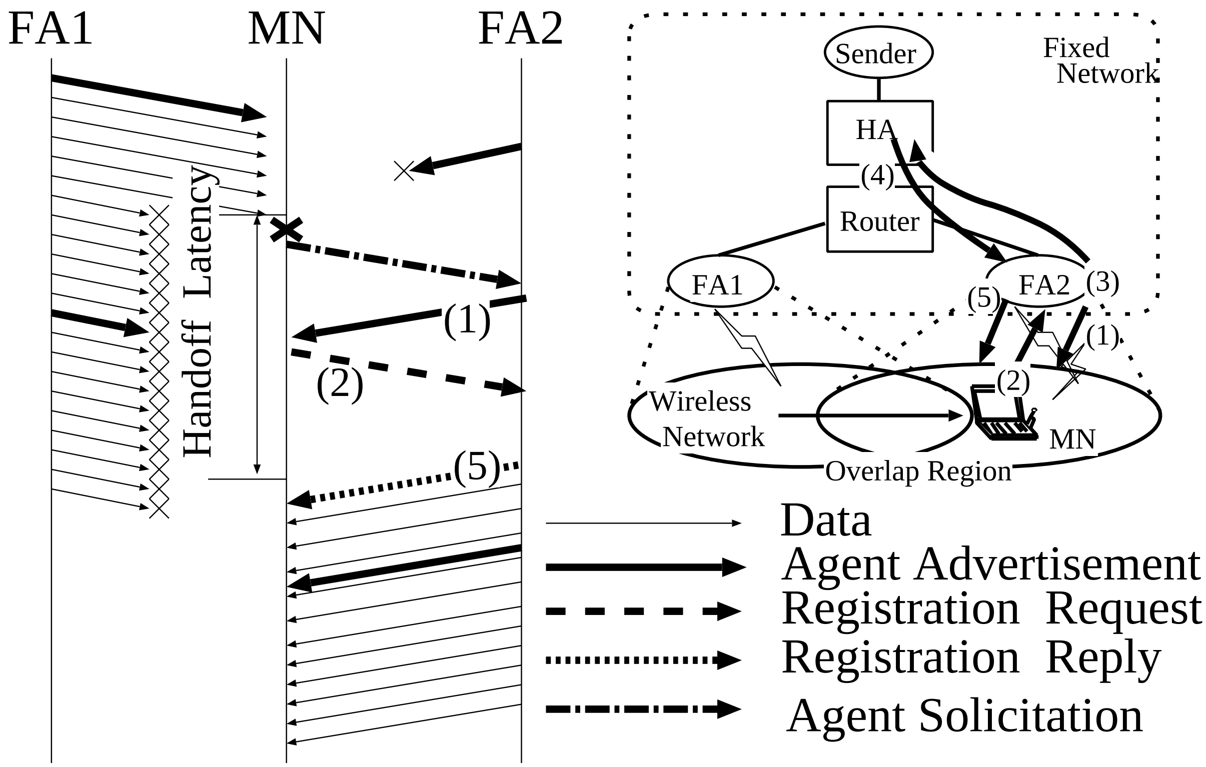

Figure 3.1: Handoff procedures on communication (fast handoff).

In the fast handoff method, when the MAC media in the MN changes the current FA from FA1 to FA2, it informs the change to the Mobile IP agent. Then, the Mobile IP agent can send an agent solicitation to require an agent advertisement from FA2. From the procedures, the MN can send a registration request to FA2 before it waits three periods of receiving an agent advertisement. However, this method can only shorten the above period and the other handoff latency is not avoidable.

3.3 Design of Seamless Handoff Method

To solve the handoff problem, the seamless handoff method allows that the MN has two network interfaces to deal with two care of addresses in each layer, MAC and IP. These care of addresses are called primary and secondary in the following, respectively. Moreover, FA

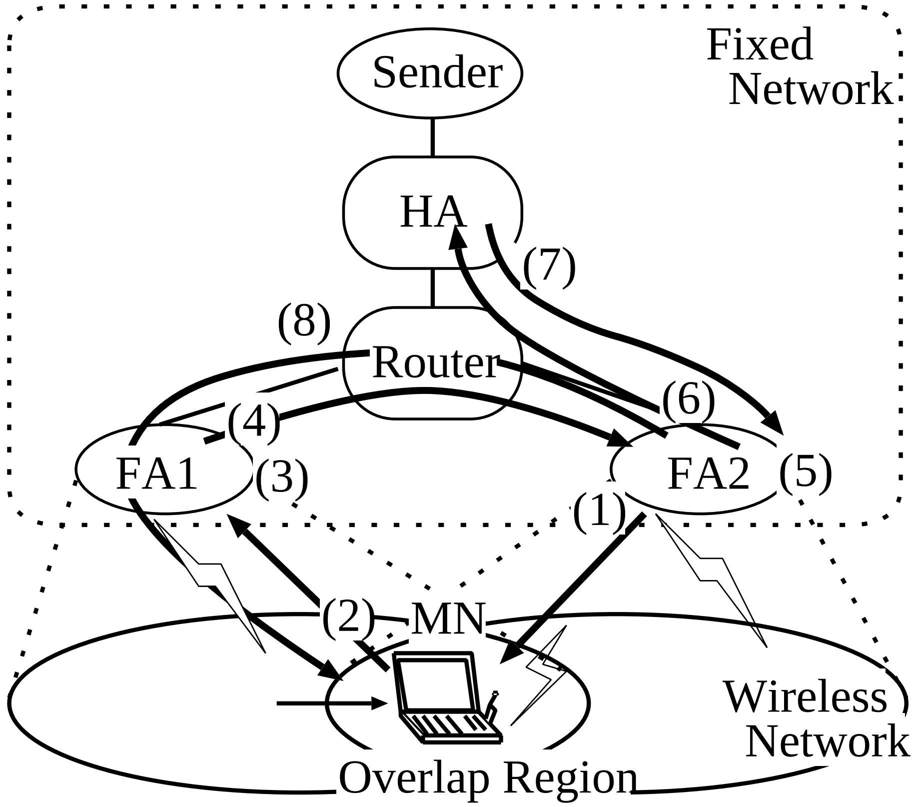

Figure 3.2: An example of routing switching with the seamless handoff method.

and HA can also deal with the two care of addresses to manage the two communication routes between the MN and the HA at IP layer. From these modifications, when the MN receives an agent advertisement from a non-primary FA, it can send a registration request to its HA via the primary FA. Therefore, the MN can complete the registration for the non-primary FA as the secondary FA before it disconnects to the primary FA. The next section shows handoff procedures of seamless handoff method.

3.3.1 Registration Procedures for Seamless Handoff Method

Fig. 3.2 shows a handoff mechanism in the seamless handoff method when the MN changes its connection point from FA1 to FA2. (1) When the MN receives an agent advertisement from FA2 in the overlap region, (2) the MN sends a registration request to FA2 via

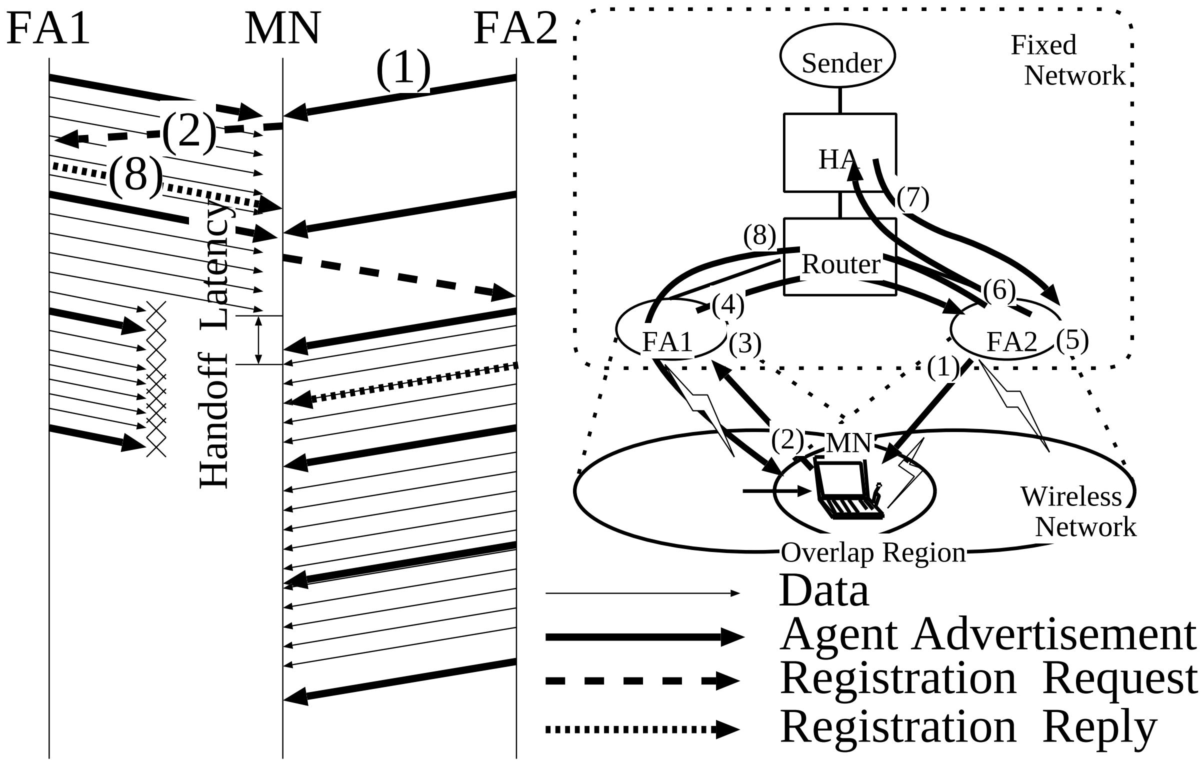

Figure 3.3: Handoff procedures on communication (seamless handoff).

FA1. (3) FA1 registers the request and (4) forwards it to FA2. (5) When FA2 accepts the registration request, (6) it forwards the request to HA. (7) HA registers the care of address of FA2 as secondary one and (8) sends a registration reply to the MN through FA2. Through these procedures the MN can register its secondary FA to the HA, only when the MN receives an agent advertisement broadcasted from FA2.

In the seamless handoff method, FA with stronger signal is called a primary FA, and the other is a secondary FA. the MN sends a registration request to the primary FA rather than to the secondary one, because the request may not reach to the latter FA and the MN can save the power consumption.

Fig. 3.3 shows a data flow diagram during the handoff procedures on communication. Since these registration procedures are completed until the MN changes the primary connection point, the handoff latency is very small. Further more, since the MN can deal two network interfaces, it can receive the communication data from the previous primary FA after changing the primary FA. This fact also makes the handoff latency be much smaller.

However, if no overlap region exists, the MN can not send its registration requests for FA2 (secondary) through FA1 (primary). In this case, the MN must send the registration request to FA2. Next, FA2 requests information to FA1 to confirm the MN's request, and forwards the registration request to the HA if the confirmation is successful. The above registration procedures for no overlap region are almost same with Reference [12] which is proposed for a secure and fast handoff method.

3.3.2 Data Flow Controls in Seamless Handoff Method

As shown in Fig. 3.3, the handoff latency may not be completely eliminated if the communication data can not be delivered soon after the MN changes the primary FA.

For reliable communication, the seamless handoff method uses the control that the HA bicasts the data for the two care of address of the MN. However, the bicast produces large overhead in certain conditions. For example, even if no packet loss occurs between the MN and FAs, the primary and secondary FAs forward the same data to the MN. To reduce this overhead, the secondary FA does not forward them until it directly receives a registration request from the MN in the proposed method.

Each data packet has information to find which FA should forward the packet or not. The information is constructed by the HA and the FA, and inserted to each packet's header. When the MN observes any packet loss from the primary FA or that the signal strength from the primary FA is weaker than the one from the secondary FA, it sends a registration request to the secondary FA to inform the MN's state, and the both FAs start forwarding the packets to the MN. This request of the MN is notified from the secondary FA to the primary FA via the HA so as to exchange the route of primary and secondary FAs. In the method, the secondary FA always waits the registration request to start forwarding the packets, however, these data are not buffered in it since these are real-time communication data.

Suppose the MN can not receive any packets from the both primary FA and secondary FA. Then the MN needs an agent advertisement from the other FAs to recognize them, however, it may not always be able to obtain as soon as it want, since the agent advertisements from the FAs are broadcasted periodically. To overcome this problem, a fast agent solicitation method is also proposed in the seamless handoff method. In this method, the MN is allowed to send an agent solicitation, when the MN detects a lost packet in spite that it has more than one valid care of address. From the above controls, the seamless handoff method can provide no disrupted communications with low overhead at the wireless link even if the MN randomly moves through multiple overlap regions.

3.4 Computer Simulation Experiments

To evaluate the seamless handoff method in the previous section, computer simulation experiments are executed by implementing the proposed methods with network simulator ns-2.1b7-current (May/15/2000)[6] and show that the seamless handoff method provides no handoff latency with low overhead in a wireless network on constant bit rate (CBR) traffic over real-time transport protocol (RTP).

3.4.1 Simulation Conditions

Two handoff methods are simulated in this section, i.e., the seamless handoff method in the previous section and the fast handoff method in Section 3.2.

Simulation experiments are executed on two network models. Each model consists of fixed network and wireless network whose topology and link speed are illustrated in Figs. 3.4 and 3.7. The common conditions for all simulations are shown in below.

[Traffic Conditions]

- The number of senders is one.
- The number of MNs in each model is also one.
- The sender sends data to the MN by CBR traffic model over RTP. The IP packet size is 256 octets whose transfer rate is 384Kbps and then the packet sending interval is 5.33ms.

[Topological Conditions]

- Media access control at the wireless network in Figs. 3.4 and 3.7 is IEEE 802.11 implemented in [6].
- For wireless communication between the MN and FA in each cell, transfer error is ignored and propagation delay is calculated as (distance between the MN and FA) / (speed of the light).

[General Conditions]

- An agent advertisement from each FA is periodically broadcasted every 0.2s within the whole cell area of the FA, including the overlap region of both cells, and the MN's registration life time[2] in each FA is 1.0s.

- Each authentication delay at the registration requests in the FAs and HA is ignored.

- In the overlap region, the MN can receive packets from the both FAs in the seamless handoff method.

- In the two methods, the MN can send an agent solicitation or a fast agent solicitation message, if it detects packet loss or more than 10ms of packet arrival interval.

- In the fast handoff method, the MN's MAC media is not allowed to deregister the FA for 0.1s after the MN registers the FA.

- In the seamless handoff, the MN's MAC media is not allowed to deregister the secondary FA for 0.1s after it registers the FA by the latest agent advertisement from the FA.

- The processing delay in Logical Link Control of the FAs is 25μs.

- In the MN, HA, FAs and router, each processing delay for routing, copying, encapsulating and decapsulating a packet is 10μs, respectively. The one of the registration procedure for the MN is 100μs.

- Every experiment result satisfies 5% confidence interval with 95% confidence level.

The additional common conditions for simulation model 1 (Fig. 3.4) are shown in below.

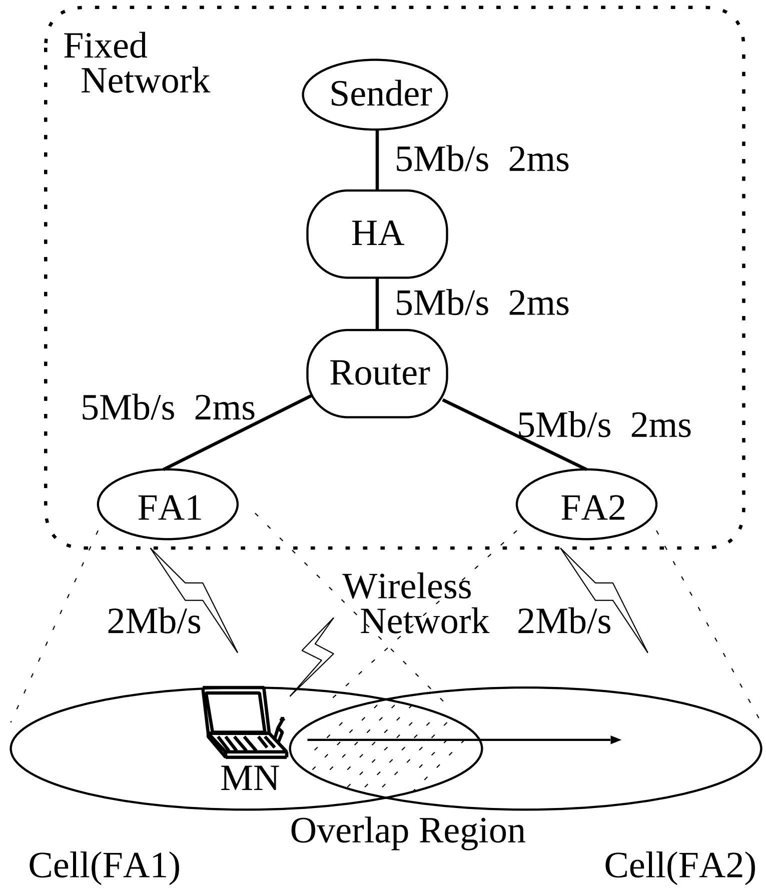

Figure 3.4: Simulation model 1.

- These are one HA and two FAs. Each FA has a circular cell.
- Each FA is placed at the center of the cell, and 360m apart from each other. The MN moves straight forward from FA1 to FA2.

3.4.2 Overlap Model

At first, the radius of each cell is set to 200m and then the simulation model has an overlap region. Fig. 3.5 indicates the average handoff latency (i.e., the average packet arrival intervals when the MN switches the primary FA from FA1 to FA2) in the CBR communication, where the two handoff methods are described as "seamless" and "fast," respectively. In the rest of this section, all simulation results are also described in this

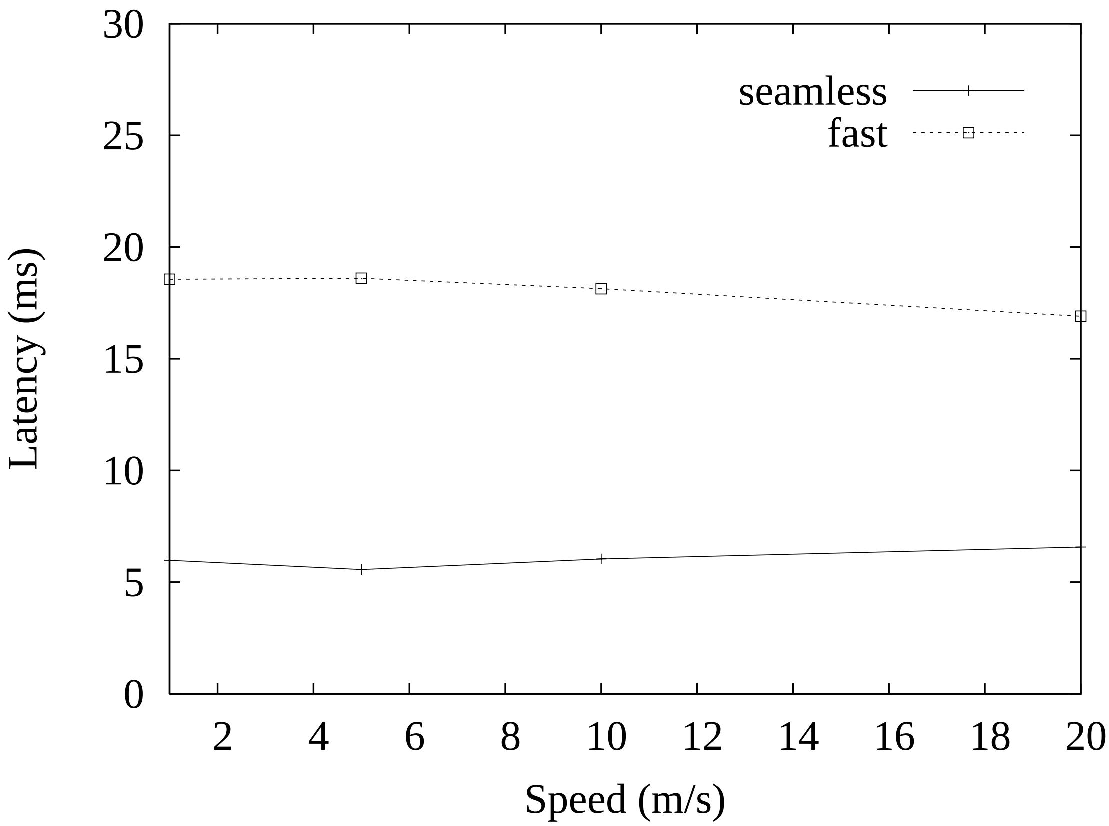

Figure 3.5: Average handoff latency (overlap model).

order. In this figure, the average handoff latencies of the two handoff methods are 5.6–6.6ms and 16.9–18.6ms. From the results, the handoff latency is almost unchange for the speed of the MN. In the simulation conditions, 20m/s (= 72km/h) is not quite fast to receive an agent advertisement from FA2 within the overlap region. Thus, the MN can complete the registration procedures for FA2 in the overlap region before it is completely released from FA1. Therefore, the handoff latency does not depend on the speed of the

Table 3.1: Average overhead of the seamless handoff method (overlap model).

	1m/s	5m/s	10m/s	20m/s
copied in HA	7161.3	1611.0	897.5	561.3
duplicated in wireless	15.4	16.0	14.7	16.0
duplicated in MN	14.4	15.0	13.7	15.0
lost	0.0	0.0	0.0	0.0

Table 3.2: Average overhead of the fast handoff method (overlap model).

	1m/s	5m/s	10m/s	20m/s
copied in HA	0.0	0.0	0.0	0.0
duplicated in wireless	0.0	0.0	0.0	0.0
duplicated in MN	0.0	0.0	0.0	0.0
lost	3.4	3.4	4.3	3.1

MN in this simulation. It is also emphasized that these results show the communication in the seamless handoff method is not interfered by the handoff procedure, since the sending packet interval of the sender is 5.33ms and each FA also produces another latency to periodically broadcast an agent advertisement.

The average handoff latency of the fast handoff method is at least 10ms lager than the seamless handoff method because the MN can not receive data from the HA during the handoff procedures which take about one round trip time between the MN and the HA, and the data are transfered into FA1 not to FA2 during the period.

Tables 3.1 and 3.2 show the average overhead (i.e., the average of the total numbers of extra or loss packets) for the two methods in the simulation. “copied in HA” is the average number of duplicated packets by bicasting from the HA. “duplicated in wireless” is the average number of packets which are included in “copied in HA” and also flowed into the wireless network. “duplicated in MN” is the average number of packets received by the MN as duplication.

The average packet losses of the both methods are 0.0 and 3.1–4.3 packets at “lost” in the tables, respectively. The seamless handoff method realizes no packet loss, since the MN can receive all packets from FA2 when it is departed from the overlap region.

“duplicate in wireless” in Table 3.1 arises for eliminating packet loss and carrying all packets in the correct order when the primary FA and secondary FA are exchanged.

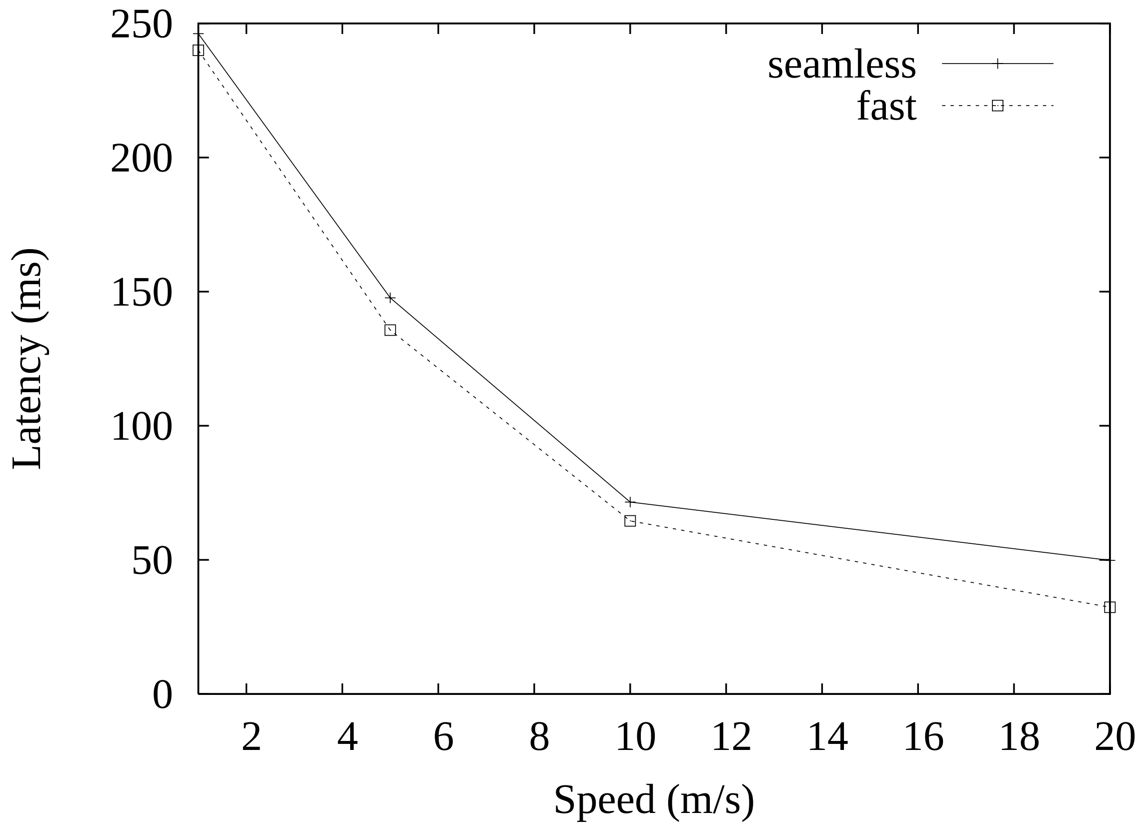

Figure 3.6: Average handoff latency (no overlap model).

561.3–7161.3 packets of the average overhead are also observed as “copied in HA,” and then these packets are flowed from the HA into the fixed network during 3.0–38.8s, which is correspond to the time when the MN stays in the overlap region.

Table 3.3: Average overhead of the seamless handoff (no overlap model).

	1m/s	5m/s	10m/s	20m/s
copied in HA	188.0	194.9	226.5	223.1
duplicated in wireless	185.3	172.9	186.5	185.0
duplicated in MN	0.0	0.0	0.0	0.0
lost	46.1	27.3	14.4	9.3

Table 3.4: Average overhead of the fast handoff method (no overlap model).

	1m/s	5m/s	10m/s	20m/s
copied in HA	0.0	0.0	0.0	0.0
duplicated in wireless	0.0	0.0	0.0	0.0
duplicated in MN	0.0	0.0	0.0	0.0
lost	45.0	25.4	13.1	6.0

3.4.3 No Overlap Model

The same experiments in the previous subsection are executed, however, the radius of the each cell is 180m, i.e., there is no overlap region in this subsection. Fig. 3.6 shows that the average handoff latencies of the two methods are 49.8–246.1ms and 32.3–240.0ms, respectively. Since there is a small disconnected area from the both FAs, the faster speed of the MN, the lower average handoff latency. The seamless handoff method results about 10ms larger average handoff latency than the ones of the fast handoff method, since no overlap region gives extra propagation delay between FA1 and FA2 in the procedures.

Tables 3.3 and 3.4 show the average overhead for the two methods in the simulation. "lost" in Table 3.3 is a little larger than the ones in Table 3.4. In Table 3.3, the seamless handoff produces 188.0–226.5 packets of "copied in HA," i.e., flows these extra packets during 1.0–1.2s, in spite that there is no overlap region. This is because the HA recognizes FA1 as a valid FA during the registration life time of FA1, and then continues to forward packets to FA1 after the registration from FA2 is completed.

3.4.4 Multiple Overlap Model

Fig. 3.7 shows the simulation model 2 with a multiple overlap region where FA3 is added to Fig. 3.4, and the MN also moves straight forward from FA1 to FA2. The common conditions for the simulation model 2 are shown in below.

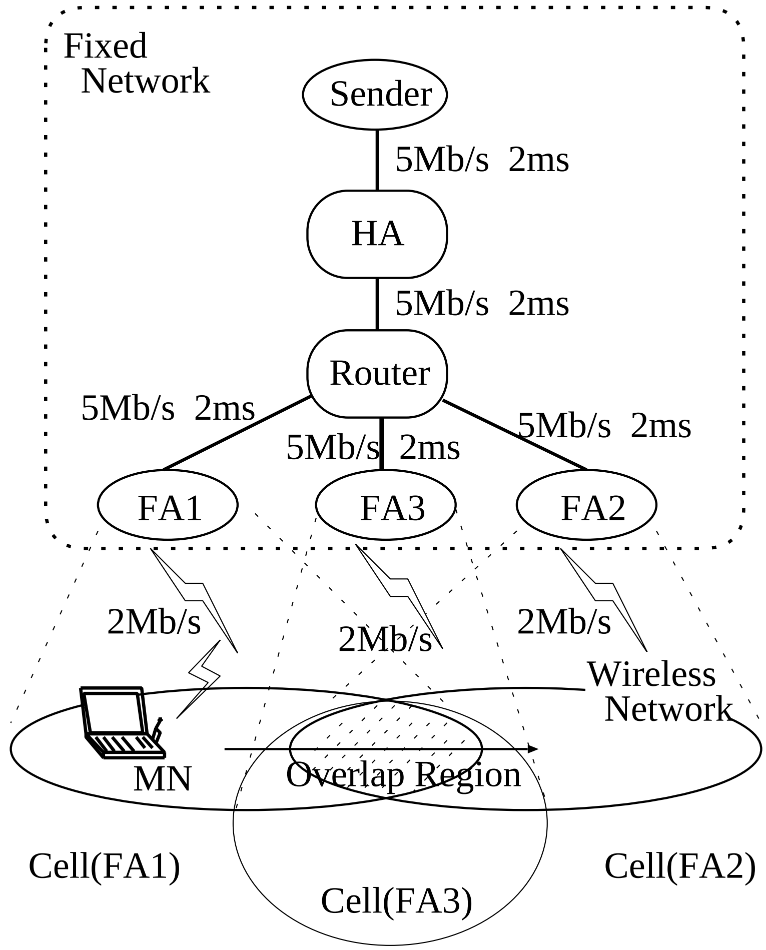

Figure 3.7: Simulation model 2.

- FA3 is placed in the same distance from FA1 and the FA2, and 180m apart from the midpoint between FA1 and FA2.
- The radius of the cell in each FA is 200m.

Fig. 3.8 depicts the average handoff latencies, and Tables 3.5 and 3.6 show the average overheads of the two method. From the figures, the handoff latencies are 5.3–5.3ms and 18.3–20.7ms, respectively, which are almost same with Fig. 3.5. However, from the tables, "copied in HA" of the seamless handoff method is about three times of the results in Table 3.1. Since the overlap region expanded by the cell of the FA3, extra packets are flowed into fixed networks during 9.3–168.5s. "lost" is not changed from Tables 3.1 and 3.2 in the

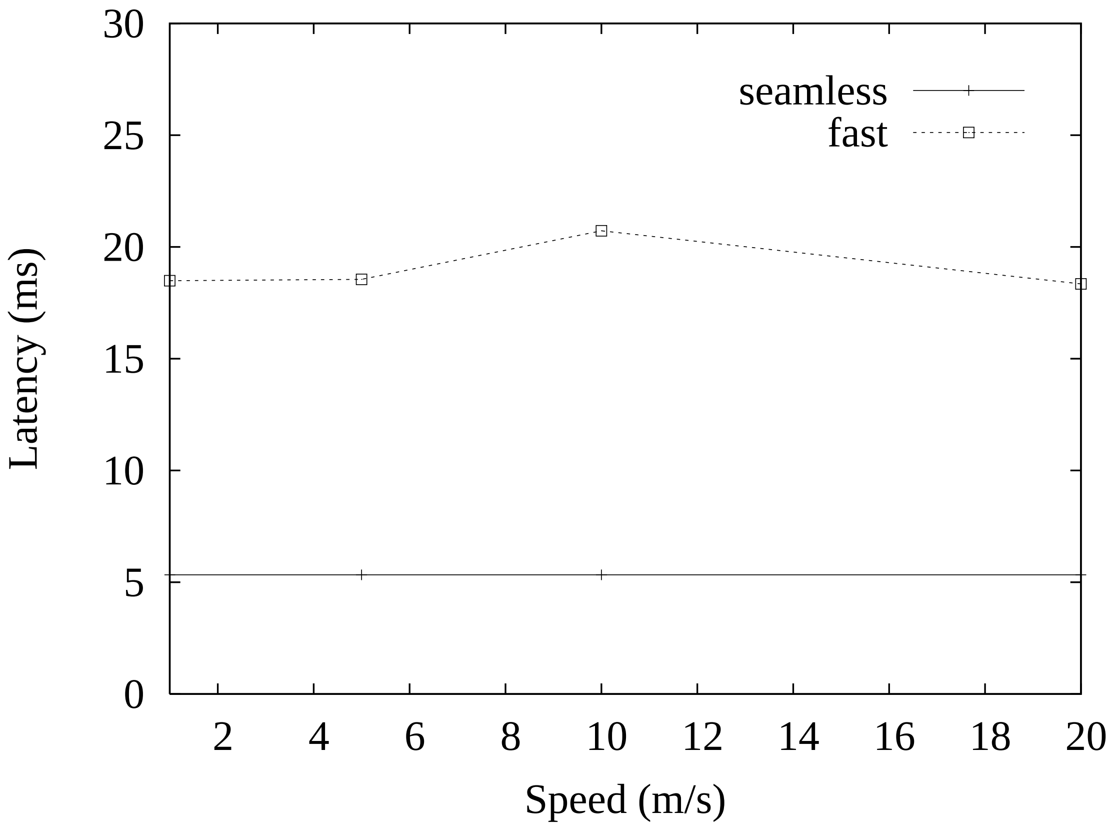

Figure 3.8: Average handoff latency (multiple overlap model).

two methods, though about two packets are increased at 10 and 20m/s in the fast handoff method. The reason of the latter result is that the MN can not switch the communication route correctly with the timing of receiving the agent advertisement from FA2 or FA3 in case that the MN moves fast. These results also show that the seamless handoff method provides seamless data stream with low overhead in wireless networks, even if the MN moves through the multiple overlap region.

Table 3.5: Average overhead of the seamless handoff method (multiple overlap model).

	1m/s	5m/s	10m/s	20m/s
copied in HA	31600.3	6475.0	3362.9	1750.4
duplicated in wireless	12.7	12.7	23.6	23.3
duplicated in MN	10.9	11.9	21.6	22.3
lost	0.0	0.0	0.0	0.0

Table 3.6: Average overhead of the fast handoff (multiple overlap model).

	1m/s	5m/s	10m/s	20m/s
copied in HA	0.0	0.0	0.0	0.0
duplicated in wireless	0.0	0.0	0.0	0.0
duplicated in MN	0.0	0.0	0.0	0.0
lost	3.4	3.4	6.7	5.3

Table 3.7: Results of random motion in multiple overlap region (average).

	fast	seamless
handoff latency (ms)	24.8	5.6
received in HA	18750.0	18750.0
copied in HA	0.0	18750.0
duplicated in wireless	0.0	904.6
duplicated in MN	0.0	902.1
lost	134.8	0.0

3.4.5 Random Motion in Multiple Overlap Region

In the previous subsections, the MN moves straight at constant velocity, however, actual motion of the MN is not so simple. In this subsection, the MN randomly moves within a square of 400m^2 at the multiple overlap region in Fig. 3.7. Additional conditions to the common ones in the previous subsection are shown in below.

- The center of the square is placed at the midpoint between FA1 and FA2. The top and bottom sides of the square are parallel with the line connected FA1 with FA2.
- The initial location of the MN is randomly chosen within the square. Then, the random point is selected within the square. The MN straightly moves to the point with a constant speed randomly selected among 0.0–28.3m/s during 1 second. The above random movement is repeated for 100s without any interval.

Table 3.7 shows the simulation results for the random motion model. "handoff latency" means the average of the maximum handoff latency in each trial. "received in HA" is the average of the total number of packets from the sender for the MN received at the HA. For the seamless handoff method, there is no packet loss even if the handoff procedures occur very often. In fact, more than 30 handoff procedures produce 134.8 of the average lost packets for 100s in the fast handoff method, though, in the seamless handoff method, the average overhead in the wireless networks is increased from the results in Table 3.5. The duplicated packets in the seamless handoff method are flowed during all the simulation time because the MN always stays in the overlap region.

3.5 Conclusions for Chapter 3

In mobile communications, a handoff latency occurs when an MN changes the connection point. Mobile IP can not avoid large handoff latency since it only serves macro mobility. For the solution, this chapter proposes the seamless handoff method to provide one of the micro mobility solutions for ordinal Mobile IP. In the method, the registration procedures are completed until the MN moves from the current FA into the other FA. The simulation experiments show that the proposed registration method is effective for real-time communications in cellular Internet environments.

In the seamless handoff method, the MN can predict which FA will be connectable in the handoff procedures. This simple mobility prediction is also effective for the authentication among the communication entities and network resource reservation for the communications. These methods are discussed in Chapter 4 and Chapter 5, respectively.

Chapter 4

Fast Authentication Method for Mobile IP

In Mobile IP, MN is required authentications to identify genuine users of the network whenever the MN changes its connection point to the Internet. This chapter proposes a fast authentication method to add authentication and key distribution mechanisms to the seamless handoff method in the previous chapter. From the method, MNs can securely switch the communication route without any communication disruption.

4.1 Introduction

Mobile IP gives one of the solutions for mobile communications instead of IP, but the mobility requires authentications to identify genuine users of the network whenever an MN changes its connection point to the Internet. To satisfy the requirements for security considerations, a symmetric-key-based authentication protocol is widely used because these protocols are simple and require low computation power to the MN. However, in such methods, the keys for the authentication must be previously distributed to each communication entity and a large number of keys are required since these keys must be prepared for all the pairs of communications. This fact restricts scalability of mobile communications.

To improve the scalability problems, a public key infrastructure (PKI) is proposed. To support the infrastructure, several protocols are also proposed to provide secure key distribution mechanism through the Internet[13],[14]. These protocols are also extended to support scalable and secure communications for mobility of the MNs[15],[16],[17]. However, in the extended protocols, their authentication procedures for the MN require a quite long period. This fact causes disrupted communications in the handoff procedures.

This chapter proposes a fast authentication method as a key distribution and authentication protocol with a simple movement prediction control extended from the seamless handoff method. By the control, the MN can send its registration request through the current FA to the predicted FA to which the MN will connect. This registration procedure can be completed before the MN is disconnected from the current FA. The control can also be extended to the public-key-based authentication and gives scalable communications for MNs with no disrupted communications in the handoff procedures.

4.2 Authentication Method in Mobile IP

For secure communications, Mobile IP prepares MN-HA, MN-FA and FA-HA authentication extensions in the registration procedures[2]. By the extensions, each node can authenticate between MN-HA, MN-FA and FA-HA, respectively. In Mobile IP, the default authentication algorithm is keyed-MD5[18] with the “prefix+suffix” mode. The secret shared keys are previously distributed to the entities because Mobile IP does not have any key distribution mechanism by itself. The authentication protocol (Protocol 1) is written as below.

[Protocol 1]

(1) FA $\rightarrow$ MN: AgentAd

(2) MN $\rightarrow$ FA: RegReq, $h(\text{RegReq})_{k_{\text{FM}}}$, $h(\text{RegReq})_{k_{\text{HM}}}$

(3) FA $\rightarrow$ HA: RegReq, $h(\text{RegReq})_{k_{\text{FH}}}$, $h(\text{RegReq})_{k_{\text{HM}}}$

$k_{\text{FM}}, k_{\text{HM}}, k_{\text{FH}}$: are shared secret keys between FA-MN, HA-MN and FA-HA, respectively.

$h(m)_k$: is a hashed value of m by a one-way function h, and encrypted by a key k.

AgentAd: is an agent advertisement.

RegReq: is a registration request.

In the protocol, time stamp or nonce is omitted. □

In the Protocol 1, (1) FA periodically broadcasts an agent advertisement which has information for MN to connect to the FA. (2) When the MN receives the agent advertisement, it sends a registration request with authentication extensions ($h(\text{RegReq})_{k_{\text{FM}}}$, $h(\text{RegReq})_{k_{\text{HM}}}$) to the FA. The FA authenticates the MN by comparing the first part of the extension and registers it. (3) The FA forwards the registration request with $h(\text{RegReq})_{k_{\text{FH}}}$ and $h(\text{RegReq})_{k_{\text{HM}}}$ to the HA. The HA authenticates the FA and MN with the same way in (2), and registers a care of address for the MN. The registration reply from the HA is delivered to the MN via the FA and then the authentication of the MN is completed.

Protocol 1 can provide one of the solutions for the secure communications. However, it still suffers from scalability of the communications, because the secret shared keys should be previously distributed to the nodes and Mobile IP does not have any key distribution mechanism by itself. Moreover, when the number of nodes is m in the Internet, $m \times m/2$

keys are required to communicate each other. This also causes difficulty for the key management and scalability.

4.3 Fast Authentication Method

4.3.1 Introduction of Certificate-Based Authentication Method for Mobile IP

To be free from the scalable problem in the previous section, several key distribution protocols are proposed for mobile networks[15],[16],[17]. Especially, a certificate-based security protocol[17] is effective to exchange a session key through networks and authenticate each communication entities such as HAs, FAs and MNs. In the protocol, a certificate authority (CA) plays an important role. The CA is a widely trusted agency and distributes its public key to all communication entities. It also signs a certification including a public key of each communication entity by its secret key. Each entity can verify the certification by decrypting itself with the CA's public key.

The concept of the above protocol can be extended to Mobile IP and shows one of the examples as Protocol 2 with the following basic principles.

- Security considerations against the attack are same with Protocol 1.
- No additional message for the original Mobile IP should be required.
- The Diffie-Hellman (DH) key exchange algorithm is used to create a session key for the registration updates.
- MN and HA previously obtain the certification each other.

- The MN, HA and FA have a PKI infrastructure like in [15] to be able to obtain certificated public keys from CA and manage them. They also previously obtain the certification of the CA.

Protocol 2 in below ignores time stamp or nonce.

[**Protocol 2**]

(1) FA $\rightarrow$ MN: AgentAd, $h(\mathrm{AgentAd})_{K_{\mathrm{FA}}^{-1}}$, $(y_{\mathrm{FA}}, \alpha, N)$, $h(y_{\mathrm{FA}}, \alpha, N)_{K_{\mathrm{FA}}^{-1}}$, $\mathrm{Cert}_{\mathrm{FA}}$

(2) MN $\rightarrow$ FA: RegReq, $h(\mathrm{RegReq})_{K_{\mathrm{MN}}^{-1}}$, $(y_{\mathrm{MN}})_{K_{\mathrm{MN}}^{-1}}$, $\mathrm{Cert}_{\mathrm{MN}}$

(3) FA $\rightarrow$ HA: RegReq, $h(\mathrm{RegReq})_{K_{\mathrm{MN}}^{-1}}$, $h(\mathrm{RegReq})_{K_{\mathrm{FA}}^{-1}}$, $(y_{\mathrm{MN}})_{K_{\mathrm{MN}}^{-1}}$,

$(y_{\mathrm{FA}}, \alpha, N)$, $h(y_{\mathrm{FA}}, \alpha, N)_{K_{\mathrm{FA}}^{-1}}$, $\mathrm{Cert}_{\mathrm{FA}}$

(4) HA $\rightarrow$ FA: RegRep, $h(\mathrm{RegRep})_{K_{\mathrm{HA}}^{-1}}$, $(y_{\mathrm{HA}})_{K_{\mathrm{HA}}^{-1}}$, $\mathrm{Cert}_{\mathrm{HA}}$

(5) FA $\rightarrow$ MN: RegRep, $h(\mathrm{RegRep})_{K_{\mathrm{HA}}^{-1}}$, $(y_{\mathrm{HA}})_{K_{\mathrm{HA}}^{-1}}$

$\mathrm{Cert}_{\mathrm{A}}$: is a certification of A which includes K_{A}, ID_{A}, $\mathrm{Date}_{\mathrm{A}}$ and $h(K_{\mathrm{A}}, \mathrm{ID}_{\mathrm{A}}, \mathrm{Date}_{\mathrm{A}})_{\mathrm{CA}}$, where K_{A} is a public key of A, and ID_{A} and $\mathrm{Date}_{\mathrm{A}}$ are identification and expiration date of the certification, respectively.

K_{A}^{-1}: is a secret key of A.

N: is a prime number.

α: is a primitive element modulo of N, i.e., $\gcd(N-1, \alpha) = 1$.

$y_{\mathrm{A}} = \alpha^{x_{\mathrm{A}}} \bmod N$: is an A's public value for DH key exchange, where x_{A} is an A's secret value for the DH key exchange.

$(m, n, \ldots)_k$: is an encrypted value of the sequence m, n, $\ldots$ by a key k.

RegRep: is a registration reply. $\square$

In Protocol 2, (1) the FA creates α and N, and a public value $y_{\text{FA}} = \alpha^{x_{\text{FA}}} \bmod N$. The MN can authenticate the FA by decrypting the message authentication code $h(\text{AgentAd})_{K_{\text{FA}}^{-1}}$ with the public key K_{FA} in the certification Cert_{FA} from the agent advertisement and makes a session key $k_{\text{FM}} = {y_{\text{FA}}}^{x_{\text{MN}}} \bmod N = \alpha^{x_{\text{FA}} x_{\text{MN}}} \bmod N$ between the FA and MN. (2) The MN sends a registration request with DH public value $y_{\text{MN}} = \alpha^{x_{\text{MN}}} \bmod N$ and its certification Cert_{MN}. The FA authenticates the MN with the same way in (1), and also creates a session key $k_{\text{FM}} = {y_{\text{MN}}}^{x_{\text{FA}}} \bmod N = \alpha^{x_{\text{FA}} x_{\text{MN}}} \bmod N$. (3) The FA forwards the registration request with y_{MN}, y_{FA}, α, N and its certification Cert_{FA}. Note that α and N are shared among the MN, FA and HA. The HA creates session keys k_{HM} and k_{FH} between HA-MN and FA-HA with the same way in (1), and authenticates the FA. (4) The HA sends a registration reply with public value y_{HA} of DH for the FA, MN and its certification Cert_{HA}. The FA can authenticate the HA and create a session key k_{FH} with the same way in above. (5) The FA forwards the registration reply to the MN. Then, the MN can create a session key k_{HM}. From these procedures, the session keys for the registration updates are securely distributed.

Above method succeeds in providing scalable communications for the MN but, in general, public-key-based authentication takes longer time than the one of a symmetric key. Moreover, in handoff procedures, the MN must release its current FA to register itself to the new FA and the new care of address to the HA, and then the MN can not communicate with any FA. This fact arises large disrupted communication time. Next subsection introduces a fast authentication method which conceals the large disruption time.

4.3.2 Key Distribution and Authentication Mechanism

This subsection introduces the fast authentication method to conceal handoff latency with authentication of the MN. The method is extended from the seamless handoff method, and then MN, FAs and HA are assumed to be able to deal with two care of addresses of the MN. Therefore, the MN has two network interfaces and can set up a new communication route before it completely release the old care of address. The basic principles of the fast authentication method are same as Protocol 2, and in addition, the following principles are added.

- The key distribution and the authentication are completed before the MN changes the current FA (FA1).
- The MN can register its new FA (FA2), even if the MN stays in an area where a registration request from the MN is not reached to FA2 by its weak signal strength, but the MN can receive an agent advertisement from FA2.
- Protocol 2 is used only for the first time registration request to create session keys which is used at later of registration updates by Protocol 1, and execute the first time authentication among the entities.
- Periodical agent advertisements contain no certification not to waste bandwidth.

Fig. 4.1 shows procedures of the fast authentication method. The method is added two new messages before Protocol 2(1). Two message routes of (2) and (5) differ from the ordinal one. In Fig. 4.1, MN has a connection to FA1 and moves into the cell of FA2. {1} The MN receives an agent advertisement from FA2 in the overlap region between the cells. However, the agent advertisement has no certification of FA2 and the MN can not

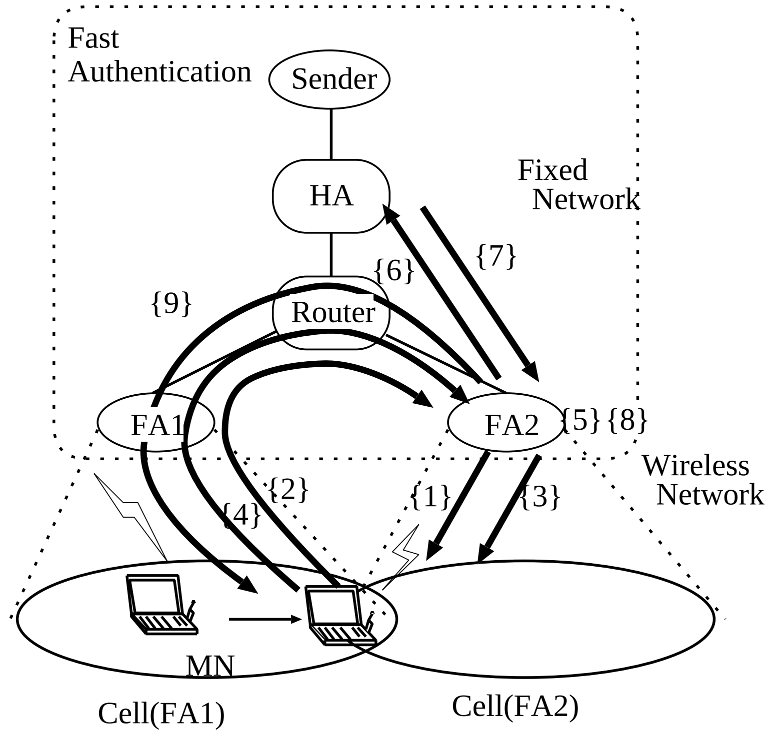

Figure 4.1: An example of key distribution and authentication by the fast authentication method.

authenticate FA2. If the MN already has the certification, go to {4}. Otherwise, {2} it sends a message to require a certification of FA2 via FA1, not directly to FA2. {3} When FA2 receives the request, it broadcasts an agent advertisement with its certification (Protocol 2(1)). Then, the MN authenticates FA2. {4} The MN sends a registration request to FA2 through FA1 (Protocol 2(2)). {5} When FA2 accepts the request, it forwards the registration request from the MN to the HA (Protocol 2(3)). Then, {6} the HA registers the care of address of FA2, and {7} sends a registration reply to FA2 (Protocol 2(4)). {8} FA2 receives the registration reply, and {9} forwards it to the MN through FA1 (Protocol 2(5)).

From these procedures, session keys are securely distributed for the registration updates and the authentications are also executed at the same time. In this method, the FA with stronger signal is called a primary FA and the other one is a secondary FA. However, the MN registers FA2 as its secondary FA at the first time, even if its signal strength is stronger than the one of FA1.

In the proposed method, the MN sends all registration requests and communication data to the primary FA rather than to the secondary one except for the broadcast packets. There are two reasons to adopt such a way. At first, the request may not reach to the latter FA by the weak signal strength. Second, the construction of the logical link control in the MN is simplefied. The information of the primary FA and secondary FA are also registered to the HA. When the HA interrupts data from the sender, the data are sent to the MN via the primary FA. Suppose the both FAs are swapped since their signal strength are changed. Then, the MN sends a registration update message to notify this fact. Although the HA does not recognize this modification until the message is received, it can receive the data from the HA via the secondary FA which is the previous primary FA.

The fast authentication method is expected to provide an effective authentication and registration mechanism to Mobile IP. However, if no overlap region exists, the MN can not send registration requests to FA1. In this case, it must send the request to the secondary FA as in Protocol 2.

4.4 Computer Simulation Experiments

To evaluate the proposed authentication method in the previous section, computer simulation experiments are executed with ns-2.1b7. In the fast authentication method, the MN can distribute its public keys and register for the new FA which may not become the primary FA of the MN. This overhead is also estimated in three mobility models of the MN.

4.4.1 Simulation Conditions

Two handoff methods are simulated. The first is the proposed method in the previous section. The second one is the fast handoff method in 3.2 with the authentication method of Protocol 1 for every registration update. Indeed, the second method with Protocol 2 produces much larger handoff latency than the one with Protocol 1. The average latency of the one with Protocol is observed 130ms under the same conditions in the followings. These simulation results are omitted here.

The network model for the simulation experiments consists of a fixed network and wireless network. The topology and each link speed for the experiment is illustrated in Fig. 4.2. The simulation conditions are in below.

[Topological conditions]

- The number of senders is one and the unicast is flowed via the HA to the MN (not using bicast from HA).
- The number of MNs in each model is one.
- Medium access control at the wireless network in Fig. 4.2 is IEEE 802.11 implemented in [6].
- For a wireless communication in each cell, the transfer error is ignored and propagation delay is (distance between the FA and MN) / (speed of the light).

[Traffic conditions]

- The sender sends packets to the MN with a same interval (5.33ms) over UDP and the IP packet size is 256 octets. Thus, the transfer rate is 384kbps.

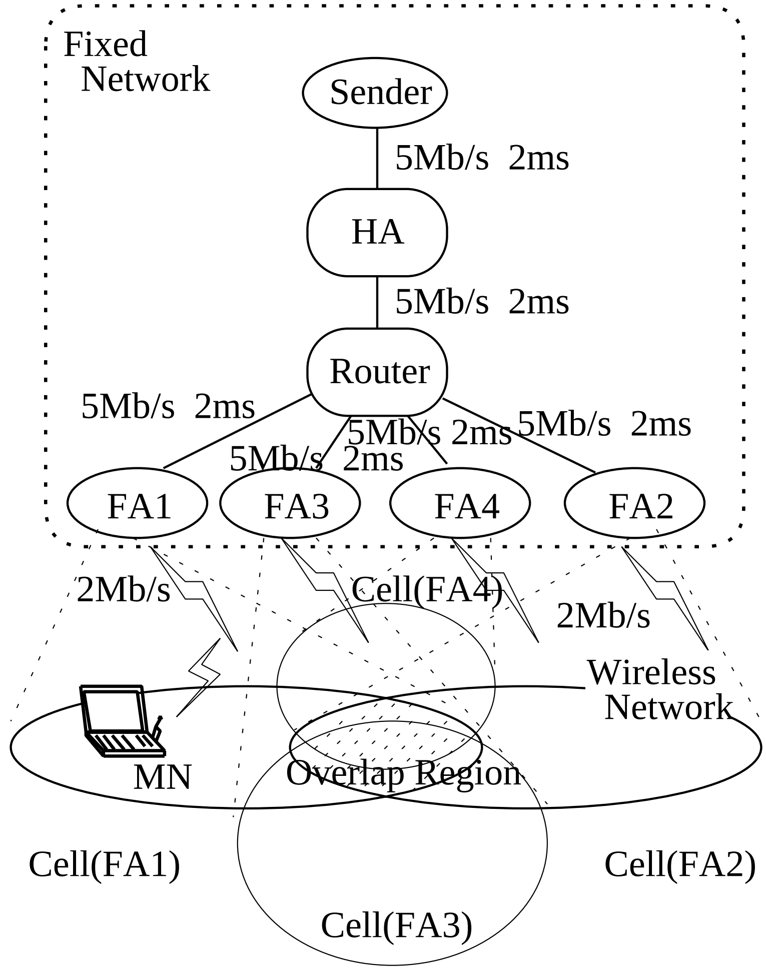

Figure 4.2: Simulation model 3.

- A packet size of an agent advertisement, registration request and registration reply are 128, 128 and 32 octets, respectively. These packet includes an authentication extension, such as the hashed values in Protocol 1.

[Security conditions]

- In the proposed method, each certification size is 512 octets. When it has parameters to share secret values, such as $(y_{\mathrm{A}\,K_{\mathrm{A}}^{-1}})$ and $(y_{\mathrm{A}},\ \alpha,\ N)_{K_{\mathrm{A}}^{-1}}$, the certification size is added to 256 and 512 octets for the extension, respectively.

- All registration requests are accepted by each entity in the authentication procedures.

- A delay of the registration procedure with the certificate-based authentication is 50ms at the MN and 25ms at the FA and the HA.

[Geographical conditions]

- There are one HA and four FAs. Each FA has a circular cell whose radius is 200m.

- The MN moves randomly within a square of 129600 ($= 360 \times 360$) m^2 at the multiple overlap regions in Fig. 4.2.

- At each midpoint of all sides of the square, an FA (FA1–FA4) is placed.

[Mobility conditions]

- The initial location of the MN is randomly chosen within the square.

- A movement of the MN is selected in combination of four speed models and three direction selection models. For the speed, a constant speed is selected as uniform random value among 0–1.0m/s, 0–5.0m/s, 0–10.0m/s or 0–20.0m/s. For the direction, the MN straightly moves within the square during 1s to a random direction whose angle is restricted within ± 180, ± 90 or ± 45 degrees from the current direction of the MN. When it stays less than 1m apart from any side of the square,the MN can select from all angles. The movements are repeated for 20000s.

[General conditions]

- An agent advertisement from each FA is periodically broadcasted every 0.2s within a whole cell area of the FA. The MN's registration life time[2] in the FAs and HA are 1.0s.

max. speed	1m/s	5m/s	10m/s	20m/s
proposed-all	5.7	5.5	5.9	5.7
proposed-half	5.5	5.4	5.5	5.4
proposed-quarter	5.4	5.4	5.4	5.5
fast-all	22.0	22.2	22.0	22.8
fast-half	21.5	22.1	22.0	21.7
fast-quarter	20.5	21.8	22.0	21.5

Table 4.1: Average handoff latency (ms).

- In the fast authentication method, the MN decides its primary and secondary FAs by the signal strength of the agent advertisement, and changes its primary FA only when it receives a stronger signal of an agent advertisement than the current one. However, 0.2s interval is required for changing the primary FA from the last change.

- The MN can broadcast an agent solicitation, if the MN detects that a packet loss occurs or a packet arrival interval becomes larger than 10ms.

- MAC medium of the MN does not deregister FA for 0.2s since it registers the FA.

- In the MN, the HA, the FAs and the router, each processing capacity for routing is 1000000 packets per second, and copying, encapsuling and decapsuling a packet take 10μs, respectively. A delay of the registration update by Protocol 1 at each HA and FA is 100μs.

- Every result satisfies 5% confidence interval with 90% confidence level.

4.4.2 Simulation Results

At first, handoff latencies are shown. In general, a handoff latency is used as two meanings. The first one is a latency occurred during registration procedures, i.e., the period is from the time the MN receives an agent advertisement of the new FA to the time the MN receives the registration reply from the new FA. The authentication delay at each node is included in the period. The second meaning is the packet arrival interval which no data packet arrives at the MN during switching the communication route. In the following, the second latency is just called handoff latency.

Table 4.1 shows the results of the handoff latencies, where the fast authentication method and fast handoff method are described as “proposed” and “fast,” respectively. “all,” “half” and “quarter” mean the available mobility model of the MN. In the table, the average handoff latencies of the two handoff methods are 5.4–5.9ms and 20.5–22.8ms, respectively. The fast authentication method results almost no handoff latency and no packet loss is observed in spite that it needs 149.0–153.8ms to complete the authentication procedures (the fast handoff method needs 16.1–16.7ms). If the MN stays for 150ms + 13ms + 200ms (handoff procedures, round trip time between the MN and the HA, and the interval of an agent advertisement) at an overlap region, no disrupted communications is served. The next evaluates the overhead of the fast authentication method to achieve these results.

Fig. 4.3 shows the average utilization rate of an authenticated secondary FA, i.e., a ratio that the secondary FA becomes as the primary within the life time. From this graph, “all,” “half,” and “quarter” are 0.055–0.196, 0.130–0.461 and 0.210–0.678, respectively. For all mobility models, the rate gradually increases for the higher speed of the MN and also decreases for the narrower direction range at every speed of the MN. Since higher speed or narrower direction range makes the MN completely move into a new cell with higher probability. From the figure, at least 5.5%–19.6% of the secondary FA are used as the primary FA from the results of all range. The rate may seem to be low even if the MN distributes the keys for only one FA. However, the prediction efficiency of the

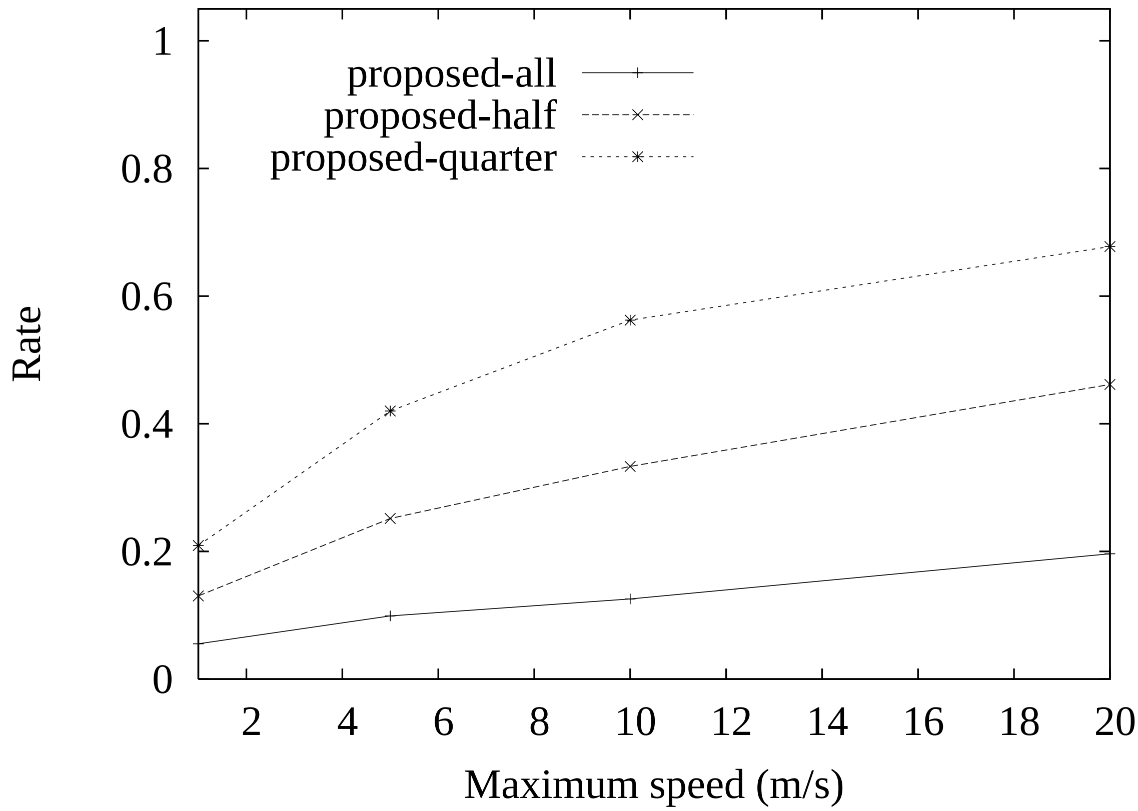

Figure 4.3: Average utilization rate of previous authenticated secondary FA.

fast authentication method should be rather higher than the multicast based handoff method such as [8] and [20], which are typical extension methods for Mobile IP, because the previous preparation is done for several FAs in these methods.

Fig. 4.4 shows the average period that the secondary FA becomes the primary one. In this graph, "all," "half" and "quarter" are 10.4–116.7s, 7.1–99.1s and 6.5–86.8s, respectively. In lower speed, the MN dose not need to connect to the secondary FA for longer period than that of the higher speed. In this case, many update messages for the secondary FA wastes resource of the networks. To improve the problem, as one of the solutions, the registration life time is set 1.0s for each FA in this simulation but the life time should be longer.

Table 4.2 shows the number of registration requests from the MN to the secondary

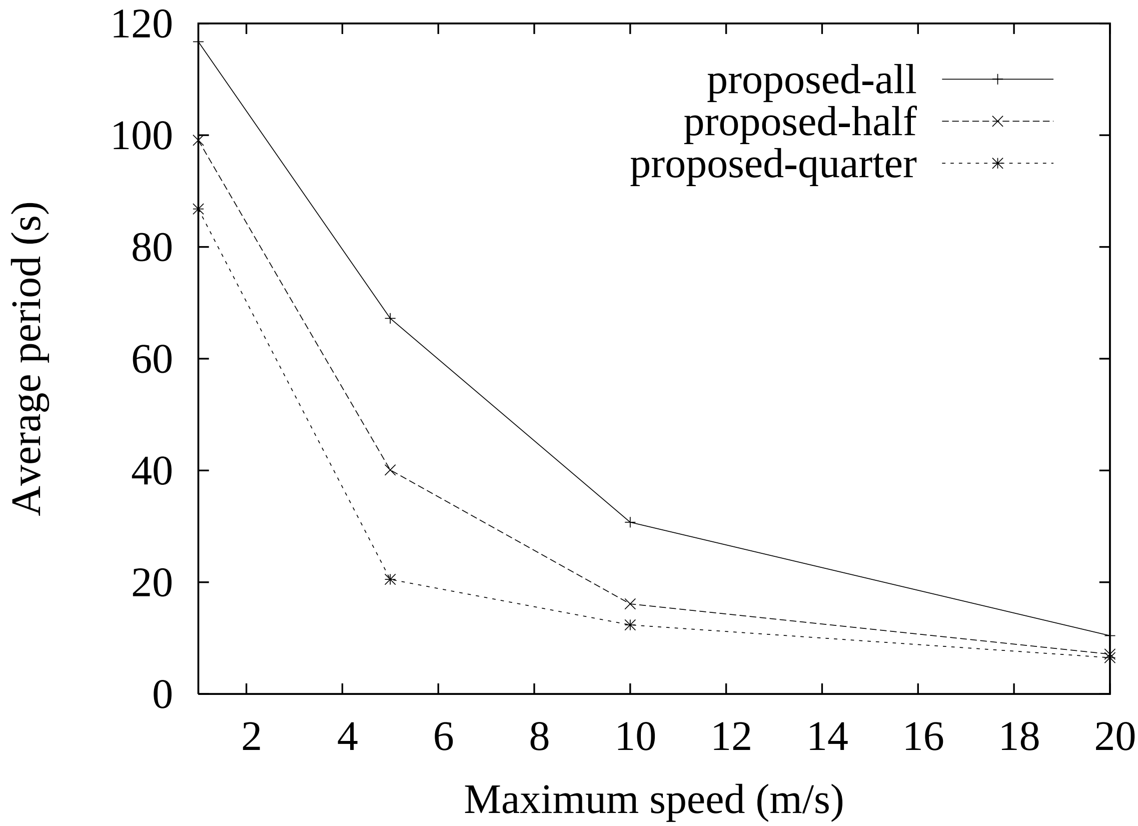

Figure 4.4: Average period for the secondary FAs to be the primary.

FA. The number in parentheses in the table show that the number of certificate-based authentications to the secondary FA. In the fast authentication method, the MN registers to two FAs in the overlap region (covering 79.4% in the square) and must update until their registrations are expired.

4.5 Conclusions for Chapter 4

In mobile networks, a handoff latency occurs when MN changes the connection point. The handoff latency is further long in case that the handoff procedures include authentications amang the communication entities. To reduce it, this chapter proposed the fast authentication method for cellular Internet environments in Mobile IP. The proposed

max. speed	1m/s	5m/s	10m/s	20m/s
proposed-all	134837.5	84417.0	84118.5	82864.0
	(129.0)	(378.9)	(576.6)	(1136.7)
proposed-half	92107.0	83310.5	85218.3	83588.5
	(113.0)	(310.5)	(573.4)	(1146.9)
proposed-quarter	94122.0	84417.0	81941.1	82617.2
	(187.7)	(312.5)	(572.4)	(1194.3)

Table 4.2: Average number of registration requests for the secondary FA.

protocol is realized with simple movement prediction control of the MN in the seamless handoff method. The simulation experiments show that the fast authentication method is effective for secure and real-time communications during handoff procedures. However, the simulation results depends on the simulation model, especially geometrical cells arrangement. The proposed method should also be evaluated with more complicated models with many FAs, Mobile IP with the route optimization and hierarchical BS models.

Chapter 5

Adaptive Resource Reservation and Admission Control for Cellular Internet Environments

This chapter proposes an adaptive resource reservation method and admission control which MN reserves the bandwidth for the secondary BS by using mobility prediction of the seamless handoff method. From the method, the MN can reserve the bandwidth from the overlap region between the cells. In order to implement the controls for Mobile IP, a Diffserv-based QoS mechanism is also proposed.

5.1 Introduction

In order to provide high quality communications for the multimedia traffic, the QoS negotiation is typically required every time when a client starts the communications. If enough resources for the new communication are not prepared for the client, the communication can not be established. This is called "new call blocking."

The negotiation is additionally required every time MN changes its connection point from the current BS to the other BS. In the following, the latter BS calls the new BS. However, from the characteristic of mobile networks, the new BS does not guarantee to provide the current quality of the communication to the MN. In the worst case, the communication would be disconnected at the new BS, which is called "handoff call blocking."

For high quality of the mobile communications, the cellular network is required for lower rates of the new call blocking and handoff call blocking. In general, the handoff call blocking rate prefer to be lower than the new call blocking rate, since the current communication has higher priority than the new communications when the network resources are shared.

To reduce the handoff call blocking rate, it is effective to reserve network resources for handoff calls which will be occurred in the near future. This concept is called "guard bandwidth[21]." However, when the guard bandwidth strategy is used, the network utilization tends to become low. Thus, in the guard bandwidth concept, it is very important how effectively to reserve the bandwidth of handoff calls for the adjacent cells from the current cell. Previously, many improved systems are proposed for the guard bandwidth strategy. Reference [22] considers the sufficient number of adjacent cells to periodically exchange their state information each other in a distributed control. Reference [23] proposes the admission control and resource reservation algorithm from the adjacent cells by using local and remote information for multimedia traffic. In reference [24], a shadow cluster concept is proposed to reserve the bandwidth to the cluster of the cells to which will be connected in the future. Mobility prediction methods are used for efficient resource reservation from the adjacent cells in references [25] and [26].

These methods improve the handoff blocking rate from the expense of the higher new call blocking rate. However, since these previously proposed concepts only subject how to reserve the resources in macroscopic aspect, and the improvement is limited.

The proposed method reserves the resource from one of the adjacent cells by using simple movement prediction control which is extending procedures of the seamless handoff method. In the cell overlap region, the MN can selects the mainly connecting BS in some suitable ones. The secondary BS is selected to reserved the bandwidth by the MN. An admission control is also proposed to manage the bandwidth by using the information.

5.2 Adaptive Resource Reservation and Admission Control

This section shows basic idea of an adaptive resource reservation method and an admission control.

5.2.1 Adaptive Resource Reservation

In cellular environments, overlap region is prepared for the MN to continuously communicate during moving between BSs (Fig. 5.1). In the overlap region, the MN can select its mainly connecting BS among them. This selection is typically based on the signal strength from the BS or utilization rate of the bandwidth in the BS.

In the proposed adaptive resource reservation method, the MN senses the signal strength to select the primary BS and secondary one. From this information, the MN can recognize the information of the secondary BS to reserve the bandwidth of the secondary BS from the overlap region via the primary BS. Moreover, the secondary BS can efficiently prepare the bandwidth for the MN. The proposed method reserves the bandwidth from

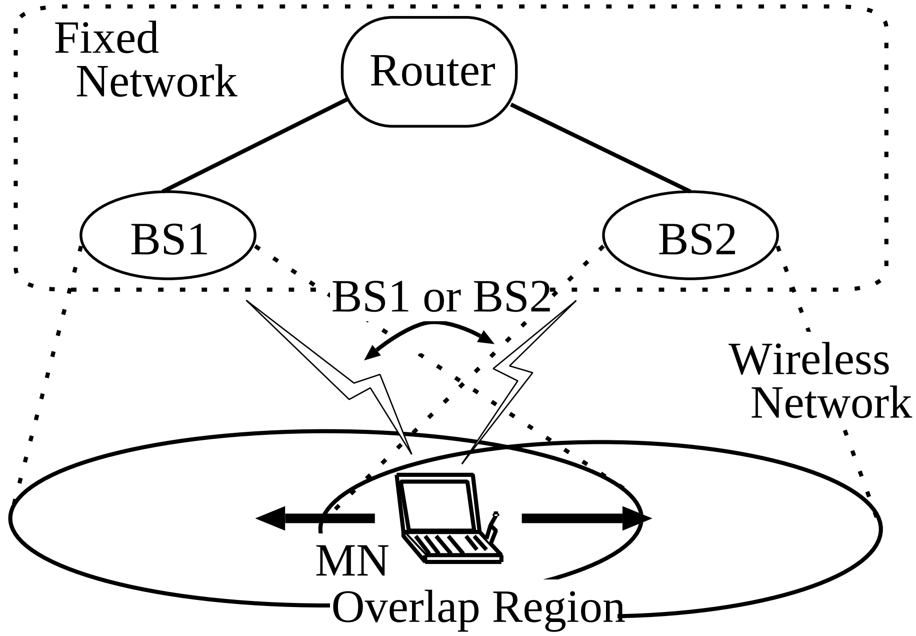

Figure 5.1: The MN acquisition of connectable BSs from these signal strength.

only one adjacent BS and the reserved bandwidth is used with high possibility since the MN can change the connection point to the secondary at any time. The next subsection shows an admission control with using above control.

5.2.2 Admission Control for Adaptive Resource Reservation

The adaptive admission control in the BS is introduced as follows. The control has two parts, i.e., the ones is for handoff calls and the other is for new calls.

[For handoff call]

if ($available\_bandwidth > required\_bandwidth$) {

Accept the handoff call.

} else {

Reject the handoff call.

}

[For new call]

if ($available\_bandwidth - (s\_guard\_bandwidth + \min(d\_reserved\_bandwidth, upper\_limit)$

$) > required\_bandwidth)$ {

Accept the new call.

} else {

Reject the new call.

}

available_bandwidth: is unused bandwidth in the BS.

s_guard_bandwidth: is statically prepared bandwidth of the handoff call.

required_bandwidth: is required bandwidth from the MN.

upper_limit: is the maximum reservable bandwidth for handoff call of the MN.

d_reserved_bandwidth: is dynamically reserved bandwidth by the MN for handoff calls.

$\square$

The admission control for the handoff call is same as the typical one since the handoff call has higher priority. The handoff call is accepted if the bandwidth is enough for the requested one. Note that *d_reserved_bandwidth* in *available_bandwidth* in the handoff call is not prepared for the every user who needs the bandwidth.

For new calls, the proposed admission control restricts the new call by the reserved bandwidth for the handoff call. The reserved bandwidth has two elements, the one is statically decided and the other is dynamically done.

s_guard_bandwidth is statically reserved bandwidth for the handoff call. If the value is set large, the new call blocking rate is large but the handoff blocking is low and network utilization is low. *d_reserved_bandwidth* is dynamically changed by the request of bandwidth from the MN in the overlap region and the maximum bandwidth is restricted by *upper_limit*. This dynamic value may result the new call blocking rate is too large or small, and the handoff blocking rate is too small or large. Since then, *upper_limit* is prepared not to reserve too much bandwidth for the handoff call and to prevent too large the new call blocking rate, and the reserved bandwidth is constructed from the two elements. Note that when *d_reserved_bandwidth* is 0, this algorithm works as "guard bandwidth" strategy[21].

5.3 System Architecture to Implement Adaptive Admission control in Mobile IP

This section proposes a system architecture to implement the adaptive admission control in the previous section to Mobile IP-based network.

5.3.1 QoS Signaling Protocol

Since the original Mobile IP has no QoS mechanisms by itself, some QoS mechanisms are proposed for mobile Internet environments. These mechanisms are categorized by the ones based on RSVP[28],[29] or Differentiated Services (Diffserv)[3],[30].

In the RSVP mechanisms, the end-to-end signaling negotiates the communication quality to every router via the communication route. Then, the communication quality of each flow is strictly guaranteed. In mobile environments, the communication route must be modified when a handoff procedure is occurred. Since the negotiation for the communication quality also has to be executed again for new route, the cost from the procedure is quite large.

The Diffserv mechanisms provide the communication quality more loosely. The DS field[3] in an IP header is used to classify the service at each router and the communication quality is guaranteed by each service class. In Diffserv, an Service Level Agreement (SLA) is contracted to define the communication quality between a user and Diffserv domain. For wireless networks, reference [30] proposes a traffic profile of each user. From registering the traffic profiles to Diffserv domains, the communication quality is effectively improved.

In the proposed system, each network manages the resources with the traffic profile based on [30]. Fig. 5.2 shows the proposed QoS architecture. The QoS agent is prepared in MN and FA to manage traffic quality and network resources. When the communication quality is negotiated between the MN and the FA, the QoS agent in the MN submits an Internet Control Message Protocol (ICMP) message with a traffic profile to that of the FA as in [30].

The negotiation procedures between the network and the MN are described in below. Suppose that the MN can recognize whether or not the network supports Diffserv with traffic profiles by periodical agent advertisements from FAs.

(1) The application in the MN requires resource reservation to the QoS agent before it starts the communication.

(2) The QoS agent in the MN submits the traffic profile to that of the FA. Then, the latter QoS agent decides whether the request is accepted or not by the admission control in Section 5.2.2. The traffic profile has the care of addresses of the MN. In the proposed system, each MN can have one or two care of addresses, where the one of them is for the FA. When the request is accepted and the profile has two care of addresses, the QoS agent in the FA forwards the profile to the other FA.

(3) The QoS agent in the FA replies the result of the request to the QoS agent in the MN.

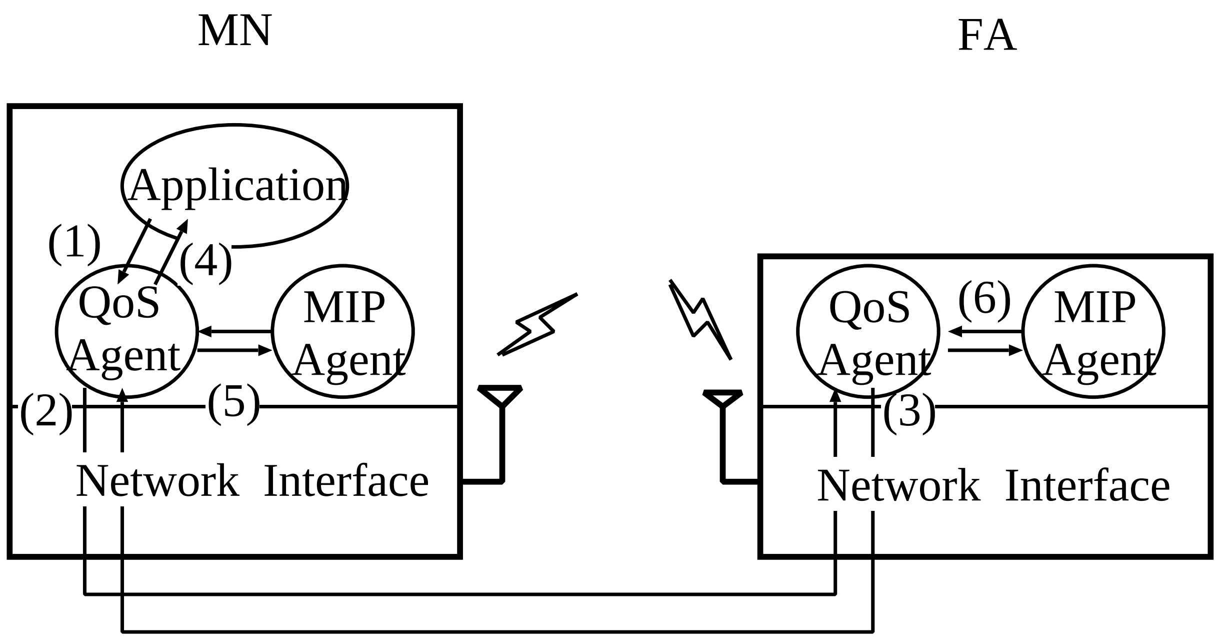

Figure 5.2: QoS architecture.

(4) The QoS agent in the MN passes the result to the application. If the request was accepted, the application starts the communication.

(5) and (6) are for updating the traffic profile. When each Mobile IP (MIP) agent in the MN and FA periodically sends and receives a registration update[2], respectively.

Resource releasing procedures are also executed by the same way. Through these procedures, the QoS negotiation is completed.

5.3.2 Registration Procedures for Adaptive Admission Control

This subsection shows modified registration procedures for Mobile IP. The procedures modified from the original registration procedures in Fig. 5.3 are explained as follows.

(1) The FA broadcasts the agent advertisement with the information described in Section 5.3.1.

(2) The MN sends a registration request to the FA. When the MN is currently commu-

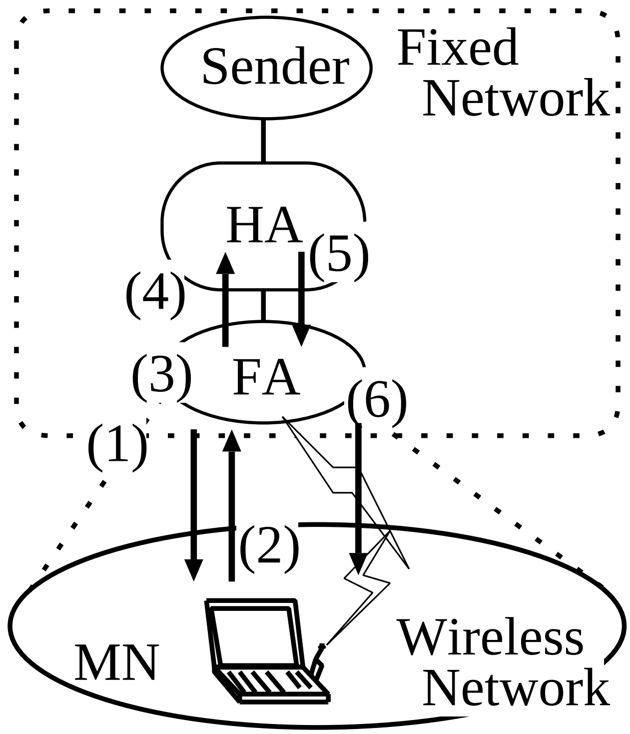

Figure 5.3: Registration procedures.

nicating with the FA, it also transmits the traffic profile update message with the request.

(3) The FA registers the MN and updates the traffic profile.

(4) The FA forwards the registration request to the HA of the MN.

(5) The HA registers the MN, and then sends a registration reply to the FA.

(6) The FA forwards the registration reply to the MN.

From (1) to (6), the original registration and QoS information update are completed.

For the proposed admission control, above method is extended that the MN can recognize the information of the secondary FA and registers to it. This control is realized by extending the seamless handoff method. One of the care of address is assigned to the current FA. The other one is for the new FA which is predicted that the MN will connect in the next. If the prediction is much proper, the care of address of the new FA is very useful for the previous reservation of network resources and adaptive admission control.

The modified registration procedures to the new FA is depicted in Fig. 5.4. In the figure, it is assumed that the MN is changing its connection point from the current FA (FA1) to the new FA (FA2). In the overlap region, the MN can receive agent advertisement from the both FAs. The communication to the current FA is duplex, however, the one to the secondary may simplex from the FA to the MN.

(1) FA2 broadcasts the agent advertisement.

(2) The MN sends a registration request to FA2 via FA1, not directly to FA2. This is because that the request may not reach to FA2, and the MN can send a registration request to the new FA only if the MN can receive the agent advertisement.

(3) FA1 updates the registration request and traffic profile of the MN.

(4) FA1 forwards them to FA2.

(5) FA2 registers and updates the registration request and traffic profile.

(6) FA2 forwards the registration request to the HA.

(7) The HA registers the MN and sends a registration reply to FA2.

(8) FA2 forwards the registration reply via FA1 to the MN.

Through above procedures, the MN can register itself and update its QoS profile to the new FA. Then, the seamless communication is also provided to the handoff procedures. In the new FA, the information is used for the admission control in Section 5.2.2.

In the next section, the computer simulation experiments are shown to evaluate the proposed method.

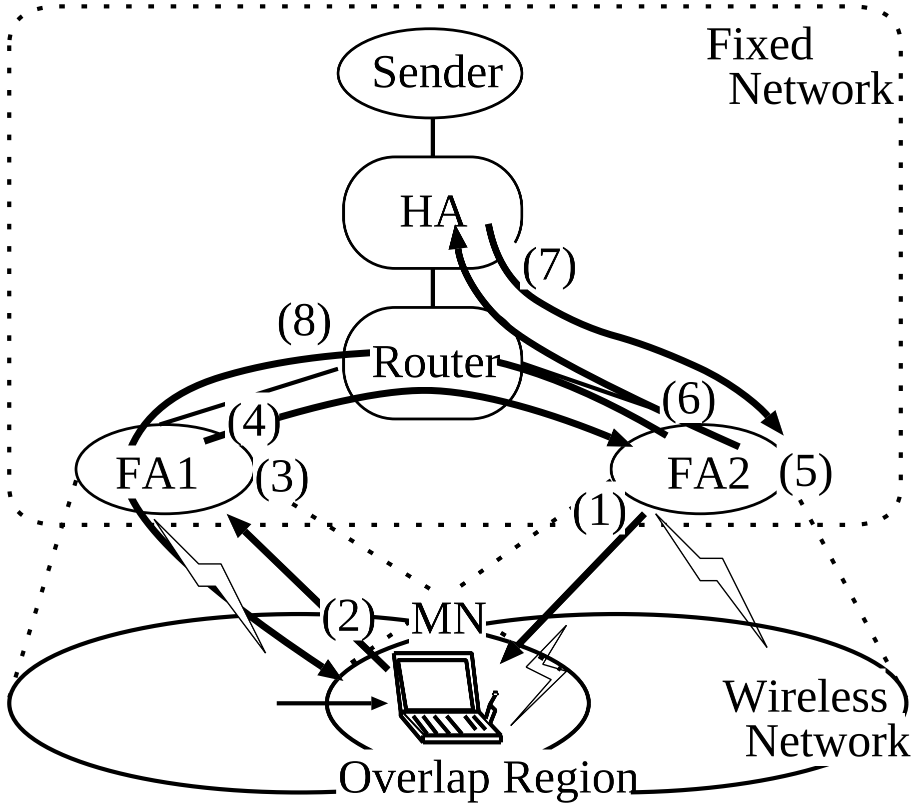

Figure 5.4: Registration procedures for adaptive admission control.

5.4 Computer Simulation Experiments

To evaluate the proposed resource reservation method and admission control, the computer simulations experiments are executed by network simulator ns-2.1b7.

The simulations are executed by changing the parameters in Section 5.2.2 and the dynamic part for the reserved bandwidth is newly added to the proposed method. At first, the simulation conditions are described and then, the simulation results are discussed.

5.4.1 Simulation Conditions

The network model for the simulation experiments consists of a fixed network and wireless network as shown in Fig. 5.5.

[Admission control conditions]

The parameters defined in Section 5.2.2 are in below.

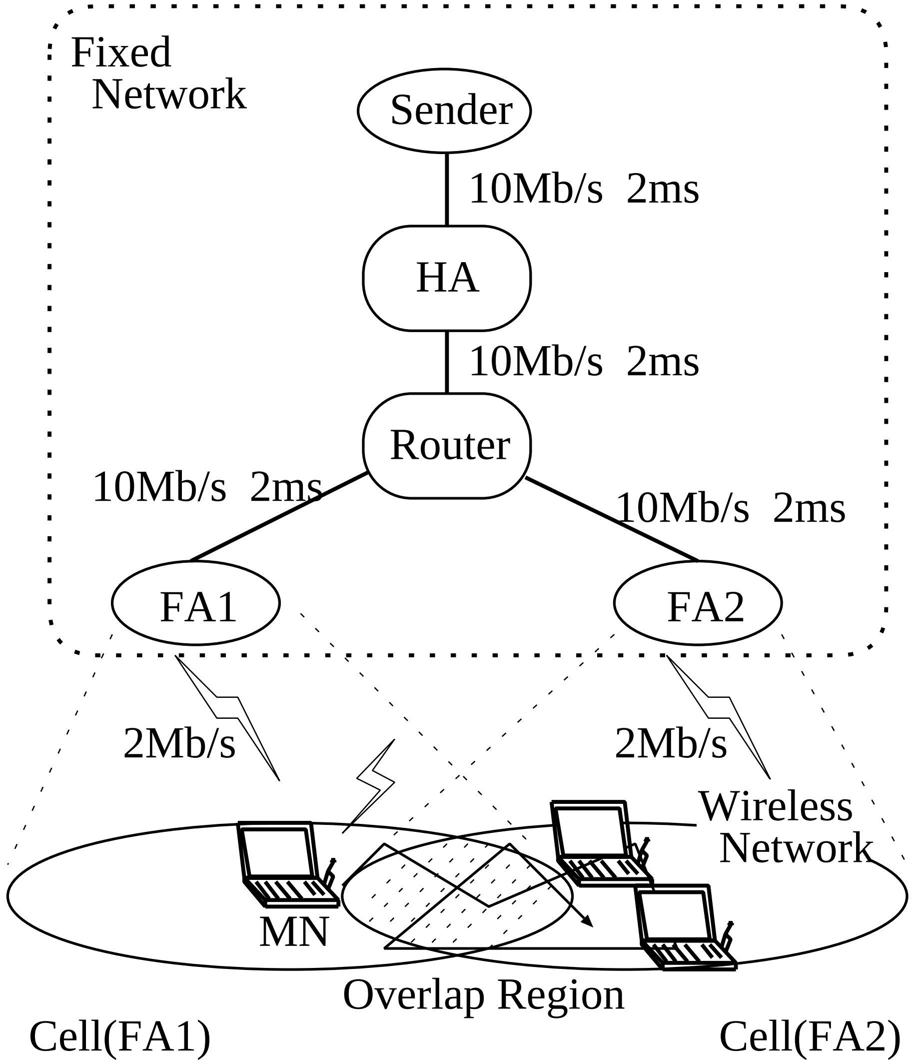

Figure 5.5: Simulation model 4.

- Total available bandwidth in the FA is 64Kbps $\times$ 6. In addition, $n \times$ 64Kbps is prepared for handoff calls, where $n = 0$, 1 or 2.

- *s_guard_bandwidth* and *d_reserved_bandwidth* are changed to 0, 1, 2, 3 and 0, 1, 2, 3, respectively. *available_bandwidth* is depended on the number of communicating users.

[Geographical conditions]

- FA1 and FA2 are placed at center of opposite side of the square (280m $\times$ 280m), respectively.

- The overlap region covers 44% of the square size.

[Topological conditions]

- The number of MNs is 10. There are one HA and two FAs. Each FA has a circular cell whose radius is fixed to 200m.
- The link delay and bandwidth are depicted in Fig. 5.5 and medium access control at the wireless network is IEEE 802.11.
- For wireless communications, the transfer error is ignored and propagation delay is (distance between the FA and MN) / (speed of the light).

[Traffic conditions]

- Each MN has a unicast connection from the sender through the HA.
- The sender sends packets to the MN with a same interval over UDP. The IP packet size is 256 octets and the transfer rate is 64Kbps.
- A new call occurs from each MN at the exponential distribution with the average time of 5, 10, 20, 30, 60, 120 or 180 seconds.
- A new call hold time also obeys the exponential distribution with the average time of 60 seconds.

[General conditions]

- An agent advertisement from each FA is periodically broadcasted every 0.2s within a whole cell area of the FA. However, the MN responds to the message each 1.0s if it is communicating, or 2.0s otherwise. The MN's registration life time[2] in the FAs and HA is 3.0s.

- Packet sizes of an agent advertisement, registration request and registration reply are 128, 128 and 32 octets, respectively.

- The MN can broadcast an agent solicitation when the MIP agent detects that the MAC layer changes the current BS (for fast handoff).

- Every result satisfies 10% confidence interval with 90% confidence level.

[Mobility conditions]

The mobility is considered for the simple vehicular mobility in an urban area.

- The MN moves within a square of 280×280 m^2 including the overlap regions in Fig. 5.5.

- The MN straightly moves within the square. The direction is randomly selected, and the angle is restricted within ± 90 degrees from the current direction of the MN. However, the MN can select from all angles, when it stays less than 1m apart from any side of the square.

- The speed follows the exponential distribution with the average speed of 10m/s but is restricted within 40m/s.

- The speed and direction are updated with a period in the exponential distributed random value whose average is 10s.

[Handoff algorithm conditions]

A handoff is executed by the condition of the signal strength from the BSs and the handoff algorithm in the MN is in below.

if ($current\_BS == BS_1$)

  if ($10 \log(S_2/S_1)^2 > \Delta$) $current\_BS = BS_2$

else $current\_BS = BS_1$

else if ($current\_BS == BS_2$)

if ($10 \log(S_1/S_2)^2 > \Delta$) $current\_BS = BS_1$

else $current\_BS = BS_2$

where $current\_BS$ is the MN's current BS. S_1 and S_2 are the signal strength from BS_1 and BS_2, respectively. Δ is the hysteresis parameter = 6dB. □

5.4.2 Simulation Results

For each admission control, *s_guard_bandwidth* are *d_reserved_bandwidth* are changed 64Kbps × 0, 1, 2, expressed by "0," "1" and "2," respectively. In addition, each method is expressed by combination of the static part (*s_guard_bandwidth*) of "s" or dynamic part (*d_reserved_bandwidth*) of "d" in the following results. For example, "s1", "d1" and "s1-d1" mean the cases that static part is 64Kbps × 1, dynamic part is 64Kbps × 1, and static part is 64Kbps × 1 and dynamic part is 64Kbps × 1, respectively. In Figs. 5.6 to 5.9, the x-axis means the new call rate for one cell, whose value is calculated from the simulation results. For simply showing the simulation results, only "s1", "s2", "s1-d1", "s1-d2" and "s1-d3" are depicted in the results. The results in these figures are obtained for 100000s of the simulation time.

Fig. 5.6 shows the average new call blocking rate. In all methods, the new call blocking rate is increased for higher new call rate and preparing large bandwidth of handoff call. In "s2", "s1-d2" and "s1-d3," almost same results are obtained since almost same bandwidth is prepared for the handoff call in these methods. If the handoff call blocking rate of "s1-d2" and "s1-d3" are improved from the one of "s2", the proposed method is effective. This is because that the network resource is largely shared for the handoff call than for

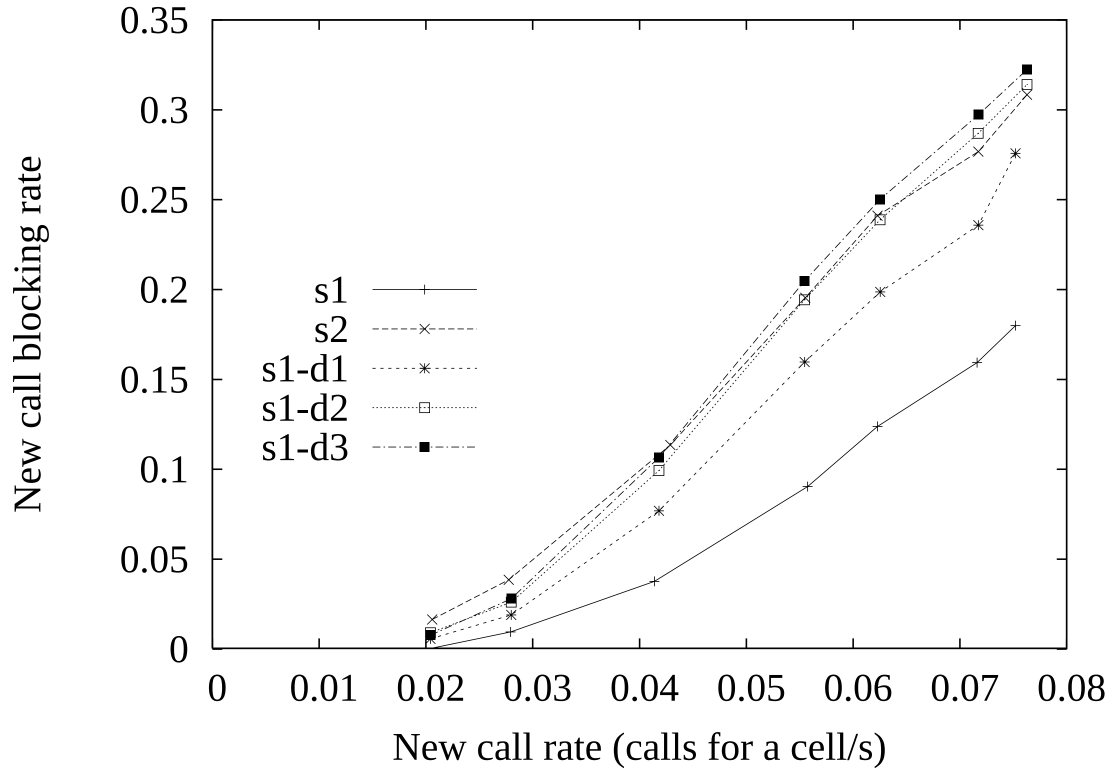

Figure 5.6: Average new call blocking rate.

the new call.

Fig. 5.7 depicts the handoff call blocking rate. In all methods, the handoff call blocking rate is increased for higher new call rate. In the case that larger bandwidth is prepared, the smaller handoff call blocking rate is served. For "s1-d2" and "s1-d3", the handoff call blocking rate is smaller than that of "s2" and improved 40–50 % at the maximum in spite that these are same new call blocking rate. This fact shows that the proposed method works well since the propose of the admission control is lower handoff call blocking rate with efficient network utilization. Especially, when the new call rate is higher, i.e., the cells are more crowed, the proposed method is much effective.

Fig. 5.8 shows the throughput of the entire network. From the simulation conditions, the maximum bandwidth of the total traffic for each cell is restricted to 64Kbps $\times$ 6, and

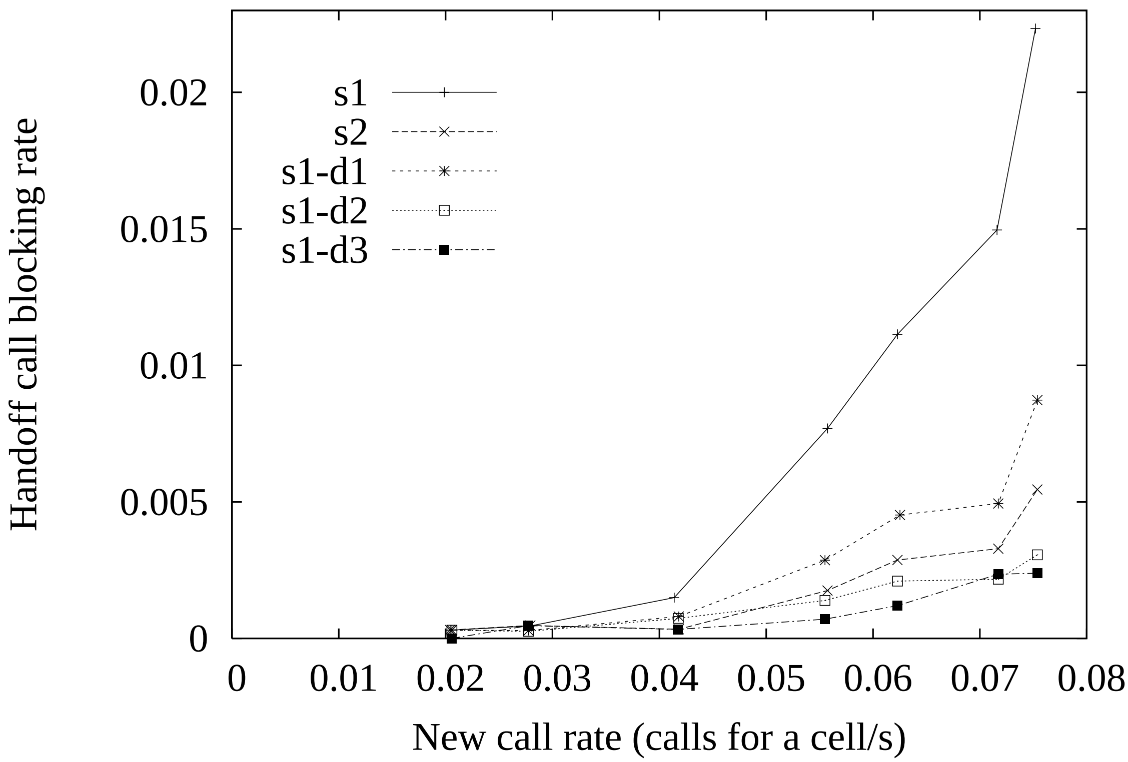

Figure 5.7: Average handoff call blocking rate.

then the one for the entire network is 64Kbps $\times$ 12. For smaller new call rate, all the results in the figure are almost same but the reserved bandwidth decreases the throughput for larger new call rate. For "s2', '"s1-d2" and "s1-d3", the throughput is almost same, however, in the proposed method, the bandwidth of the entire network is much shared for handoff calls not for new calls. This result also supports the purpose of the admission control, and thus the bandwidth is effectively used by the proposed admission control.

Finally, Fig. 5.9 depicts the overhead for realizing the proposed method. The overhead in this section is defined by the average number of registration requests and updates for new FA from one MN during the simulation time. In the figure, the results have almost same feature with that of throughput. When the reserved bandwidth is wide, the throughput decreases from Fig. 5.8. In this case or when the new call rate is low, the

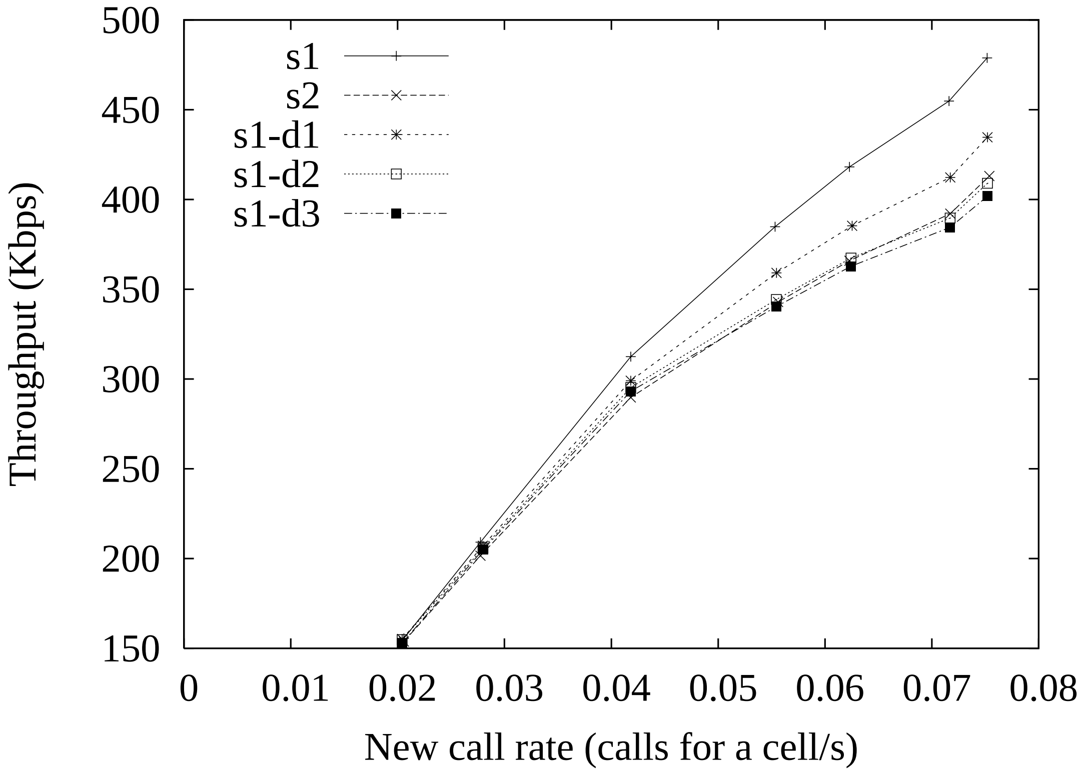

Figure 5.8: Average throughput.

number of communicating MNs are fewer, and then the number of registration updates for the current and new FAs are smaller. Because the overlap region is 44% of the square, the overhead of the number of requests is restricted to less than one time per a second for the simulation conditions even if the new call rate is higher.

5.5 Conclusions for Chapter 5

The adaptive resource reservation and admission control method based on the information from micro mobility is proposed and evaluated in this chapter. In order to implement the proposed method for Mobile IP, the system architecture including the QoS signaling protocol and registration procedures are also proposed. The simulation experiments show that the proposed algorithm provides the lower average handoff call blocking rate than

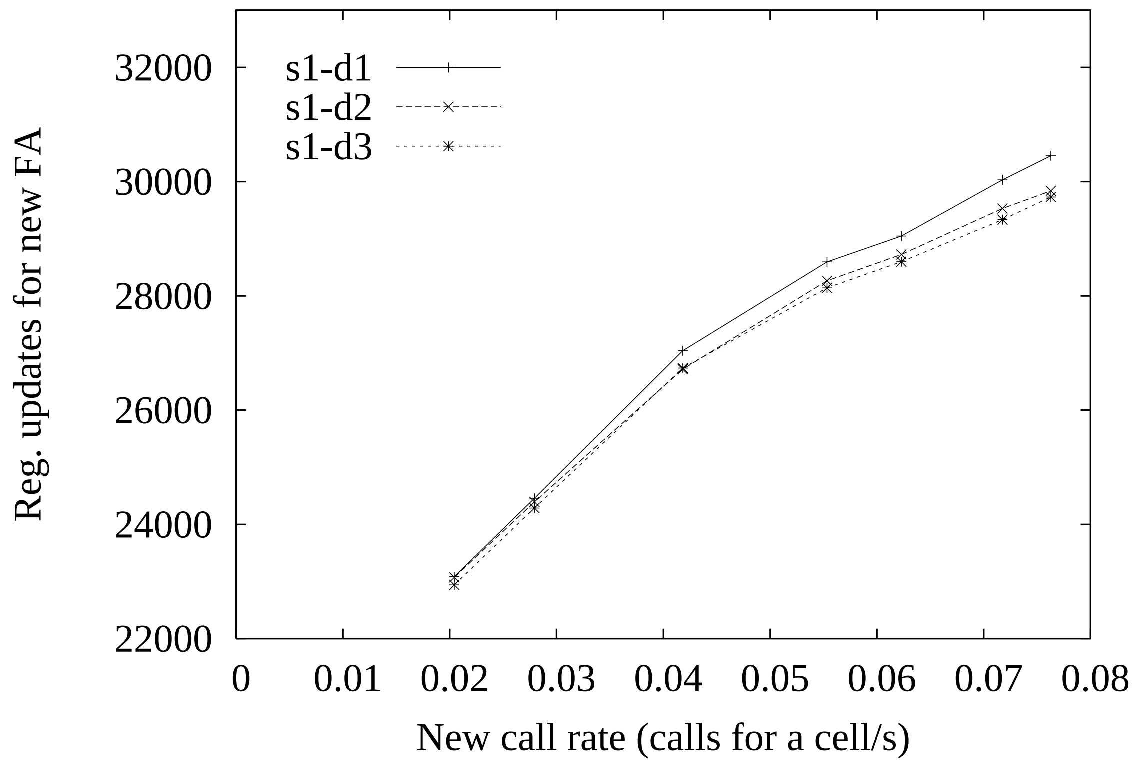

Figure 5.9: Average overhead.

the previous method in spite of low overhead for the expansion. In the simulation model, the traffic from the sender to the MNs is uniformed. The network topology is also simple. More realistic experiment should be executed for the traffic with multiple QoS classes on a network with many FAs. Although the proposed method is only for using information of micro mobility, a combination model of micro and macro mobility is more effective. In fact, in the case that the MN changes the BS very often, the bandwidth for the handoff call should be prepared in advance, not just before, to improve the handoff call blocking rate. To improve this "static" part, the previous reservation part is also dynamically changed by the information of the macro mobility, i.e., the number of the MNs dwelling in the adjacent cells.

Chapter 6

Adaptive Handoff Method for Efficient Resource Utilization

As mentioned in the previous chapter, the guard bandwidth strategy is effective for improving handoff call blocking rate. This chapter proposes an another approach which the timing of handoff is dynamically changed with the numbers of registered MNs for each FA. In the method, when a cell includes quite many MNs, each MN tends to easily go out from the cell. On the other hand, in the case of a less crowded cell, every MN tries to keep the current connection as long as possible. From above control, the network resource is efficiently used and the communication quality is also improved because the proposed method works as if the network resources are traded among the adjacent FAs.

6.1 Introduction

For high quality mobile communications, the cellular network is required for lower rates of new call blocking and handoff call blocking. In general, the handoff call blocking rate is preferred to be lower than the new call blocking rate, since the current communication has higher priority than the new communications when the network resources are shared. Realizing above control, many concepts are proposed. In [22]–[26], the handoff call blocking rate is improved by the expense of the higher new call blocking rate. A queuing handoff strategy is also introduced in [21] and [31]. These methods reduce the handoff call blocking rate by waiting for the handoff until the bandwidth for the handoff call is prepared. Their resource management and handoff algorithms are separately considered, although the relations of them are quite close. Therefore, the combination system of the both methods should provide more effective results.

This chapter proposes the other approach, where the timing of the handoff is dynamically changed by the utilization rate of BSs. In a cell including quite many MNs, each MN tends to easily go out from the cell. On the other hand, in a less crowded cell, every MN tries to keep the current connection as long as possible. The control is realized by adaptively changing the hysteresis value of the "threshold-with-hysteresis" handoff algorithm[32] by comparing each utilization rate in the current BS and new BS. Although the algorithm is originally proposed to reduce unnecessary handoff when MN wanders between two BSs, the algorithm is applied to solve the handoff call blocking problem by adaptively changing the hysteresis value. From above control, the network resource is efficiently used and the communication quality is also improved because the proposed method works as if the network resources are traded among the adjacent BSs. The proposed control is also evaluated on the QoS architecture in the previous chapter.

6.2 Handoff Algorithm

When MN moves from the current cell to an another cell, the MN must change its link-level connection point to the BS in the another cell. One of the decision parameters to switch the connection point is the radio signal strength from BSs. In general, the MN connects to the BS with the strongest signal strength among the connectable BSs. When the signal strength from the current BS is weaker than the one of the other BSs, the MN switches its current BS. However, if the MN wanders between two cells, unnecessary handoff procedures are frequently executed. To reduce the handoff-oscillations, a simple threshold-with-hysteresis algorithm such as Algorithm 1[32] in below is typically used. Note that Algorithm 1 describes for the situation when MN moves between BS_1 and BS_2.

[**Algorithm 1**]

if ($current\_BS == BS_1$)

  if ($10 \log(S_2/S_1)^2 > \Delta$) $current\_BS = BS_2$

  else $current\_BS = BS_1$

else if ($current\_BS == BS_2$)

  if ($10 \log(S_1/S_2)^2 > \Delta$) $current\_BS = BS_1$

  else $current\_BS = BS_2$

where $current\_BS$ is the MN's current BS. S_1 and S_2 are the signal strength from BS_1 and BS_2, respectively. Δ is the hysteresis parameter > 0. □

The hysteresis parameter Δ is important factor in the algorithm. Too large Δ produces large decision delay of the handoff. Too small Δ causes the handoff oscillation. Reference [32] discusses the suitable value of Δ. Reference [19] further improves Algorithm 1 to reduce exceeded handoffs for complex movement of MNs by adding the handoff decision thresholds. In such improvements, however, the tradeoff between the number of handoffs and handoff decision delay is mainly discussed, and the communication quality is considered for only one user.

6.3 Adaptive Handoff Algorithm

This section introduces a new parameter to Algorithm 1 in Section 6.2 to efficiently share network resources among multiple users for realizing high utilization of the entire network. To understand the proposed control, considering the following situation, i.e., MNs stay in the cells of BS_1 and BS_2, where the cell of BS_1 is crowded but the other of BS_2 is not. Suppose that MN is currently connected to BS_1 and moves to BS_2. Since BS_1 is connected from many MNs, the new call blocking rate and handoff call blocking rate become much higher than the ones of BS_2. Then, the lower handoff call blocking rate should be served from the cell of BS_2. Therefore, from the view point of the network utilization, it is preferable that the MN should change its connection point to BS_2 as fast as possible. On the other hand, when the MN is currently connected to BS_2 and moves to BS_1, the entirely opposite control is required from the previous assumption, i.e., the MN continues to keep connecting to BS_2 as long as possible.

To realize above control, this section proposes Algorithm 2 to adaptively change the hysteresis parameter Δ in Algorithm 1. In the algorithm, if the cell is much crowded then Δ becomes smaller to let MNs easily release the connection to the current BS.

[**Algorithm 2**]

if ($current\_BS == BS_1$)

  case: BS_2 is crowded

    if ($10 \log(S_2/S_1)^2 > \Delta_L$) $current\_BS = BS_2$

    else $current\_BS = BS_1$

  case: BS_2 is not crowded

    if ($10 \log(S_2/S_1)^2 > \Delta_S$) $current\_BS = BS_2$

    else $current\_BS = BS_1$

case: BS_1 and BS_2 have the same conditions

if ($10 \log(S_2/S_1)^2 > \Delta$) $\quad current\_BS = BS_2$

else $\quad current\_BS = BS_1$

else if ($current\_BS == BS_2$)

case: BS_1 is crowded

if ($10 \log(S_1/S_2)^2 > \Delta_L$) $\quad current\_BS = BS_1$

else $\quad current\_BS = BS_2$

case: BS_1 is not crowded

if ($10 \log(S_1/S_2)^2 > \Delta_S$) $\quad current\_BS = BS_1$

else $\quad current\_BS = BS_2$

case: BS_1 and BS_2 have the same conditions

if ($10 \log(S_1/S_2)^2 > \Delta$) $\quad current\_BS = BS_1$

else $current\_BS = BS_2$

where $current\_BS$ is the current BS of the MN. S_1 is the radio signal strength from BS_1. S_2 is the radio signal strength from BS_2. Δ, Δ_L and Δ_S are the hysteresis parameters where $\Delta_L > \Delta > \Delta_S > 0$. □

It should be noted that Algorithm 2 changes the timing of the handoff to be not only delayed but also accelerated from the network conditions. Under the controls, the network resources are effectively used by MNs.

6.4 System Architecture to Implement Adaptive Handoff Algorithm

This section proposes a system architecture to implement the adaptive handoff algorithm (Algorithm 2) in the previous section to Mobile IP-based networks. The adaptive handoff algorithm is implemented on QoS architecture in Section 5.3.1. The required information for the algorithm is collected by the architecture.

In Algorithm 2, the MAC layer in MN needs to recognize whether or not the current FA and new FA can provide enough network resources to the MN. However, the network layer level QoS mechanism such as Diffserv is just treated within the IP layer. Some kinds of communications are required between the IP layer and the MAC layer to realize that the MAC layer obtains the QoS information. Therefore, the MN should have two network interfaces and deal with the network conditions so as to receive the IP level information from the current FA and new FA[19].

By using Figure 6.1, above information exchange procedures are explained as follows. In the figure, let MN stay in the overlap region between the cells of the current FA and the new FA.

(1) In each FA, the MIP (Mobile IP) agent obtains the information of the network resources from the QoS agent, and decides whether the cell of the FA is crowded or not.

(2) Each FA periodically broadcasts an agent advertisement[2] with "crowded bit."

(3) The MIP agent in the MN receives crowded bits from the FAs, and then informs them to the network interfaces.

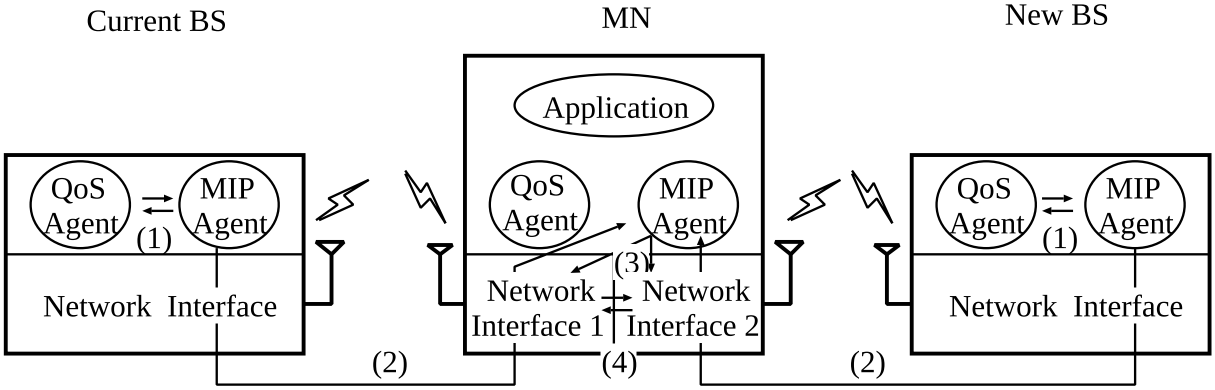

Figure 6.1: Information exchange to decide the crowded FA.

(4) Each network interface communicates each other to exchange information for Algorithm 2.

Then, the decision algorithm of "crowded," "not crowded" or "same condition" in Algorithm 2 is realized as in below.

[Crowded FA Decision Algorithm]

if ($crowded\_bit_{cur}$ == 1 && $crowded\_bit_{new}$ == 0) {

The new FA is not crowded.

} else if ($crowded\_bit_{cur}$ == 0 && $crowded\_bit_{new}$ == 1){

The new FA is crowded.

} else {

The current and new FAs are in the same condition.

}

where $crowded\_bit_{cur}$ is a "crowed bit" from the current FA. $crowded\_bit_{new}$ is a "crowed bit" from the new FA. □

In the next section, the computer simulation experiments evaluate the proposed method.

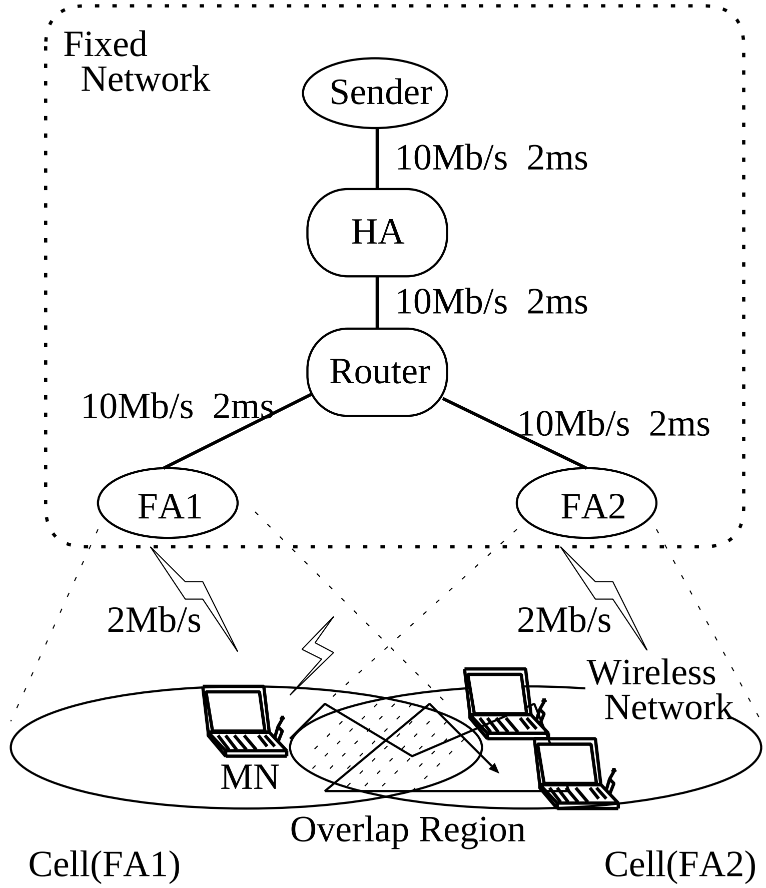

Figure 6.2: Simulation model 5.

6.5 Computer Simulation Experiments

To evaluate the proposed handoff algorithm, the computer simulation experiments are executed with network simulator ns-2.1b7. In the simulations, Algorithms 1 and 2 with the admission control of the guard bandwidth strategy[21] are compared.

6.5.1 Simulation Conditions

The network model for the simulation experiments consists of a fixed network and wireless network as illustrated in Figure 6.2.

[Handoff algorithm conditions]

- In Algorithms 1 and 2, the hysteresis values of Δ_L, Δ, Δ_S are 10dB, 6dB and 2dB, respectively.

[Admission control conditions]

- Available bandwidth in the BS is 64Kbps $\times$ 6. In addition, $n \times$ 64Kbps is statically prepared for handoff calls, where $n = 0$, 1 or 2.

- When each FA has six connections, it sets "crowded bit" in a periodical agent advertisement in Algorithm 2.

[Topological conditions]

- The number of MNs is 10. There are one HA and two FAs. Each FA has a circular cell whose radius is fixed to 200m. FA1 and FA2 are placed at the center of the opposite side of the square, respectively.

- The link delay and bandwidth are depicted in Figure 6.2 and medium access control at the wireless network is IEEE 802.11.

- For wireless communications, the transfer error is ignored and propagation delay is (distance between the FA and MN) / (speed of the light).

[Traffic conditions]

- Each MN has a unicast connection from the sender through the HA.

- The sender sends packets to the MN with a same interval over UDP. The IP packet size is 256 octets and the transfer rate is 64kbps.

- A new call is occurred from each MN at the exponential distribution with the average time of 5, 10, 20, 30, 60, 120 or 180 seconds.

- A new call hold time also obeys the exponential distribution with the average time of 60 seconds.

[General conditions]

- The MN decides the current FA and new FA by Algorithm 1 or 2.
- The MN can broadcast an agent solicitation when the MIP agent detects the MAC layer changes the current BS (for fast handoff).
- Every result satisfies 10% confidence interval with 90% confidence level.

[Mobility conditions]

The mobility is considered for the simple vehicular mobility in an urban area.

- The MN moves within a square of 280×280 m^2 including the overlap regions in Figure 6.2.
- The MN straightly moves within the square. The direction is randomly selected, and the angle is restricted within ± 90 degrees from the current direction of the MN. However, the MN can select from all angles, when it stays less than 1m apart from any side of the square.
- The speed follows the exponential distribution with the average speed of 10m/s but is restricted within 40m/s.
- The speed and direction are updated with a period in exponential distributed random value whose average is 10s.

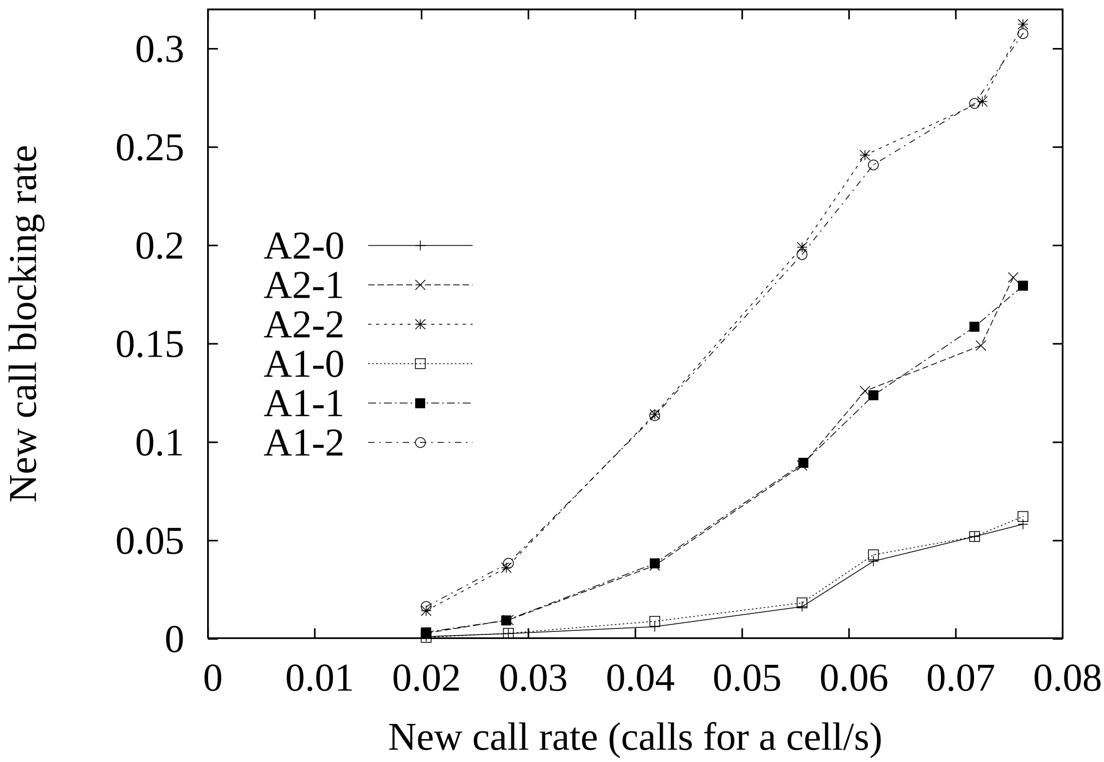

Figure 6.3: Average new call blocking rate.

6.5.2 Simulation Results

The simulations are experimented for two algorithms, i.e., Algorithms 1 and 2, which are expressed by "A1" and "A2" in the following results, respectively. For each admission control, the reserved bandwidth for handoff call is selected among 64Kbps $\times$ 0, 1, 2, expressed by "0," "1" and "2," respectively. In Figures 6.4 to 6.5, the x-axis means the new call rate for one cell, whose value is calculated from the simulation results. The results in these figures are obtained for 100000s of the simulation time.

Figures 6.3 and 6.4 show the average new call blocking rate and average handoff call blocking rate, respectively. For the new call blocking rate, in spite that all MNs, on calling or not, adopts Algorithm 1 or 2, almost same results are obtained for each reserved bandwidth. From the results of the average handoff call blocking rate and new call blocking rate, Algorithm 2 mainly uses the prepared bandwidth for reducing the

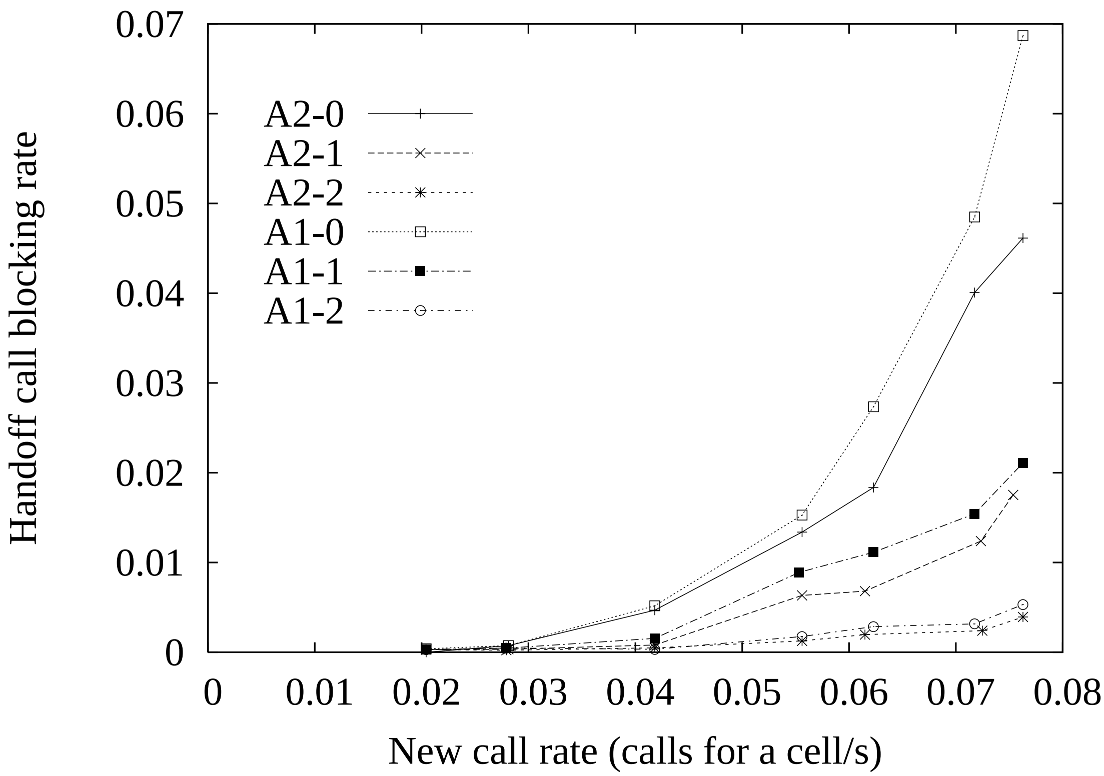

Figure 6.4: Average handoff call blocking rate.

handoff call blocking rate. This is because that the speed of the MN is so high in the mobility model that the effective area produced by Algorithm 2 for reducing the new call blocking is very small.

For the handoff call blocking rate, larger reserved bandwidth provides the smaller average handoff call blocking rate for each algorithm. For the same reserved bandwidth, Algorithm 2 reduces 20–30 % of the average rate at the maximum from Algorithm 1. Especially, when the new call rate is higher, i.e., the cells are more crowed, Algorithm 2 is much effective.

Figure 6.5 shows the average throughput of the entire network. From the simulation conditions, the maximum bandwidth of the total traffic for each cell is restricted to 64Kbps $\times$ 6, and then the one for the entire network is 64Kbps $\times$ 12. For smaller new call rate, all the results in the figure are almost same. For larger new call rate, the reserved band-

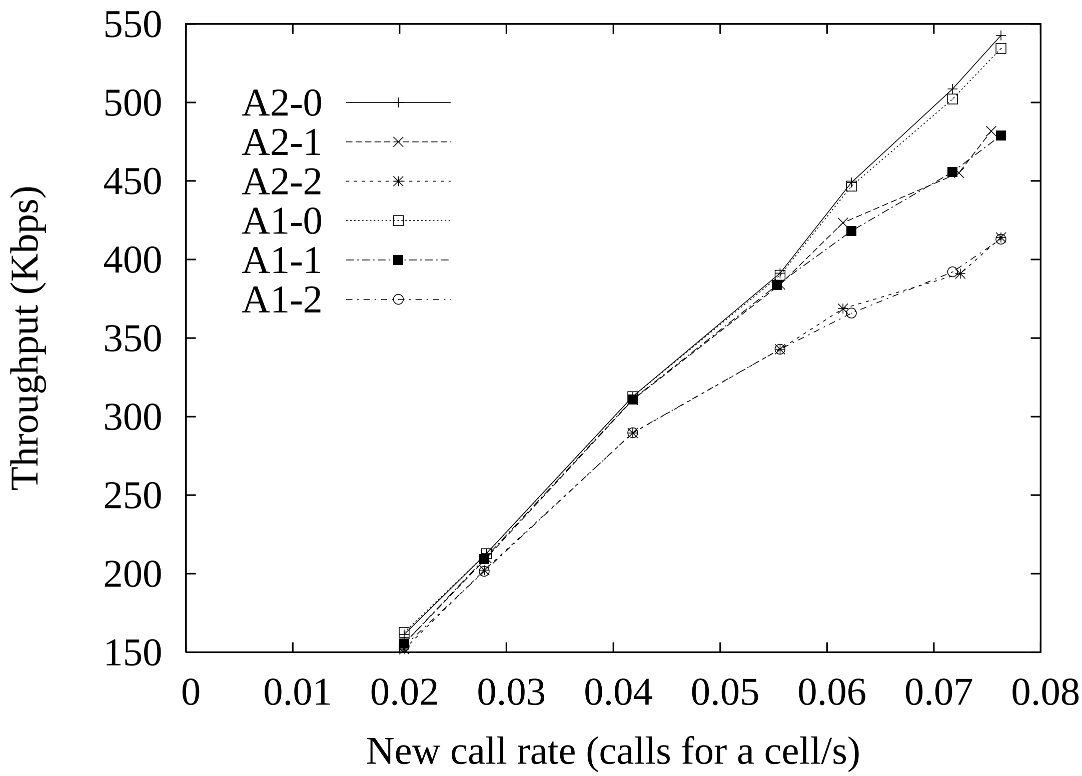

Figure 6.5: Average throughput.

width decreases the average throughput for the both algorithms. For the same reserved bandwidth, the difference for Algorithms 1 and 2 of the throughput is small. However, in Algorithm 2, the bandwidth of the entire network is much shared for handoff calls, but not for new calls. This result strongly supports the purpose of the admission control, and then the bandwidth is effectively used by the control of Algorithm 2.

Figure 6.6 shows the average handoff rate. The average handoff rate has no difference between Algorithms 1 and 2 in spite that the proposed handoff algorithm changes the timing of the handoff, since the average timing of the handoff in Algorithm 2 totally almost equals to that in Algorithm 1 from the experiment conditions.

Figures 6.7–6.9 show the proportion of the crowded conditions, i.e., "crowded," "not crowded" and "same conditions" in "Crowded FA Decision Algorithm." The rate of each condition in the figures is the average of all MNs' rates. Since the decision algorithm is

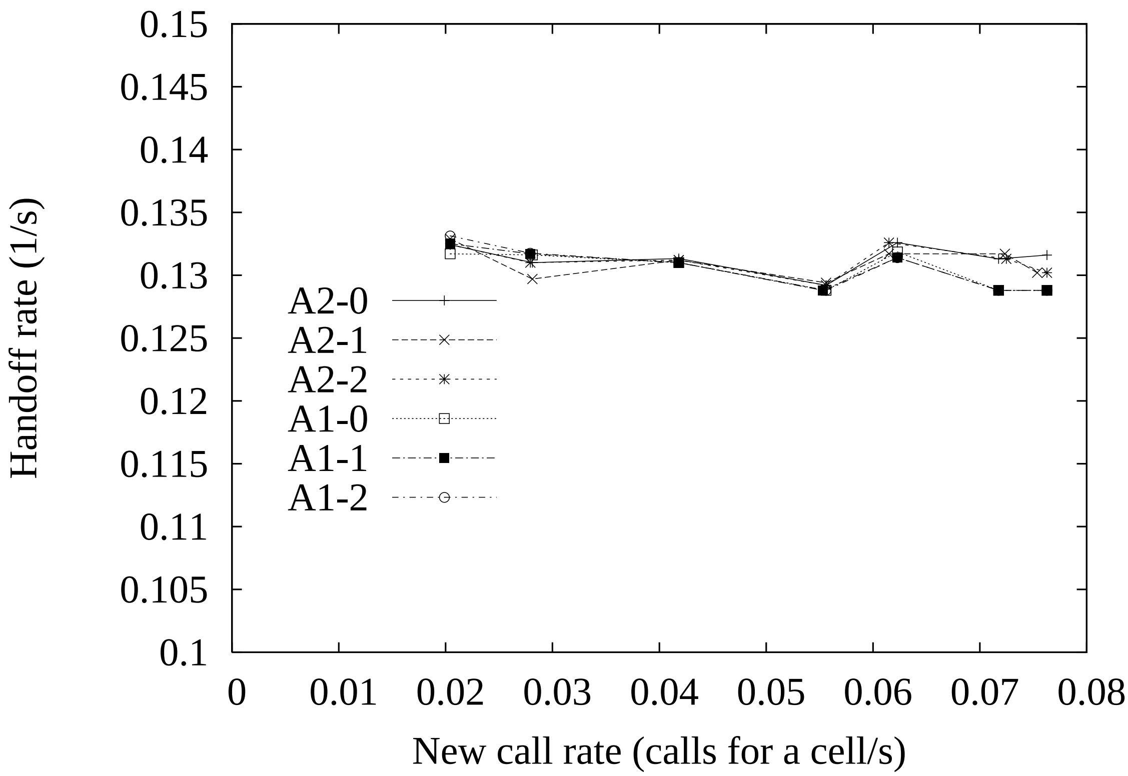

Figure 6.6: Average handoff rate.

applied every time when each MN receives an agent advertisement, in the cases that the MN does not stay at the overlap region are included in the rate of "same conditions." From the geographical conditions, the overlap region occupies 44% of the square. Thus, the maximum of total rates in the cases of "crowded" and "not crowded" is also 44%. From Figure 6.7, the control of the proposed mechanism often works at higher new call rate, since the network is very crowded in this case. From Figures 6.8 and 6.9, the rates of "crowded" and "not crowded" are decreased by larger reserved bandwidth. Therefore, the control of the proposed mechanism is restricted from the case of Figure 6.7. This is because that "crowded bit" is set when the FA has 6 connections in every experiments but this condition should be adaptively change for each reserved bandwidth for handoff calls.

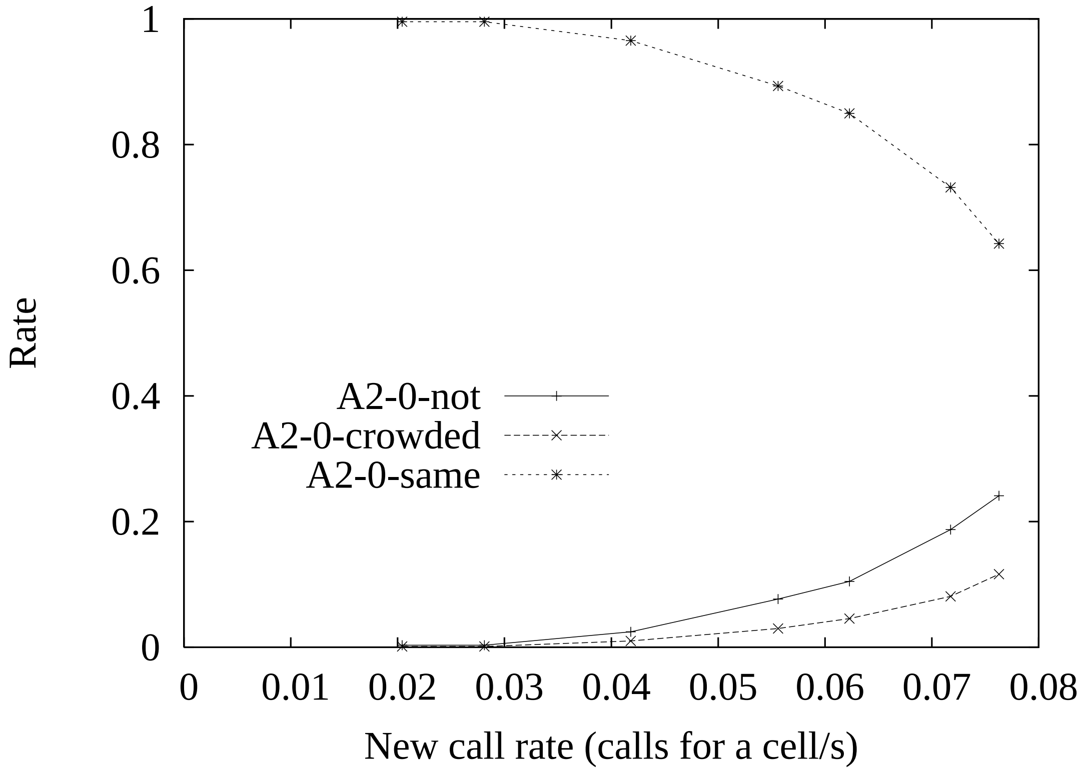

Figure 6.7: Proportion of the crowded conditions when the reserved bandwidth is 64Kbps $\times$ 0.

6.6 Conclusions for Chapter 6

This chapter proposes the adaptive handoff algorithm which the timing of the handoff are dynamically changed by the number of registered MNs in the BS. In order to implement the algorithm for Mobile IP, the system architecture including the QoS signaling protocol, crowded FA decision algorithm and registration procedures for the IP and MAC layers is also proposed. The simulation experiments show that the proposed algorithm provides the lower average handoff blocking rate in spite of low overhead for the proposed expansion. In the simulation model, the traffic model and network topology are quite simple. These experiments for the traffic with multiple QoS classes on a network with many FAs should be executed.

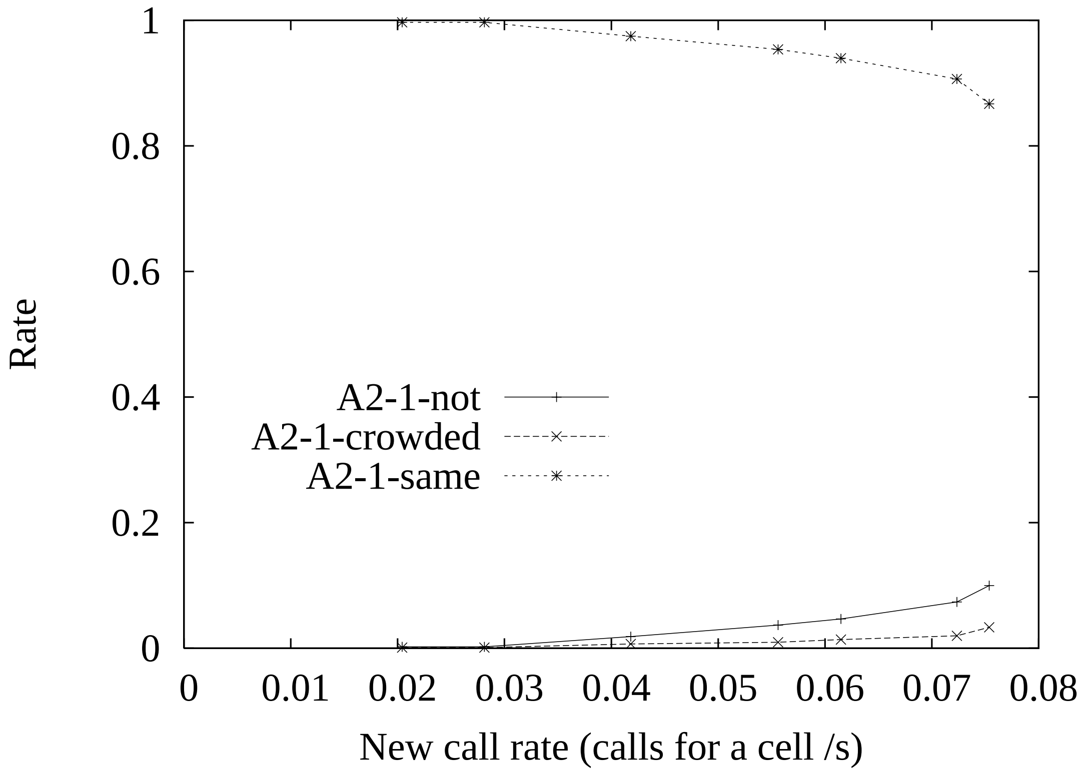

Figure 6.8: Proportion of the crowded conditions when the reserved bandwidth is 64Kbps $\times$ 1.

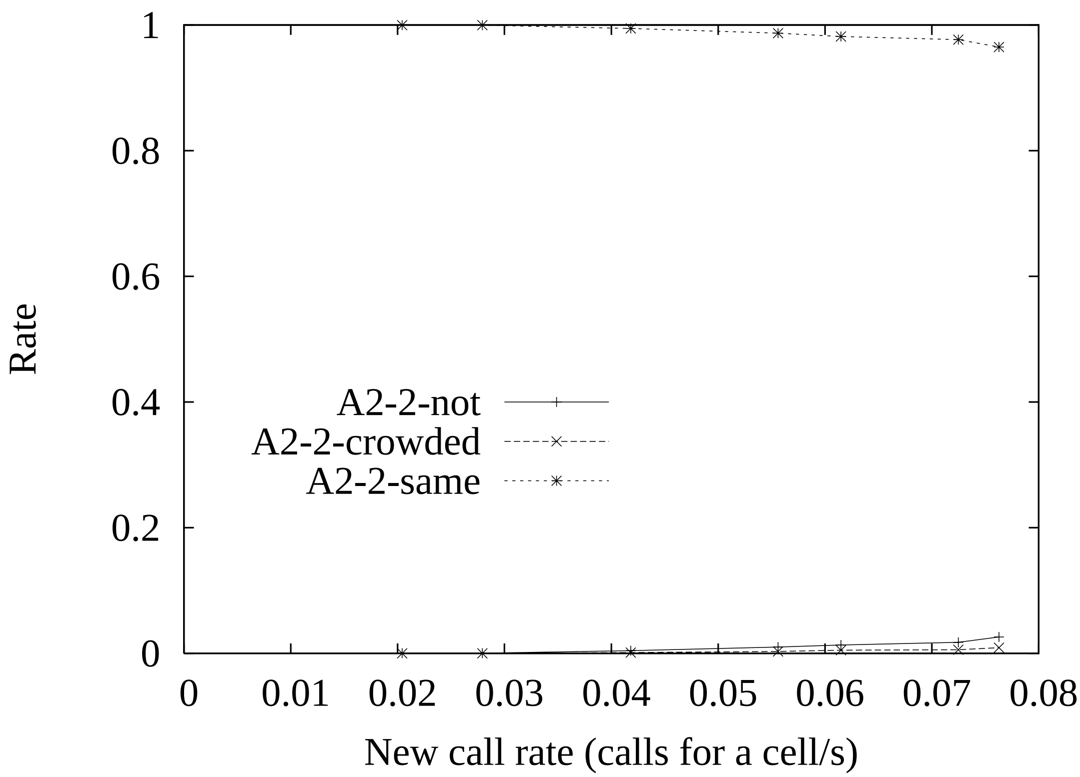

Figure 6.9: Proportion of the crowded conditions when the reserved bandwidth is 64Kbps $\times$ 2.

Chapter 7

Queue Management Method for Improving TCP Performance in Wireless Environments

This chapter proposes a queue management method for improving TCP performance over a wireless link. The proposed method is implemented between the media access control (MAC) layer and the logical link (LL) layer in BS, and designed to help the local retransmission control in the MAC layer. No change is required to TCP itself and servers. From the proposed control, any packet loss in the wireless link is completely eliminated and all packets are delivered in the order. The proposed method keeps trying to consume the available bandwidth of the wireless link at high bit error rate conditions under the ordinal TCP controls.

7.1 Introduction

In the Internet, reliable data communication is mainly derived by TCP[4],[5]. Since TCP is originally designed for wired networks, the control method assumes that a packet loss generally occurs from network congestion and then reduces the congestion window size of the traffic at time. However the needless congestion control also arises when the packet loss is appeared from wireless link error in wireless networks, and always decreases the throughput of the connection.

To improve the performance in the wireless environments, many methods are proposed to adapt the congestion control of TCP for wireless environments. Selective Acknowledgment[33] notifies the multiple lost packets by the acknowledgment (ACK). k-SACK[34] modifies the congestion detection and avoidance mechanisms of SACK with TCP New Reno for wireless environments. In Freeze-TCP[35], a receiver lets a sender keep the congestion window size over a frequently disconnected communication route without any modifications of the sender and intermediate nodes. Snoop agent[36] works at the LL layer in BS and observes a packet loss in the wireless link by checking the sequence number of TCP header for every packet. If the agent detects the loss, the lost packet is locally retransmitted from the cache in the agent. In explicit loss notification (ELN)[37], BS observes a packet loss in a wireless link, and then explicitly notifies the fact to the sender. From the notification, the sender can select the suitable congestion control for the reason of the loss. In [38], ELN is introduced to the above snoop mechanism to notify the occurrence of a packet loss in a wireless link to the sender.

These proposed methods improve TCP performance in wireless Internet environments, but also require the complicated control and a lot of modifications to the conventional systems. Furthermore, these methods are not considered under conditions with a local retransmission control in the MAC layer in spite that many kinds of MAC media have such a retransmission control for the reliable communications.

This chapter presents a queue management method between the LL layer and MAC layer in BS to realize an effective TCP congestion control over a wireless link without any modification to TCP itself nor reference of sequence number information in TCP or LL headers. From the proposed queue management, no packet loss at the wireless link seems to be observed from the end nodes, and all packets are delivered in the order.

In general, the MAC layer repeats the local retransmission control within several times for the reliable communications. The number of the maximum retransmission times is defined for fairness of each flow, however, every packet loss can not be completely covered for a high bit-error link. In this case, packet loss in the wireless link is occurred and the needless congestion control is executed for TCP connections. To release from the problem, the proposed queue management method prepares a queue for each wireless terminal between the MAC layer and LL layer at BS. In the control, when one of the head packet in the queue is transmitted to the MAC layer of the wireless link, the packet is not removed until the queue is informed an acknowledgment of the packet from the MAC layer. Even when the retransmission procedure of the packet exceeds the predefined maximum retransmission times in the MAC layer, the packet is derived at the next cycle of the queues. From the above management, a packet loss is completely eliminated and the sender never receive a duplicate ACK by the wireless link error. In the case that the wireless link error rate is too high, the TCP congestion control should be invoked for the congestion at wired networks or the overflow at the queue in BS.

In this chapter, the queue management method is proposed at first. Then, the computer simulation experiments are shown to evaluate the proposed method.

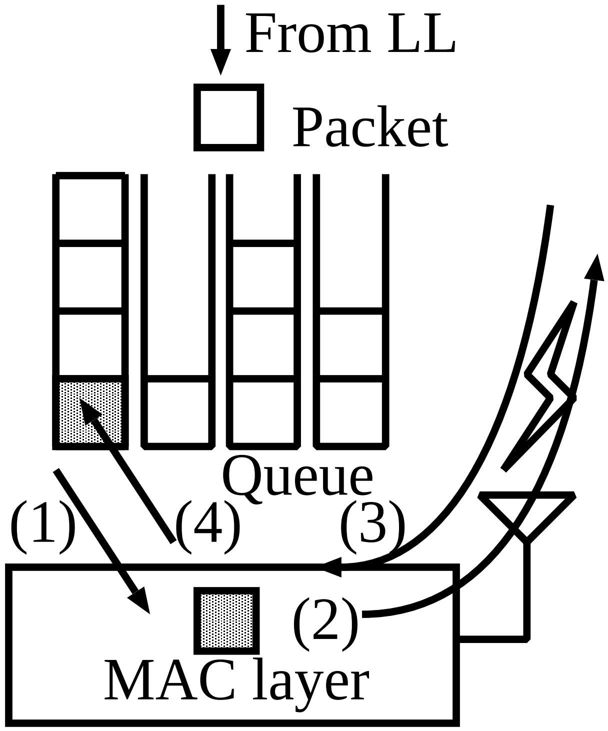

Figure 7.1: Procedures of the queue management method in BS.

7.2 Proposed Queue Management Method

For reliable communications over a wireless link, most kinds of MAC layers prepare the mechanism of the local retransmission, and the maximum number of retransmission times is defined to avoid from wasting the bandwidth by heavy repeat of the retransmissions. In the proposed method, the local retransmission mechanism is enhanced to repeatedly execute the retransmission of the wireless link, to improve TCP performance. This control is tried until each packet is correctly received at the terminal without the bandwidth occupation.

As shown in Fig. 7.1, the proposed method prepares a queue between the MAC layer and the LL layer in BS for each terminal, since the radio condition differs at each terminal. This is also because that a packet to a terminal in extremely bad radio condition may prevent the transmission of packets to the other terminals in case that a common queue is used for all terminals.

When a queue is allowed to send a packet, (1) the head packet of the queue is copied and transmitted to the MAC layer, but the original one is continued to stay at the head of the queue. In the next, (2) the MAC layer forwards the packet to the receiver. If the transmission is succeeded, (3) the acknowledgment is notified from the MAC layer in the receiver, and (4) is also informed to the queue to eliminate the packet from the head of queue. Otherwise, the MAC layer repeats retransmitting the packet, until the acknowledgment is notified.

When the local retransmission can not recover the packet loss, i.e., the acknowledgment in Fig. 7.1(3) can not be received, the retransmission times are exceeded the predefined limit. Then, the packet is eliminated from the MAC layer, and the next queue is allowed to send a packet. Since the eliminated packet is remained in the head of the queue yet, it can be retransmitted to the receiver again when the queue is allowed to send at the next time.

Note that the sequence numbers in TCP and LL headers are not referred in the above procedures. This fact should also be emphasized that the packet order is never changed by wireless link error in the proposed method. Thus, even when the wireless link error occurs heavily, TCP agent in the sender recognizes that the wireless link has narrow bandwidth or requires long round trip time. In such a case, the queue may be filled, and then lost some packets. However, since the TCP congestion control quickly adjusts the flow, the overflows of queue seldom occur. In addition, since all ACK packets are surely delivered to the sender, TCP retransmission timer is hardly expired. Otherwise, the wireless link condition is quite difficult to continue the TCP connection. To confirm this discussion, the next section shows the evaluations of the proposed method.

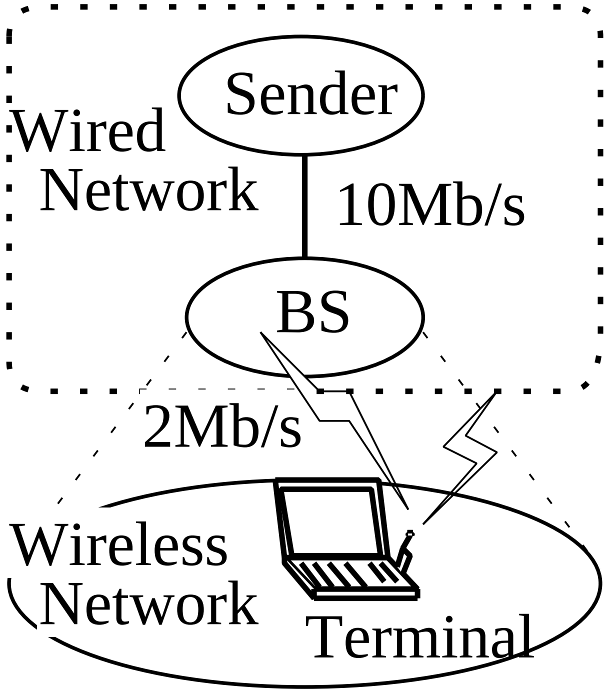

Figure 7.2: Simulation model 6.

7.3 Computer Simulation Experiments

To evaluate the proposed queue management method, the computer simulation experiments are executed with using the network simulator ns-2.1b7. Two queue management methods, the original method and proposed method, are compared for TCP communications under the network model in Fig. 7.2. In these methods, a queue is prepared for each terminal. These queues are selected by Deficit Round Robin (DRR)[39], which is implemented in ns, to send the head packet. The packet is discarded after transmitting it to the MAC layer in the original method, while the proposed method leaves as explained in the previous section.

7.3.1 Simulation Conditions

Common conditions for all simulation experiments are in below.

[Network conditions]

- The transfer speed of each link is depicted in Fig. 7.2. The propagation delay at the wired link is selected from 1ms or 100ms. The distance between each terminal and BS is fixed to 1m.

- For wireless communications, the link error rate is selected among 1/20000 and 1/1000000 error/bit for all terminals.

[Traffic conditions]

- The sender prepares a source for each terminal.

- The source in the sender continuously sends packets to the terminal by FTP under the TCP congestion control. The sizes of all IP data packets and ACKs are 1500 octets and 40 octets, respectively.

[Queue conditions]

- A queue is prepared for each terminal. The each queue is selected by DRR implemented in ns.

- All queues are constructed with a shared memory, and the total memory size to store packets is 150K bytes.

[General conditions]

- Medium access control at the wireless network is IEEE 802.11 implemented in ns.

- The proposed method is implemented in BS.

- The simulation time is 100000s for each experiment.

- Every result satisfies 5% confidence interval with 95% confidence level.

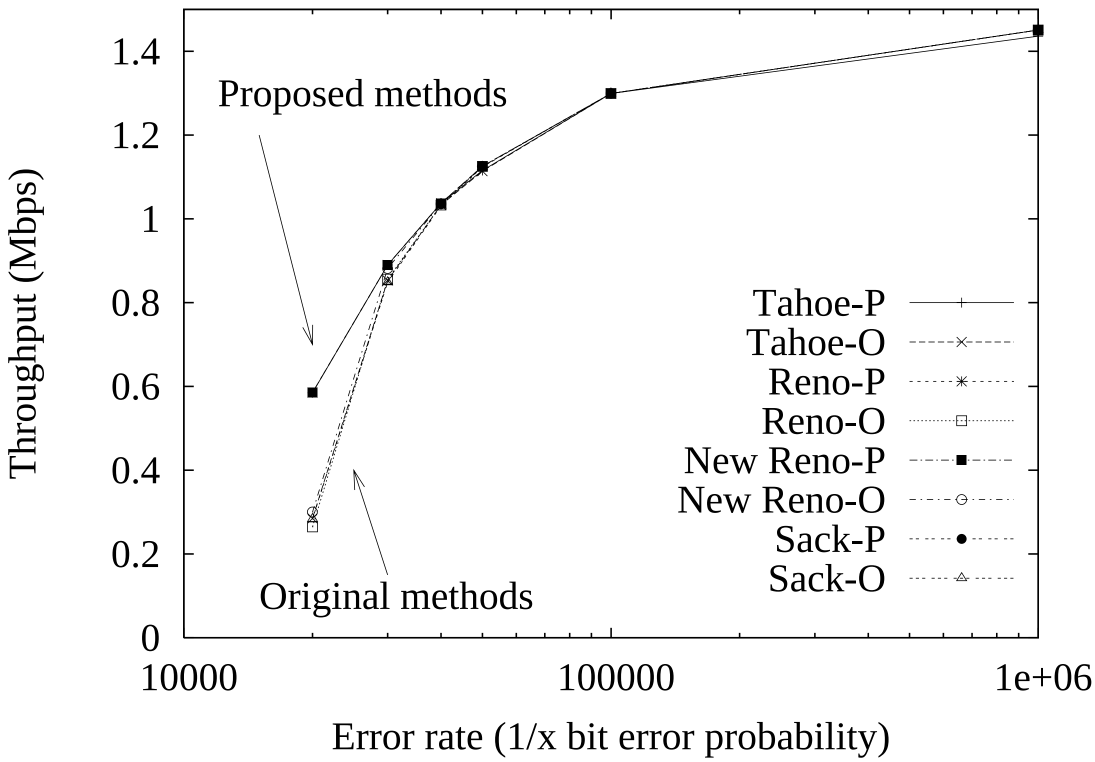

Figure 7.3: Average throughput (delay 1ms).

7.3.2 Basic Performance Evaluations (One Terminal Model)

At first, one terminal model is simulated, where four versions of TCP (Tahoe, Reno, New Reno, SACK) are used. Figs. 7.3 and 7.4 show the average throughput when the delay of the wired link is 1ms and 100ms, respectively. In these graphs, "Tahoe-P" means Tahoe with the proposed method, and "Tahoe-O" is same as "Tahoe-P" except with the original method. The other ones are just replaced the TCP version instead of Tahoe.

From Figs. 7.3 and 7.4, all methods decrease the average throughput when the bit error rate becomes higher. The proposed methods improve 150% and 500% of the average throughput against the original method at the maximum, respectively. In the case that the wired link delay is 100ms, the packet loss penalty is larger, however, the average throughput of the proposed method is clearly better than the one of the original method for the higher error bit rate. Even more, the proposed method keeps almost same average

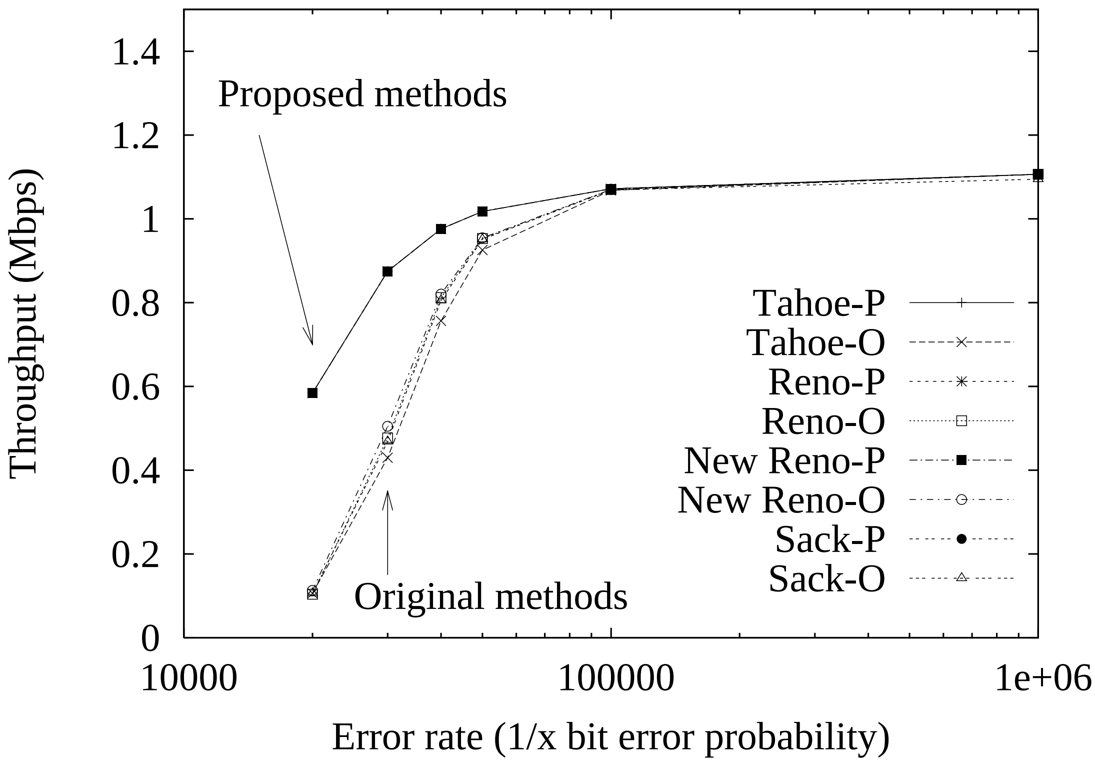

Figure 7.4: Average throughput (delay 100ms).

throughput for the high error bit rate conditions against the one of the delay of 1ms. From the results, the proposed method succeeds to reduce the end-to-end retransmission penalty of TCP. Differences among TCP versions in both methods are hardly observed, because IEEE802.11 has the local retransmission mechanism and maximumly 5 or 7 (depend on a flame size) retransmission times let the differences of the versions be very small. Note that, in [36], [37] and [38], original TCP performance becomes extremely bad when one of ten packets loss is occurred, since their evaluations ignores the local retransmission mechanism in the MAC media. However, these results show that the local retransmission control in the wireless LAN keeps 90% of the throughput from the maximum even when 1/100000 bit error (1500/100000 packets loss probability) is occurred.

In the next, the queue length of each method is discussed. Figs. 7.5 and 7.6 show the minimum queue length, Figs. 7.7 and 7.8 show the average queue length, and Figs. 7.9 and

7.10 show the maximum queue length in BS, when the wired link delay is 1ms and 100ms, respectively. Note that the total queue length is sampled for every 100ms to obtain the average queue length.

From Fig. 7.5, the minimum queue length of the proposed method is not reduced to 0 in the high error rate condition for the delay of 1ms. This means that BS always tries to send a packet to the terminal. Thus, since the proposed method keeps consuming the available bandwidth in the MAC media, the average throughput in Fig. 7.3 should be the maximum bandwidth under the high error rate condition. From Fig. 7.6, the minimum queue length for the delay of 100ms is 0 in the both methods, however, the average throughputs of Figs. 7.3 and 7.4 are almost same in the high error rate condition. In fact, the average and maximum queue lengths for the proposed method are quite larger against the original one in this condition. Therefore, the minimum queue length for the proposed method is just temporary result.

From Figs. 7.7 and 7.8, larger queue size is required for 1ms link delay than that for 100ms in all error rates, because longer link delay plays like a queue. In each link delay, the original method achieves a peak length when the error rate is 1/50000 or 1/40000. While the local retransmission covers the packet loss, TCP does not need to retransmit packets nor reduce the congestion window, and then the packets in the queue must wait for longer time until the local retransmission is completed. In higher bit error rate conditions, the local retransmission can not cover the packet loss and the congestion window tends to be small. Then, the queue length is shorten. In the proposed method, the queue length increases for the high bit error rate, because the retransmission of packets from the queue to the MAC layer occurs many times, and then the control makes packets stay in the queue for long time. In the delay of 100ms, the average queue length also increases at the

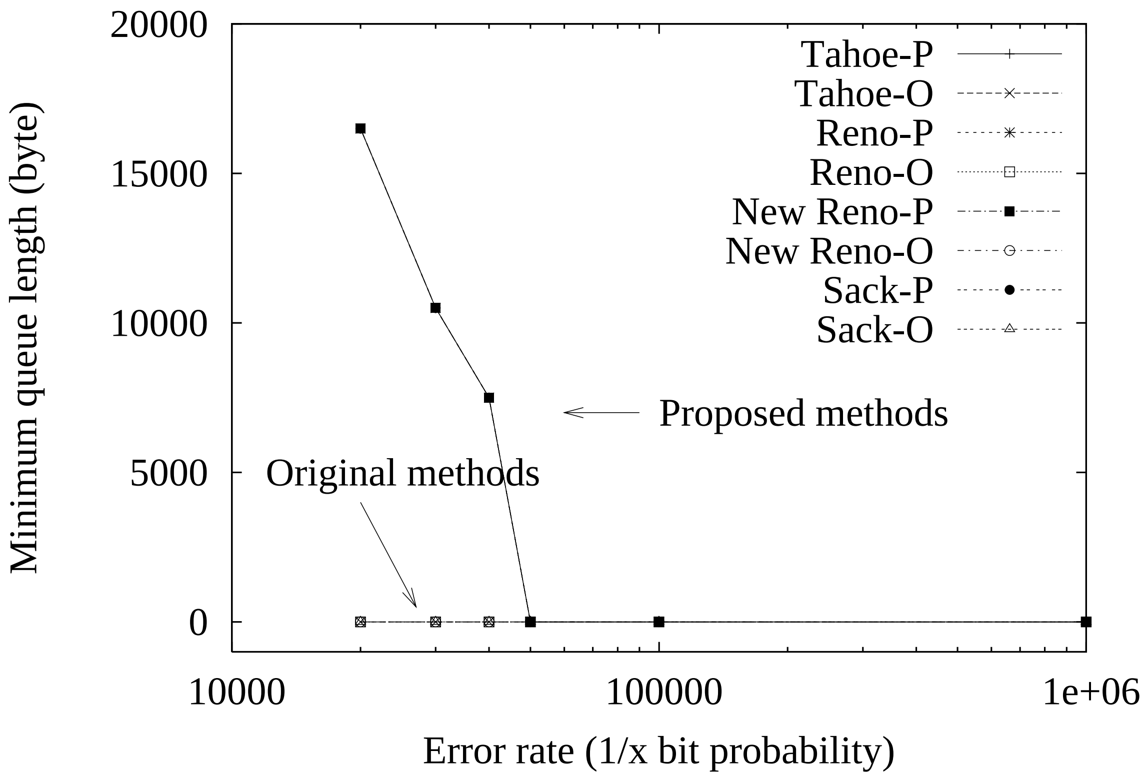

Figure 7.5: Minimum queue length in BS (delay 1ms).

high error bit conditions. Since the long end-to-end delay disturbs the rapid congestion control by many times of the local retransmission, the average queue length increases rapidly. However, the increment will be saturated by the congestion controls of TCP.

From Figs. 7.9 and 7.10, the maximum queue length is 30000 bytes (= 20 packets) in all conditions except to the original method when the error rate is 1/20000. Therefore, it is sufficient that the queue has memory capacity to be able to store more than 20 packets for one flow in the both methods.

Tables 7.1 and 7.2 show the efficiency of transmitted packets defined as (the number of the effective packets) / (the number of all transmitted packets). In the original methods, the efficiency decreases against the high bit error rate, since extra packets are retransmitted by the TCP congestion control for the wireless link errors. Under the proposed queue management control, all lost packets are completely recovered by retransmissions at the

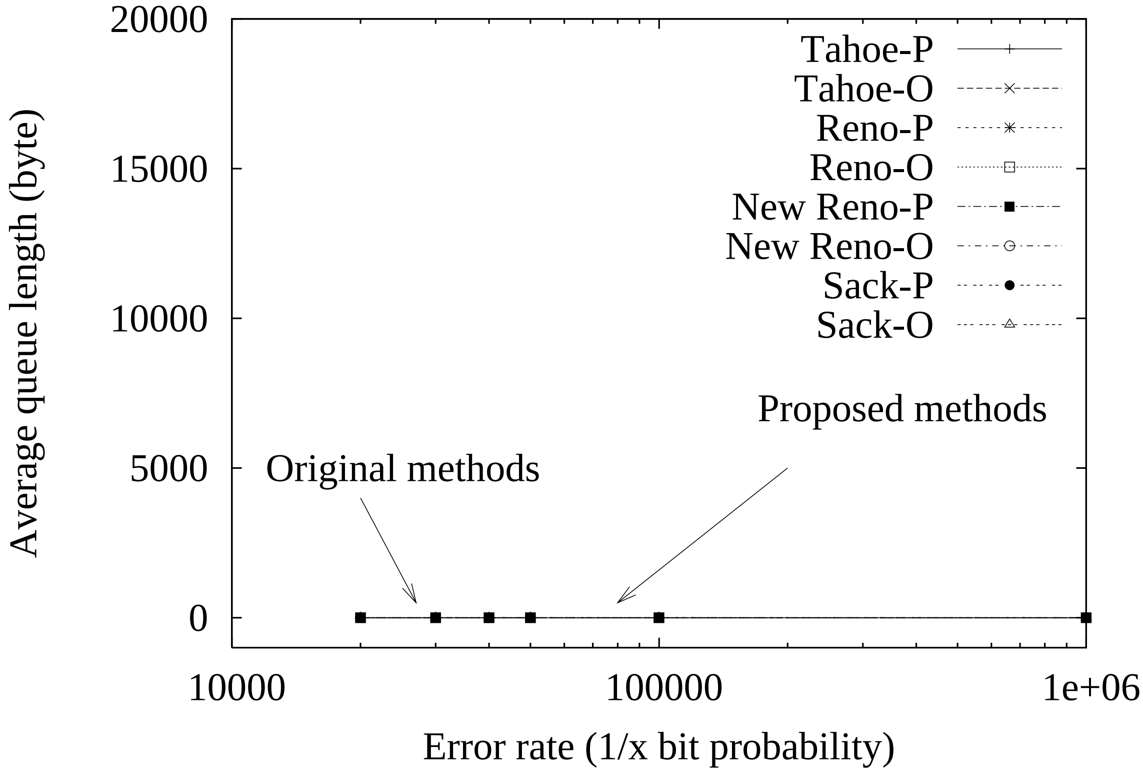

Figure 7.6: Minimum queue length in BS (delay 100ms).

MAC layer. Thus, no packet loss in the wireless link is observed at the TCP layer level and then no extra packet is retransmitted by TCP. This fact also helps for the efficient usage of the network resources.

Then, how the proposed method is influenced from the limited memory size for each queue?

Figs. 7.11–7.18 are the results of TCP Reno with the same conditions in above, where the total memory size in BS to store packets is limited to 20000 and 25000, respectively. The results of the other TCP versions are omitted since the difference among the versions are not remarkable. From Figs. 7.11 and 7.12, when the error bit rate is lower than or equal to 1/100000, the average throughput for the cases of the limited queue length is 5–10% less than the one for the case that the maximum queue size is 150000 by the congestion control of TCP. However, the average throughput is almost same for the high

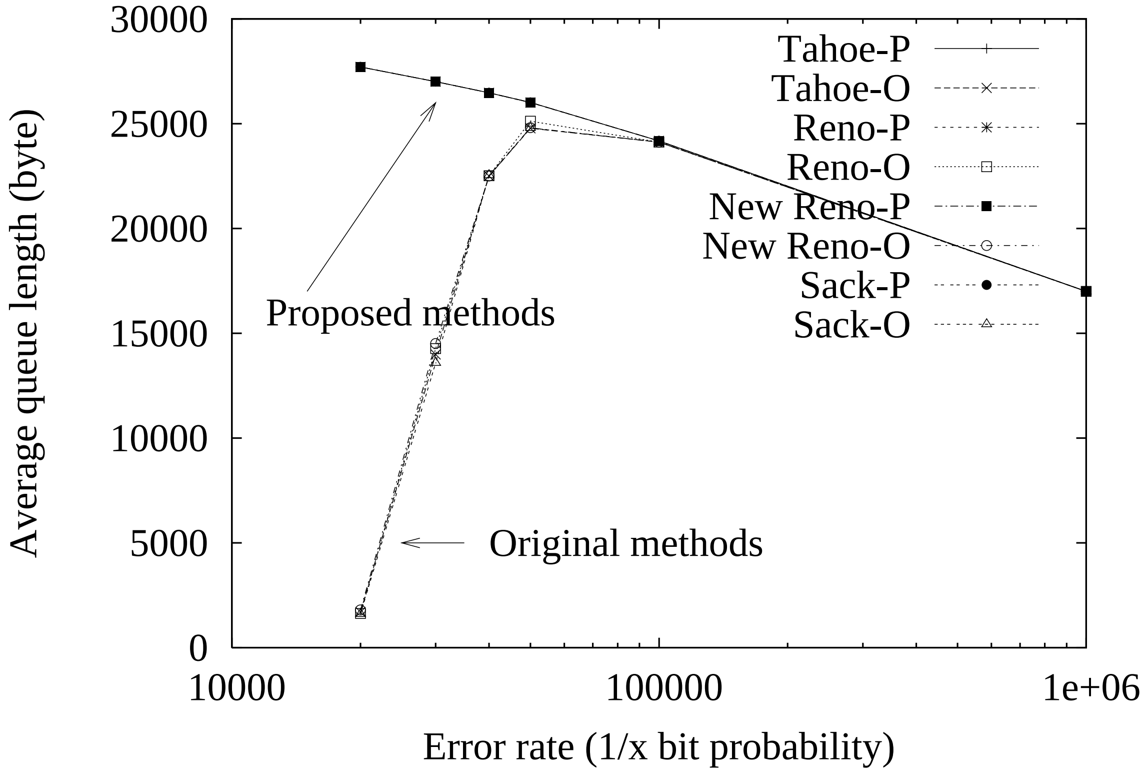

Figure 7.7: Average queue length in BS (delay 1ms).

error bit rate condition in each queue management method. This is because that the bad wireless link condition mainly influences to restrict the average throughput rather than the congestion control of TCP under the limitation of queue length. Thus, the proposed method works well, even when both network congestion and wireless link loss occur at the same time.

In Figs. 7.13 and 7.14, the minimum queue length is reduce to 0, in the cases that the maximum queue length is limited to 20000 and 25000 bytes. Then the wireless link may not keep consuming the available bandwidth.

In Fig. 7.15, the average queue length is about half of the one when the maximum queue length is 150000. It should be noted that the average queue length of the proposed method is about one packet longer than the one of the original method, when the maximum queue length is limited to 20000 and 25000 bytes in Fig. 7.15. This is because that a transmitted

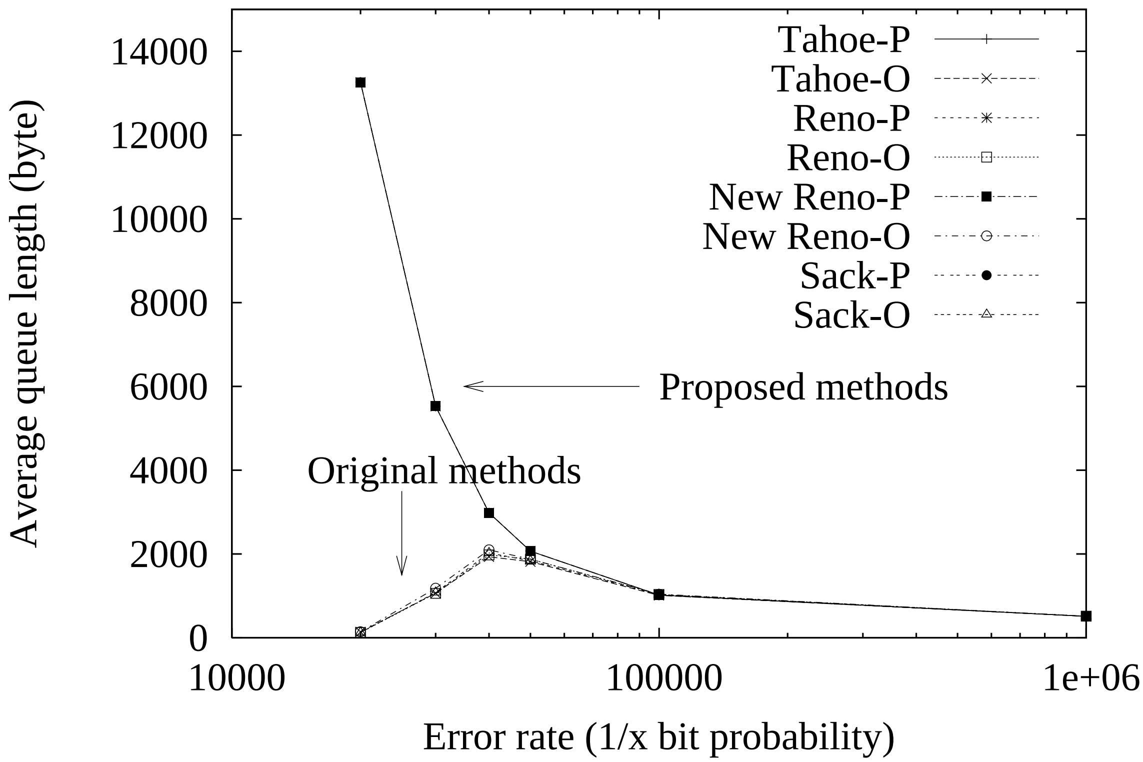

Figure 7.8: Average queue length in BS (delay 100ms).

packet is duplicated in the head of the queue in the proposed control. In Fig. 7.16, the proposed method of the average queue length of 20000 is reduced by the control of TCP. However, for each queue management method in Figs. 7.15 and 7.16, the shape of graphs is rather similar.

The maximum queue length in Fig. 7.17 is increased for the high link error rate in the proposed method. The one in Fig. 7.18 becomes also lager, however, the incremental rate must be smaller than that in Fig. 7.17. In these two figures, since the ones from the original method are rather reduced, the proposed method can also effectively use the bandwidth.

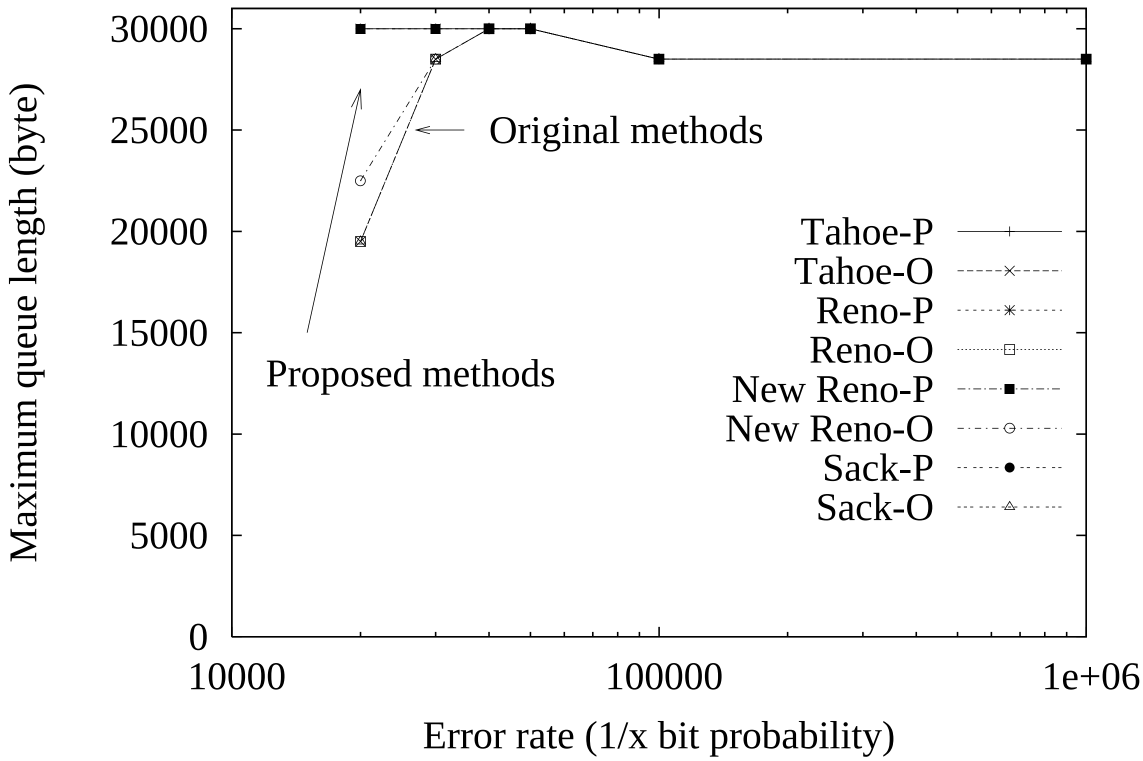

Figure 7.9: Maximum queue length in BS (delay 1ms).

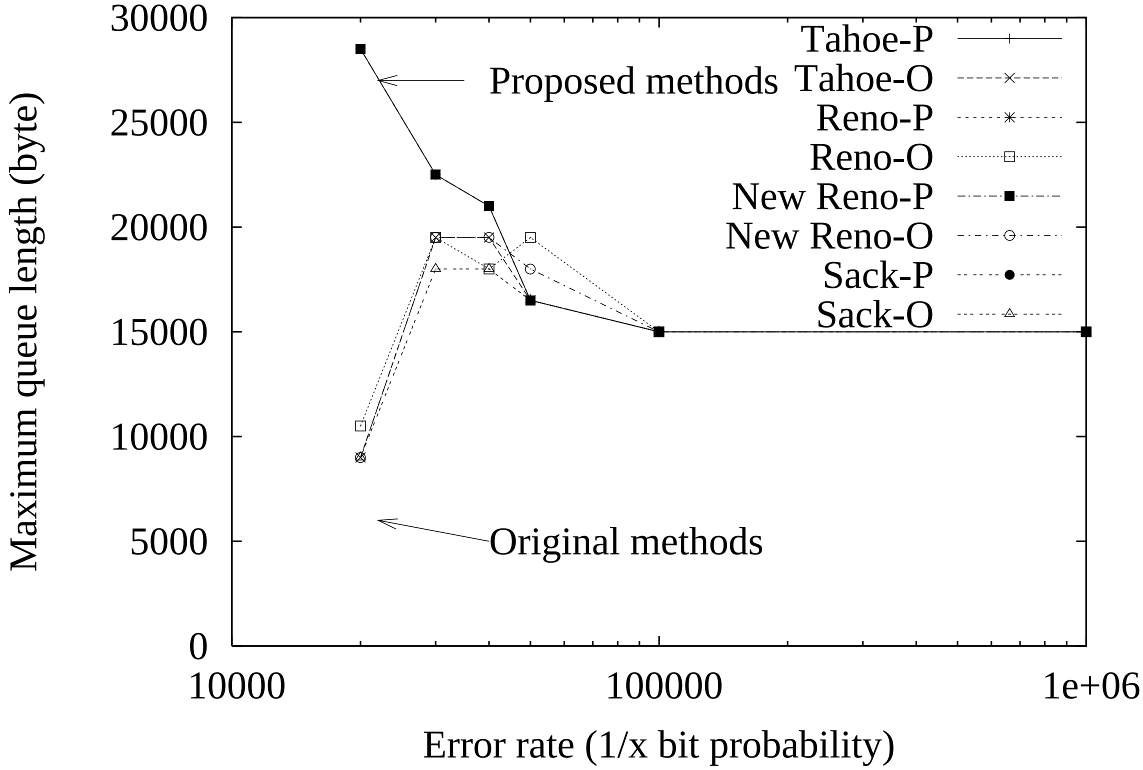

Figure 7.10: Maximum queue length in BS (delay 100ms).

Error rate ($\times 10^{-4}$)	1/2	1/3	1/4	1/5	1/10
Tahoe-P	1	1	1	1	1
Reno-P	1	1	1	1	1
New Reno-P	1	1	1	1	1
Sack-P	1	1	1	1	1
Tahoe-O	0.908	0.988	0.997	0.999	1
Reno-O	0.908	0.988	0.997	0.999	1
New Reno-O	0.908	0.988	0.997	0.999	1
Sack-O	0.908	0.988	0.997	0.999	1

Table 7.1: Average transmission efficiency (delay 1ms).

Error rate ($\times 10^{-4}$)	1/2	1/3	1/4	1/5	1/10
Tahoe-P	1	1	1	1	1
Reno-P	1	1	1	1	1
New Reno-P	1	1	1	1	1
Sack-P	1	1	1	1	1
Tahoe-O	0.908	0.988	0.997	0.999	1
Reno-O	0.908	0.988	0.997	0.999	1
New Reno-O	0.908	0.988	0.997	0.999	1
Sack-O	0.908	0.988	0.997	0.999	1

Table 7.2: Average transmission efficiency (delay 100ms).

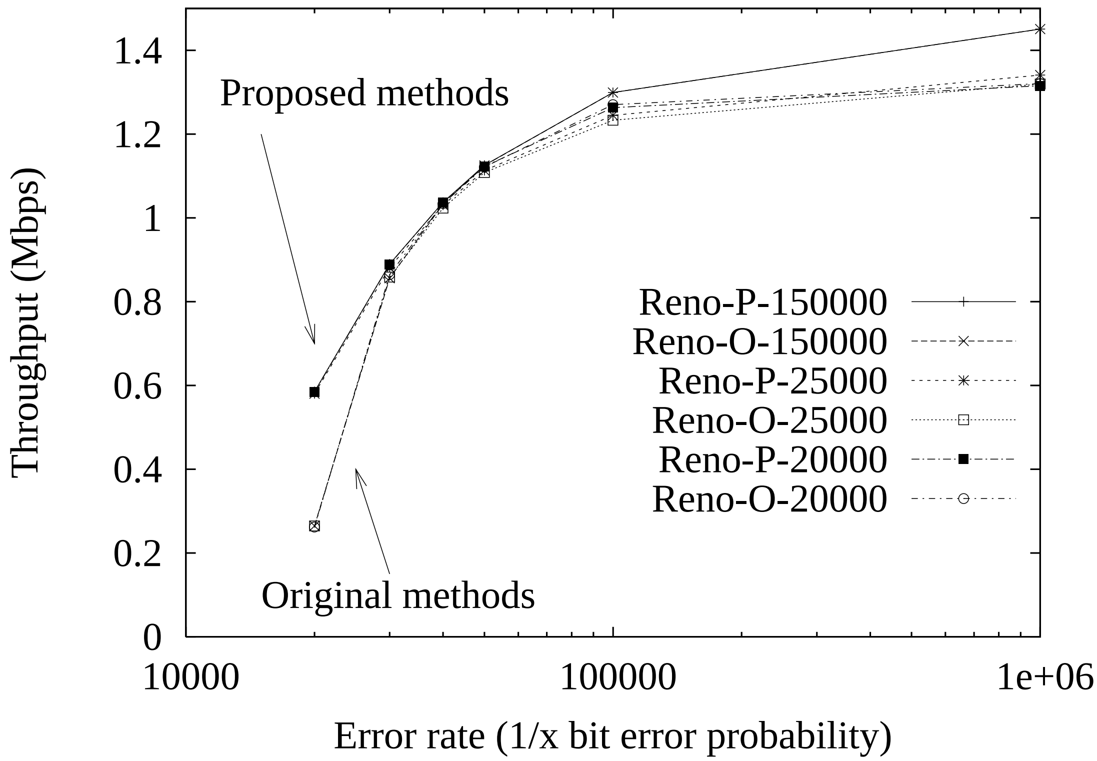

Figure 7.11: Average throughput with the limit of memory size (delay 1ms).

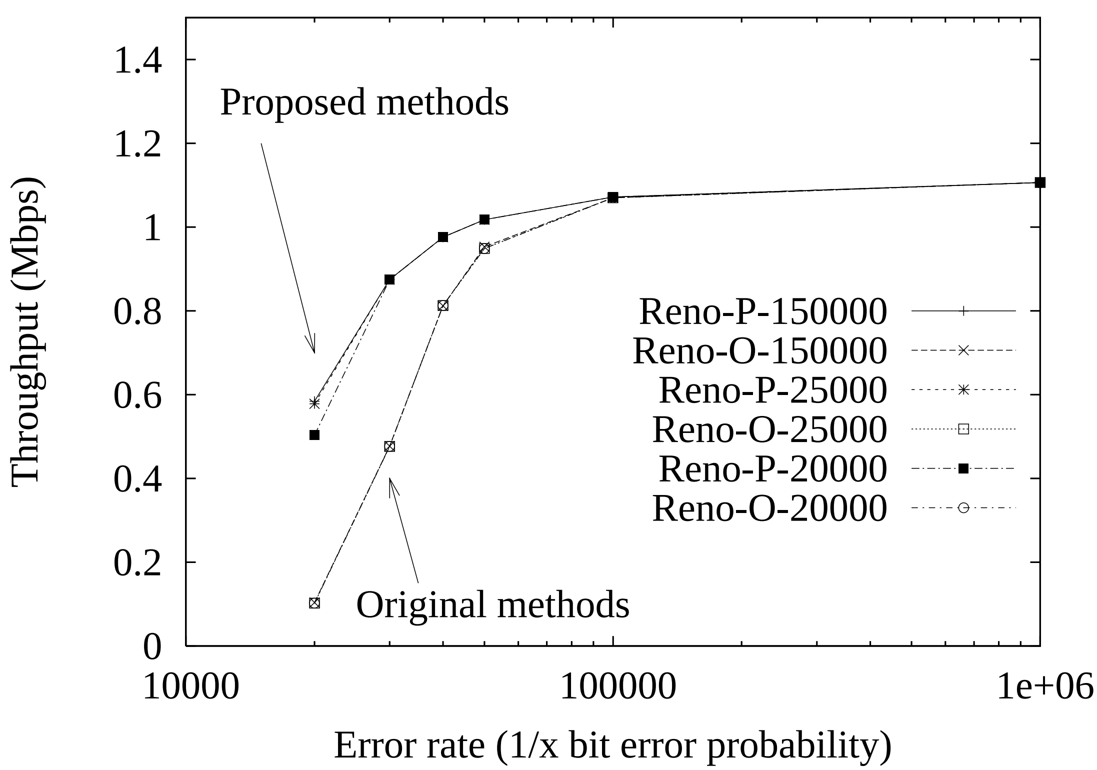

Figure 7.12: Average throughput with the limit of memory size (delay 100ms).

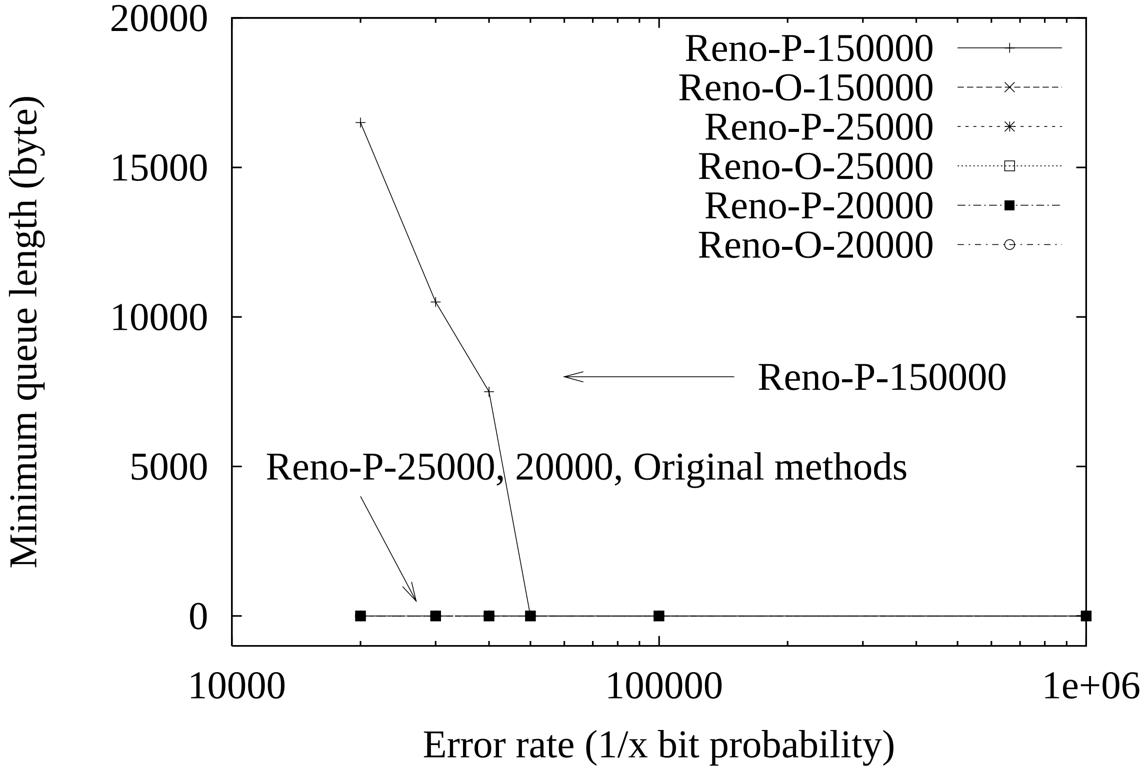

Figure 7.13: Minimum queue length in BS with the limit of memory size (delay 1ms).

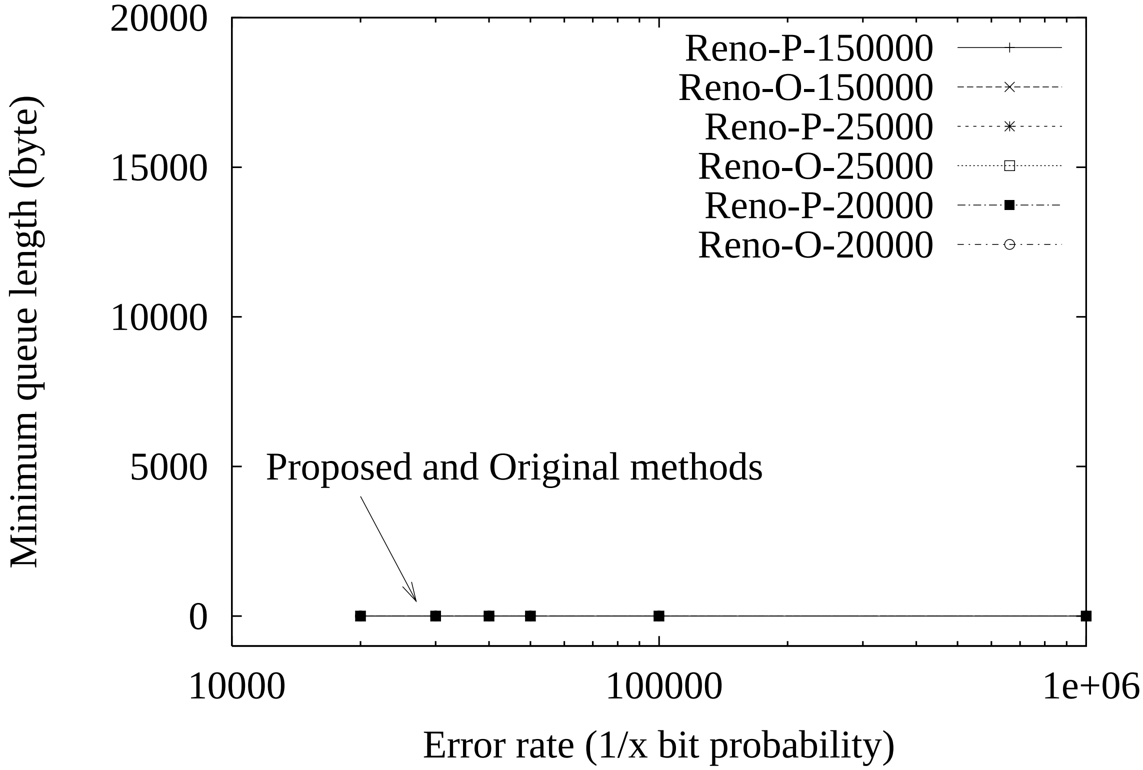

Figure 7.14: Minimum queue length in BS with the limit of memory size (delay 100ms).

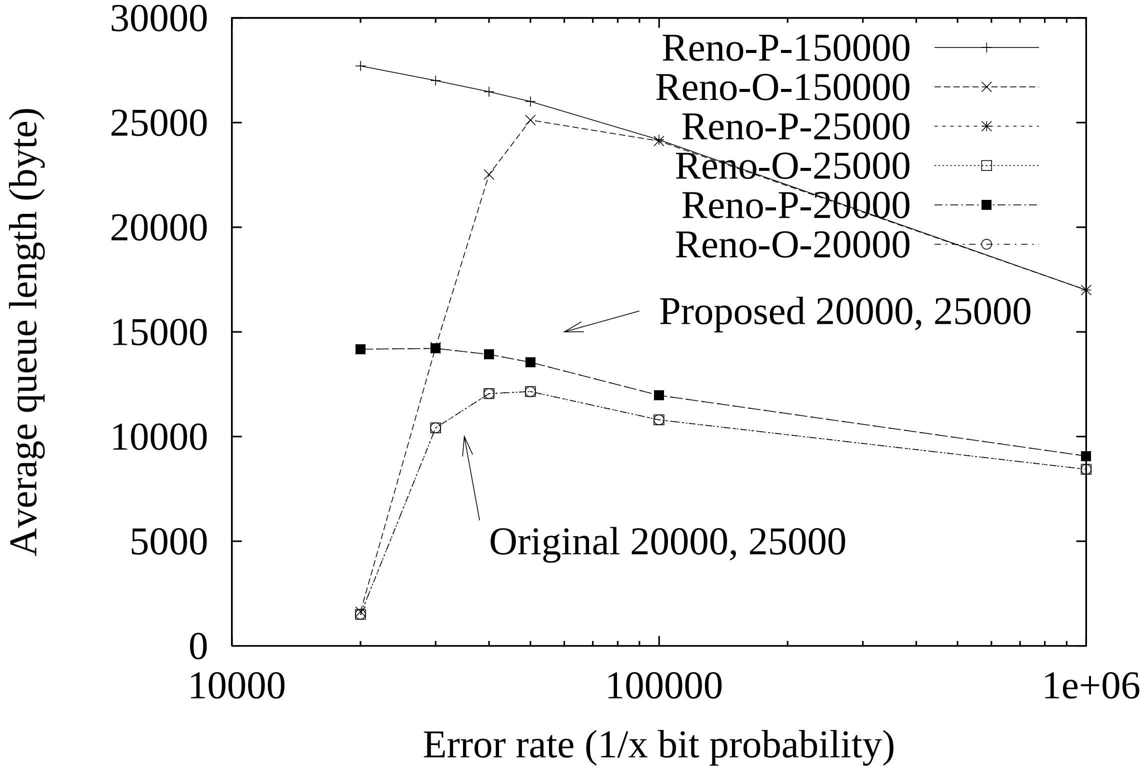

Figure 7.15: Average queue length with the limit of memory size (delay 1ms).

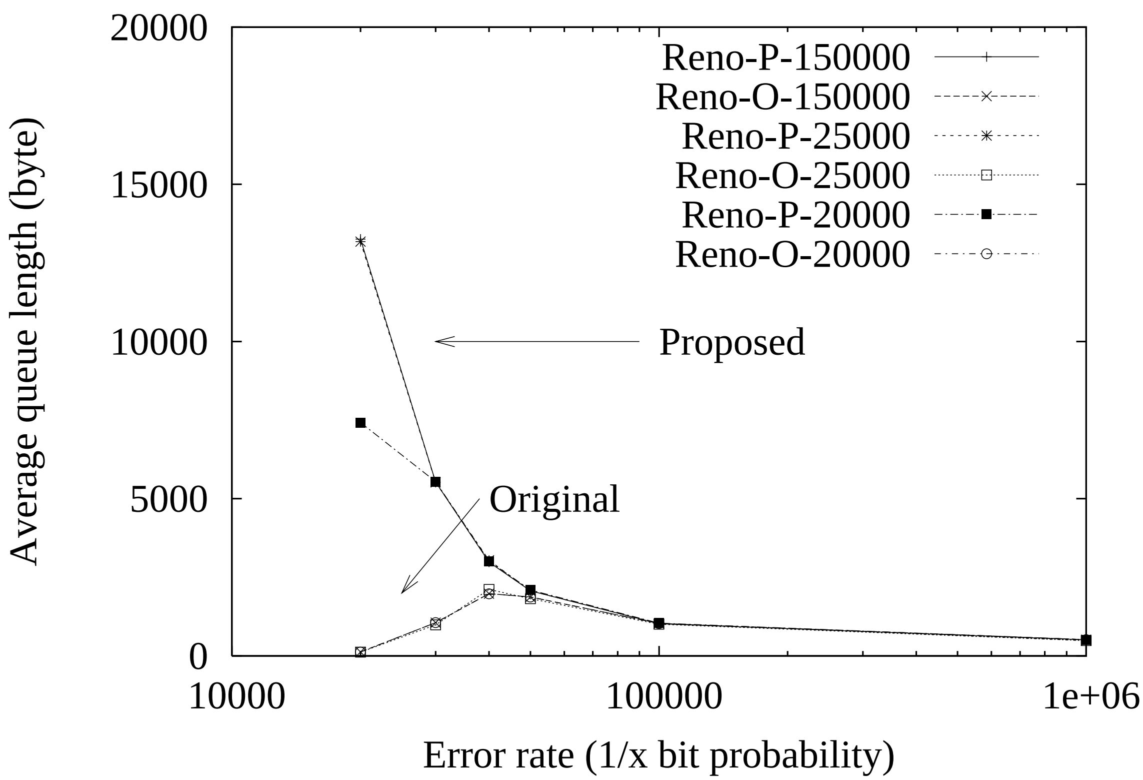

Figure 7.16: Average queue length with the limit of memory size (delay 100ms).

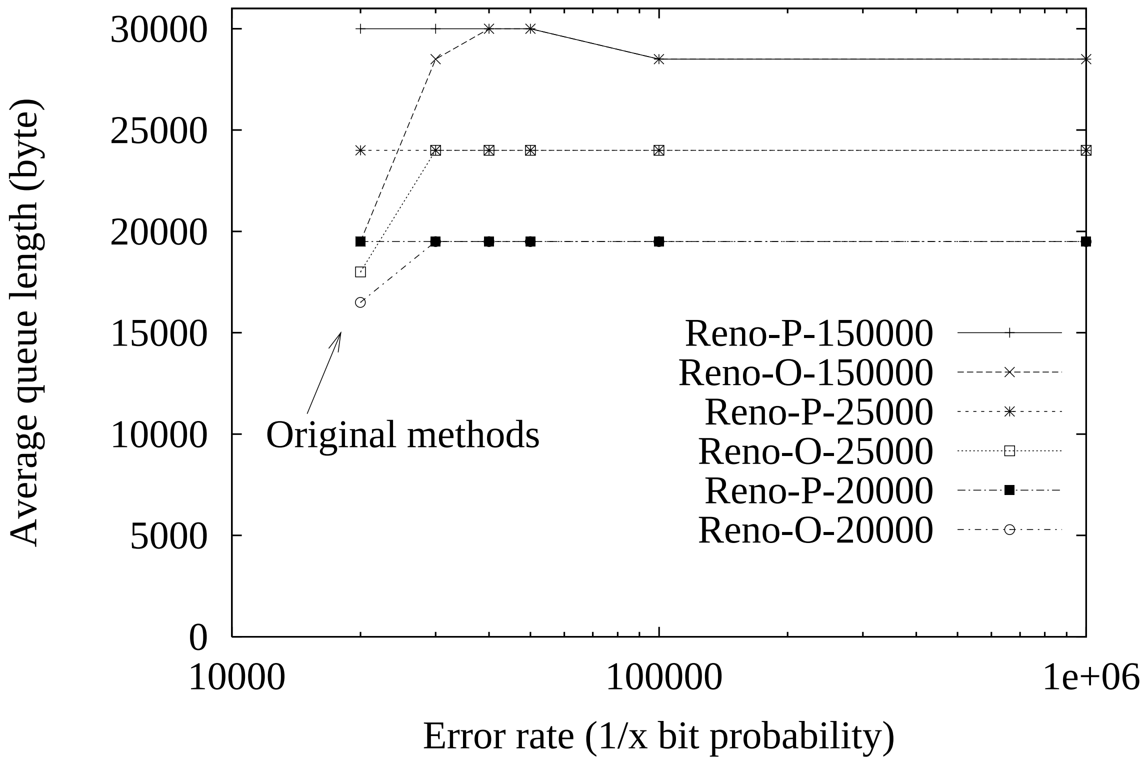

Figure 7.17: Maximum queue length in BS with the limit of memory size (delay 1ms).

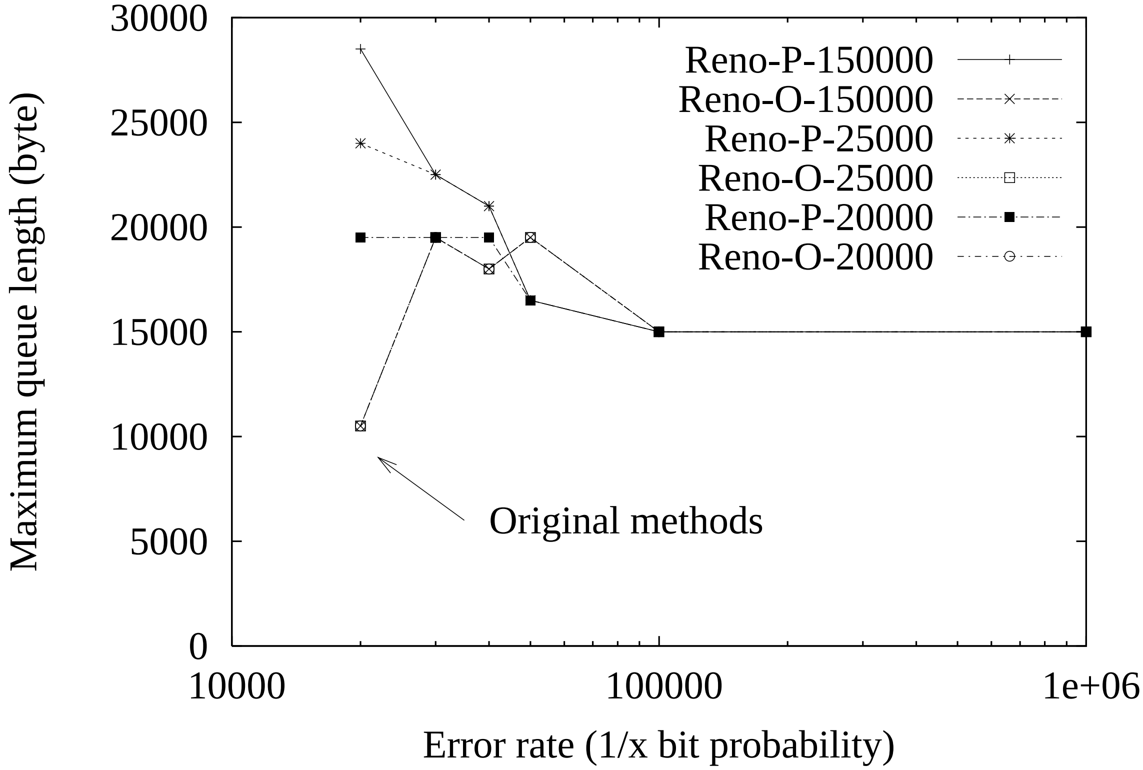

Figure 7.18: Maximum queue length in BS with the limit of memory size (delay 100ms).

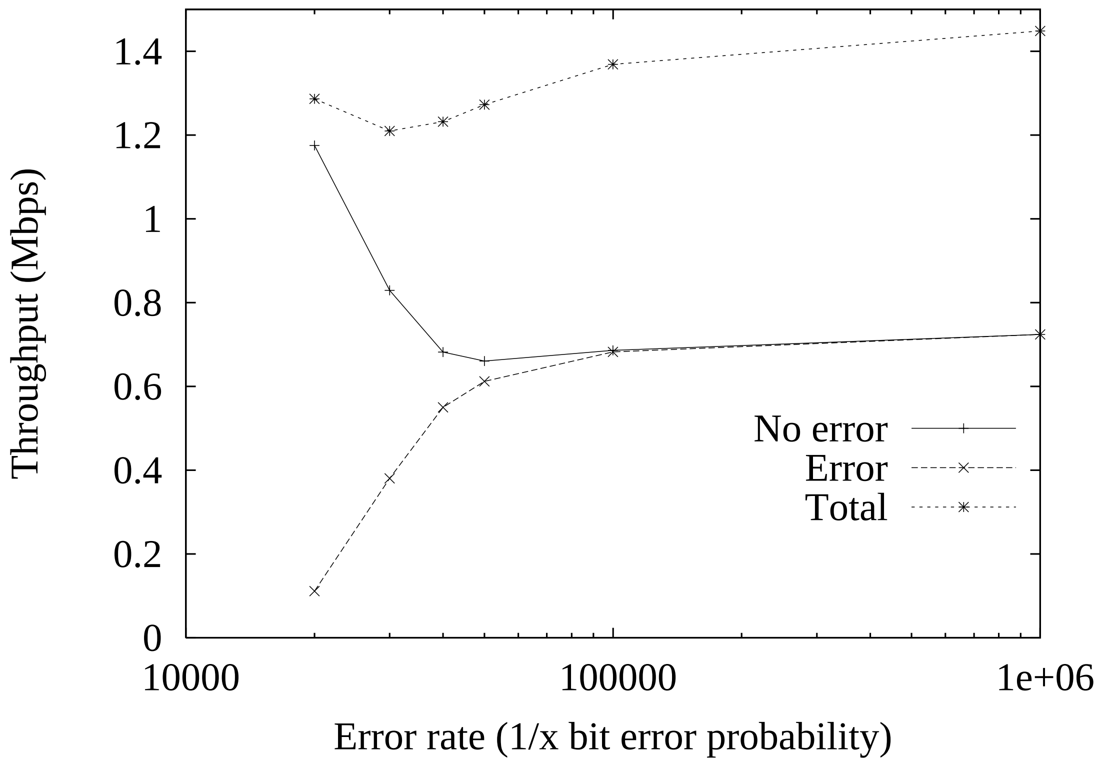

Figure 7.19: Average throughput of the original method when the error rate of two terminals is different (delay 1ms).

7.3.3 Two terminals model

Figs. 7.19, 7.20 and Figs. 7.21, 7.22 depict the average throughput of the original method and proposed method with two terminals when the wired link delay is 1ms and 100ms, respectively. The error bit rate of the one terminal is fixed to 1/1000000, while that of the other terminal is changed from 1/20000 to 1/1000000. For the experiments, only the TCP Reno version is used. When the error bit rate is less than or equal to 1/100000, two terminals fairly share the bandwidth in the wireless link for both methods. On the other hand, in the original method, the average throughput for high-error-rate terminal is quite reduced in both cases of the wired link delay in spite of using DRR as queue management control, since the other terminal obtains the rest of the bandwidth by the TCP congestion control. The tendency is increased in the case of the larger wired link

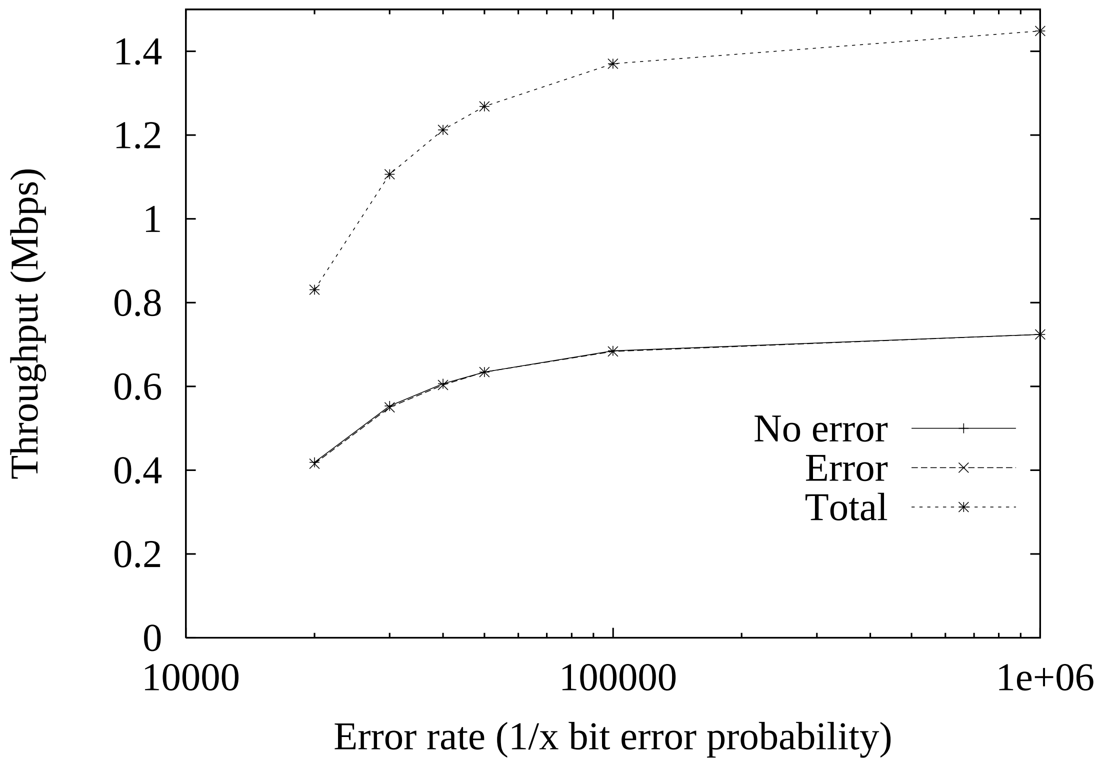

Figure 7.20: Average throughput of the proposed method when the error rate of two terminals is different (delay 1ms).

delay. In the proposed method, the bandwidth is also fairly shared between the two terminals in high bit error rate for the both cases of the wired link delay. At the sender of the TCP agent, the difference of the link condition does not interfere with the end-to-end control because the both terminals can receive packets in the order without loss, and have the same end-to-end delay. However, the average of the total throughput at the high error rate decreases 40% from the maximum. The reduction of the bandwidth is used to retransmit packets at the MAC layer for the high-error-rate terminal.

Figs. 7.23 and 7.24 show the average of the total queue length when the wired link delay is 1ms and 100ms, respectively. As shown in Figs. 7.7 and 7.8, the original method reduces the queue length against the high bit error rate in the both delay cases, since a high-error-rate terminal reduces the average throughput, and then all packets in the queue come only from the other terminal. In the proposed method, the queue length for

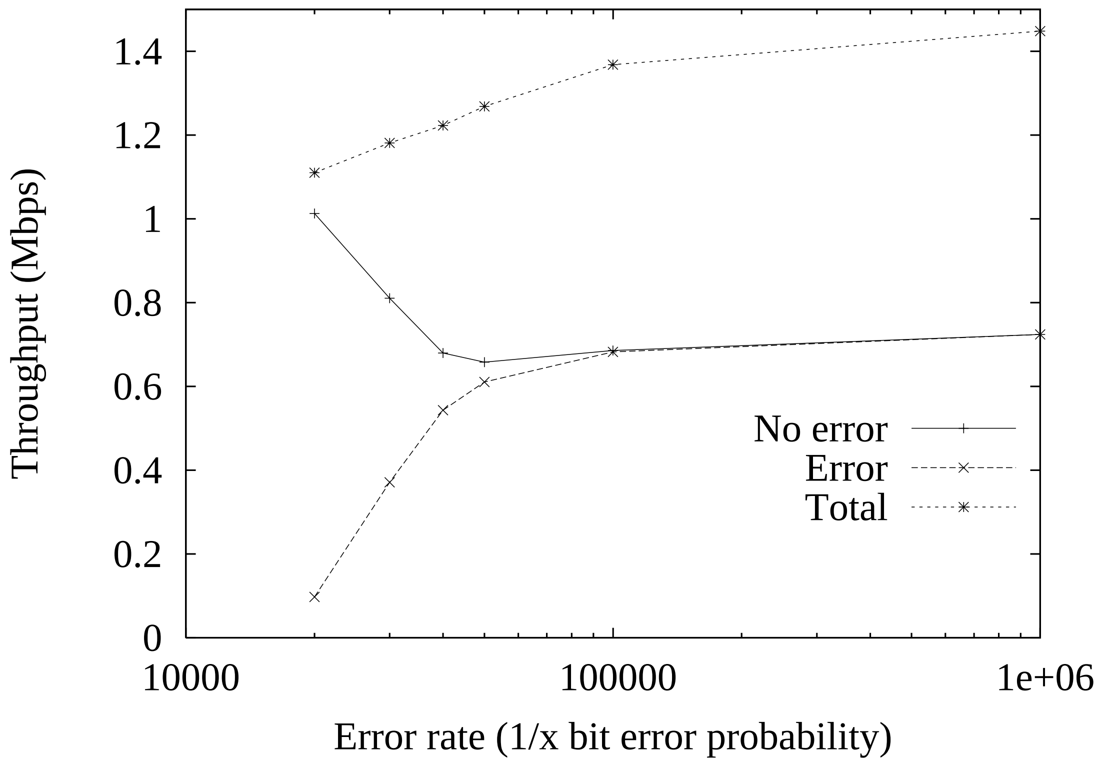

Figure 7.21: Average throughput of the original method when the error rate of two terminals is different (delay 100ms).

the delay of 1ms increases a little for the high error bit, but the increment rate is much smaller than that in Fig. 7.7 due to the effect of the statistical multiplexing. However, for the delay of 100ms, the queue length is increased against the high error bit rate. This is because that the long round trip time delays the congestion control of TCP and then the packet retransmission at the MAC layer also causes that the queue length becomes long.

Figs. 7.25 and 7.26 depict the average of the total throughput, and Figs. 7.27 and 7.28 show the average of the total queue length, when the error bit rates of the both terminals are equally changed from 1/20000 to 1/1000000 with the wired link delay is 1ms and 100ms, respectively. In the proposed method, the total throughput in Figs. 7.25 and 7.26 are almost same as the average throughput in Fig. 7.3. Therefore, the two terminals share the entire bandwidth which can be available at the wireless link at each error rate. In fact, Figs. 7.27 and 7.28 shows that the average of the total queue length is not reduced

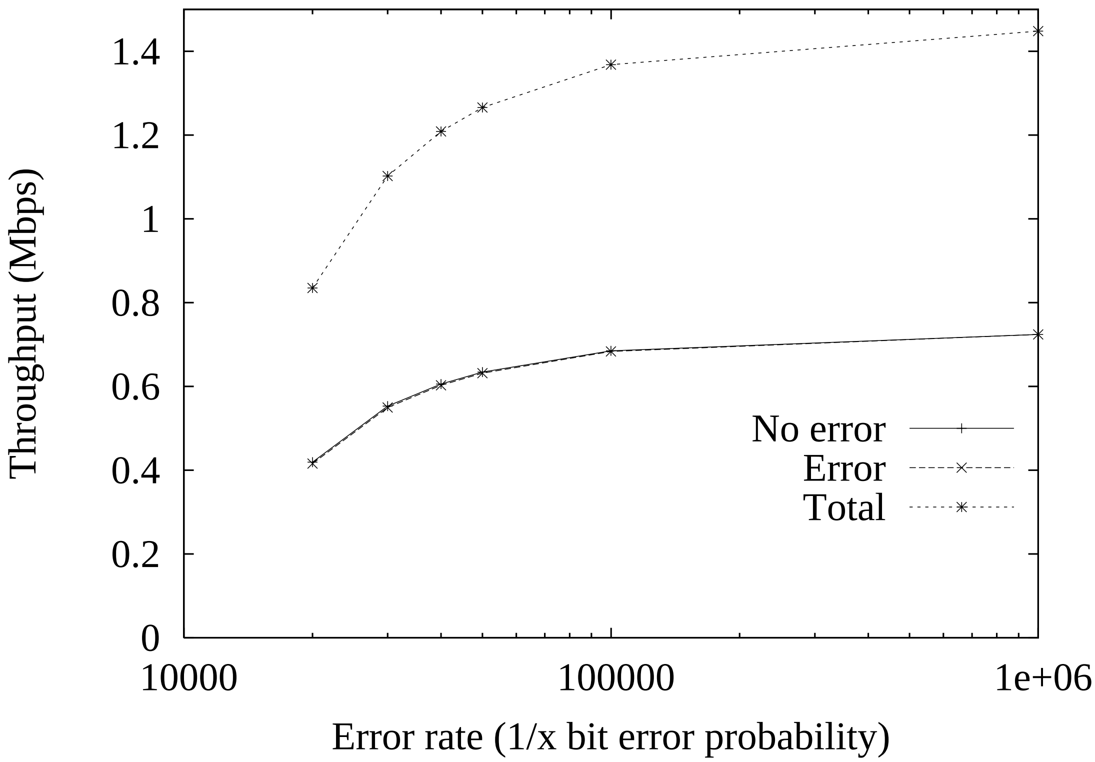

Figure 7.22: Average throughput of the proposed method when the error rate of two terminals is different (delay 100ms).

against the high bit error condition.

The original method increases the total average throughput in Figs. 7.25 and 7.26 against Figs. 7.3 and 7.4 in the high error bit rate, since the values are combined from the throughput from the two terminals. The queue length is decreased in the high error bit rate because of the needless congestion control of TCP. However, the entire bandwidth of the wireless link is not completely used at the two terminals model. These are clearly observed in results of the wired delay 100ms, since the queue length and throughput are reduced for the high bit error rate.

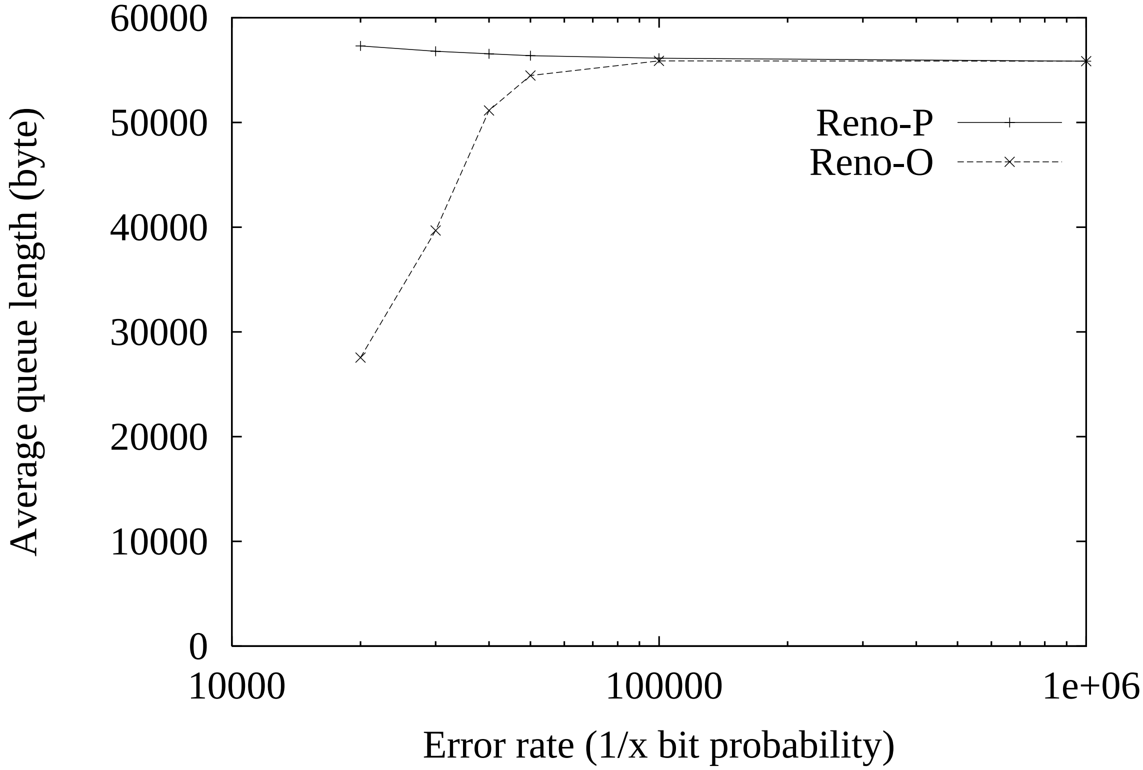

Figure 7.23: Average of total queue length when the error rate of two terminals is different (delay 1ms).

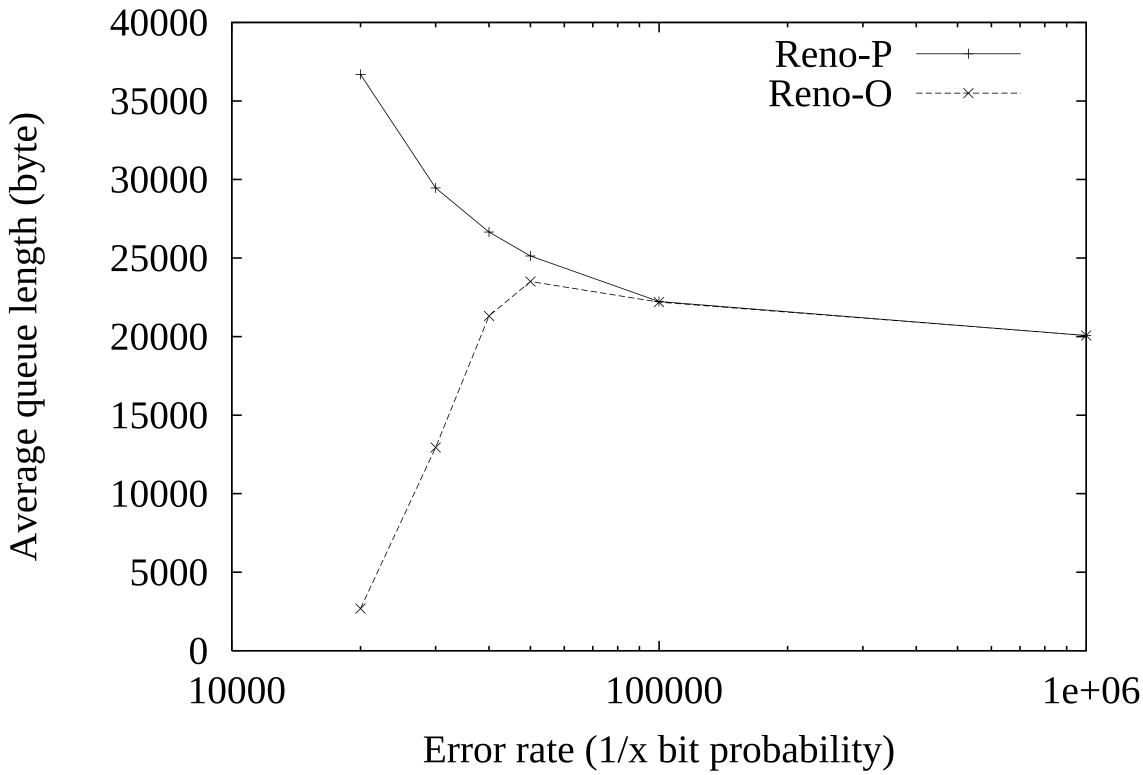

Figure 7.24: Average of total queue length when the error rate of two terminals is different (delay 100ms).

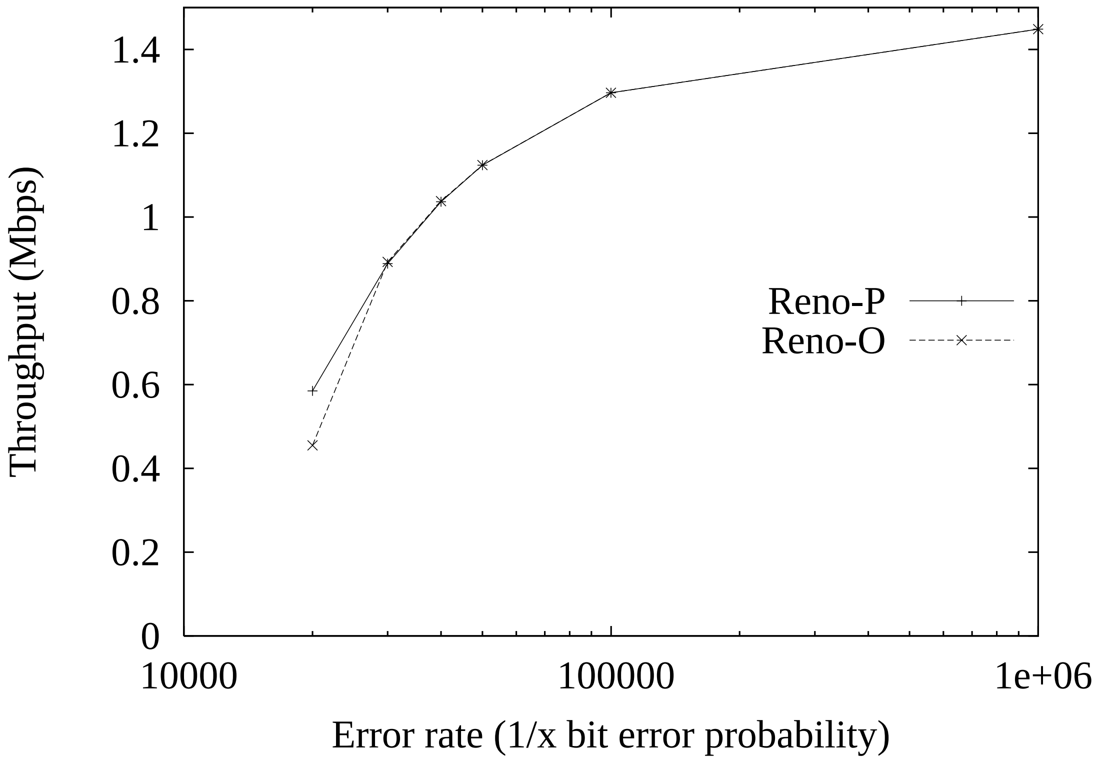

Figure 7.25: Average of total throughput (delay 1ms) when the error rate of two terminals is same.

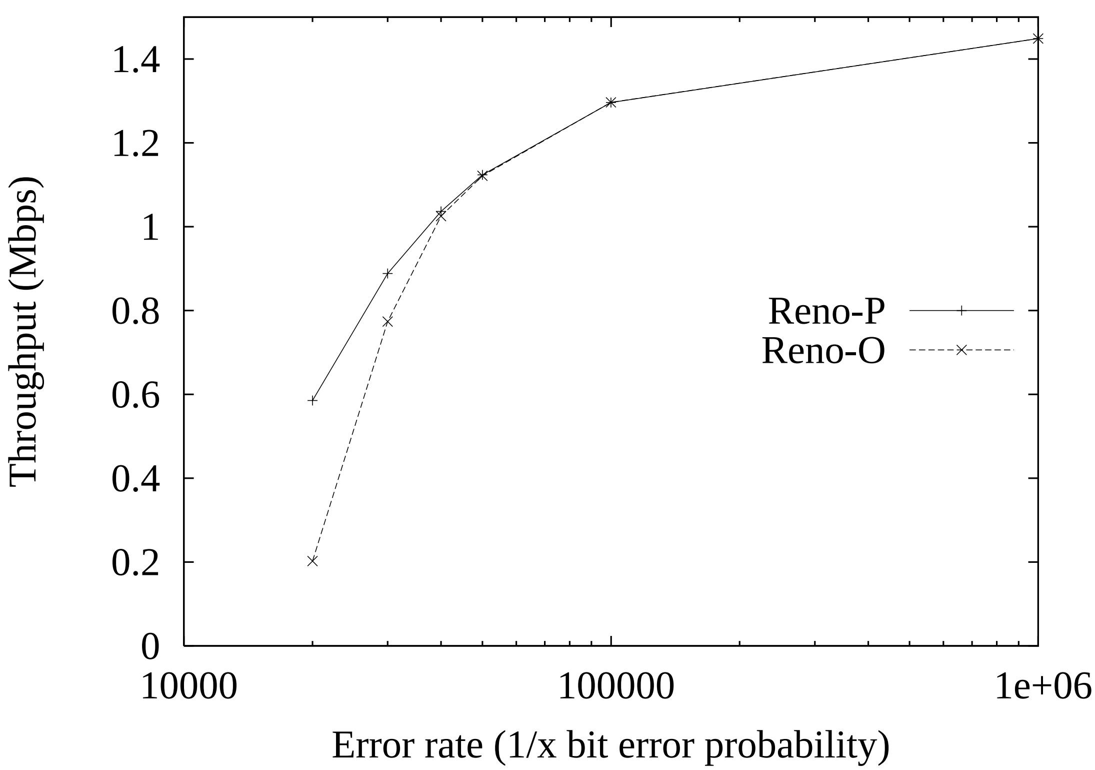

Figure 7.26: Average of total throughput (delay 100ms) when the error rate of two terminals is same.

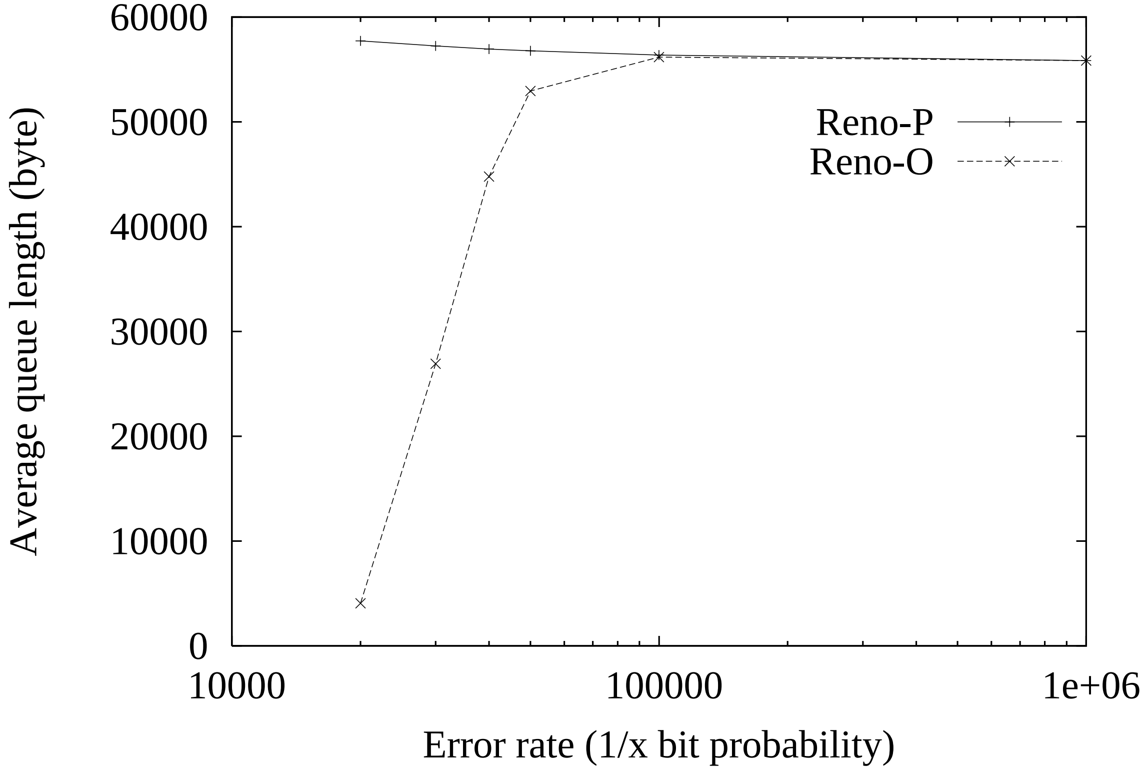

Figure 7.27: Average of total queue length (delay 1ms) when the error rate of two terminals is same.

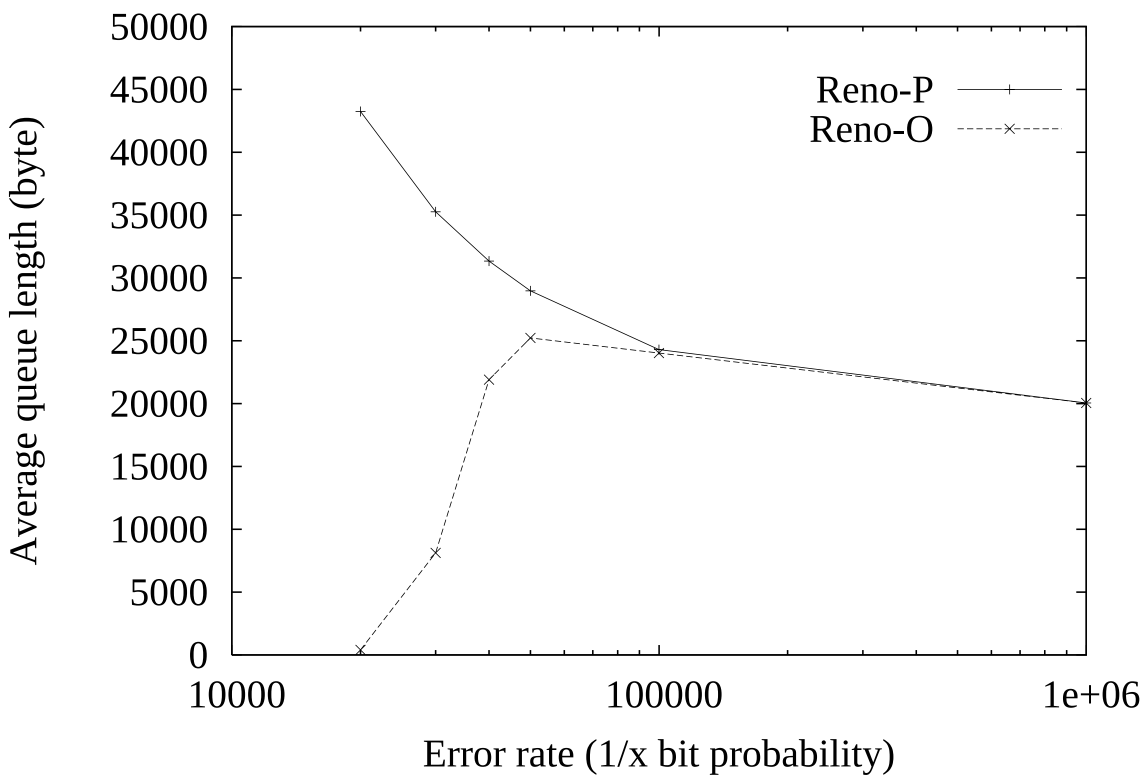

Figure 7.28: Average of total queue length (delay 100ms) when the error rate of two terminals is same.

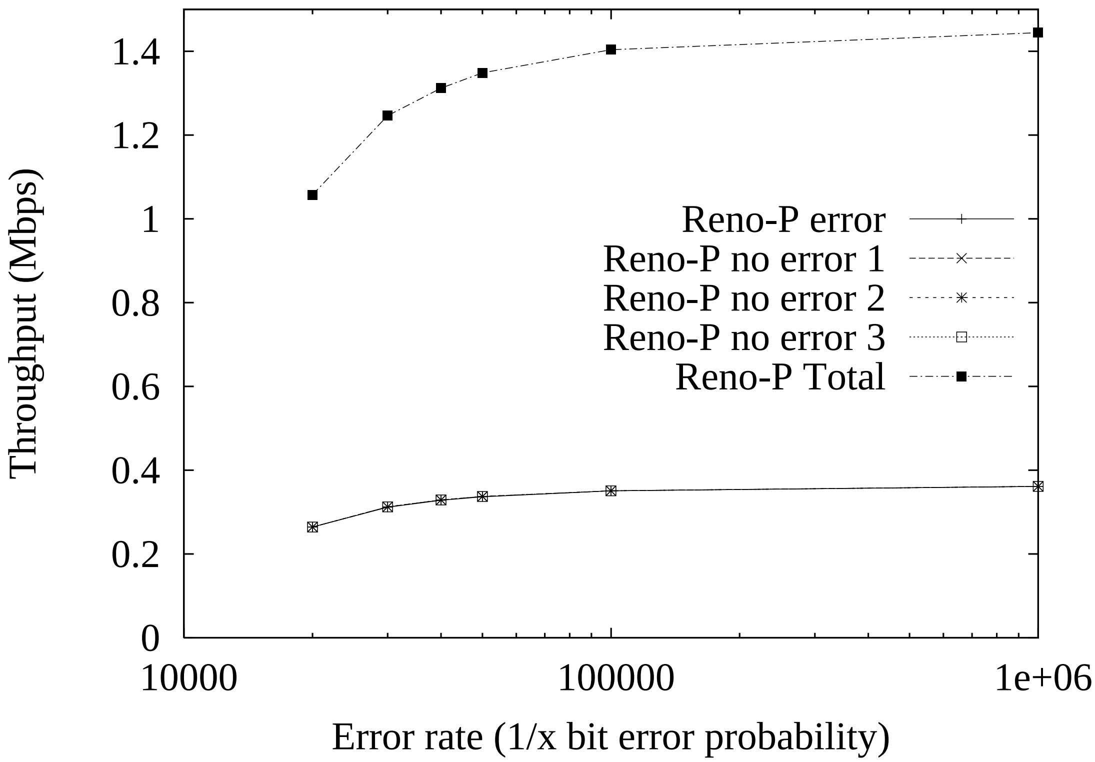

Figure 7.29: Average throughput of the proposed method (delay 1ms) for four terminals (one error) model.

7.3.4 Four terminals model

Figs. 7.29–7.34 depict the average throughput and average queue length of the proposed method and original method for the delay of 1ms and 100ms with four terminals, respectively, when the error bit rate of three terminals is fixed to 1/1000000, but that of the other one terminal is changed from 1/20000 to 1/1000000. As illustrated in Section 7.3.3, only TCP Reno is used.

When the error bit rate is less than or equal to 1/100000, the results of the both methods have no difference and the bandwidth in the wireless link is equally shared to each terminal in every condition. In the high error rate condition, the bandwidth is also fairly shared among the four terminals as in Figs. 7.29 and 7.30 for the proposed method, where the average of total throughput is decreased 30% from the maximum because of the

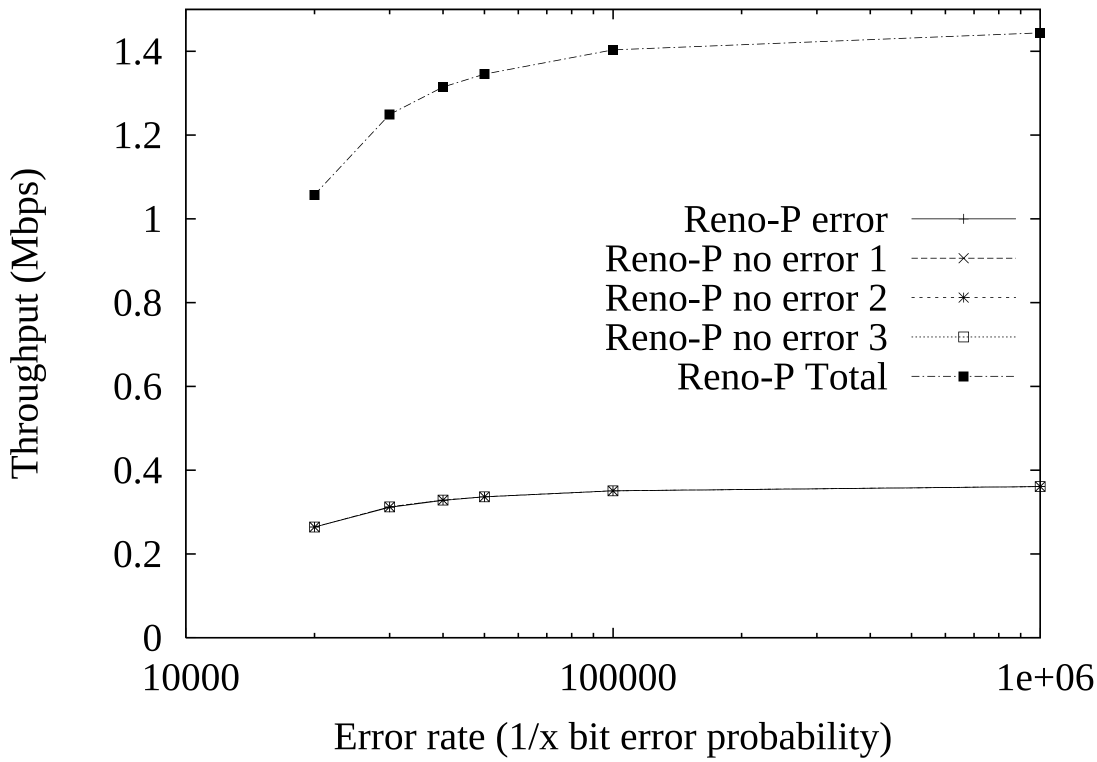

Figure 7.30: Average throughput of the proposed method (delay 100ms) for four terminals (one error) model.

retransmission control at the MAC layer. In the original method, however, the average throughput for the high-error-bit terminal is extremely reduced, since the rest of the bandwidth is shared by the other three terminals as in Figs. 7.31 and 7.32. The large difference is not observed between the delay of 1ms and 100ms. Fig. 7.33 and Fig. 7.34 show the average total queue length. In the proposed method, since only one terminal is in the high error conditions, the queue length does not increase against the high error bit rate conditions comparing to the low error, as in Figs. 7.23 and 7.24. In the original method, the average one is shorten in the high error bit rate conditions, because the terminal in the high error conditions hardly communicates, and then the queue is only used by the other three terminals.

In the next, the error bit rate of one terminals is fixed to 1/1000000, but the other three terminals are changed from 1/20000 to 1/1000000. The results are depicted in

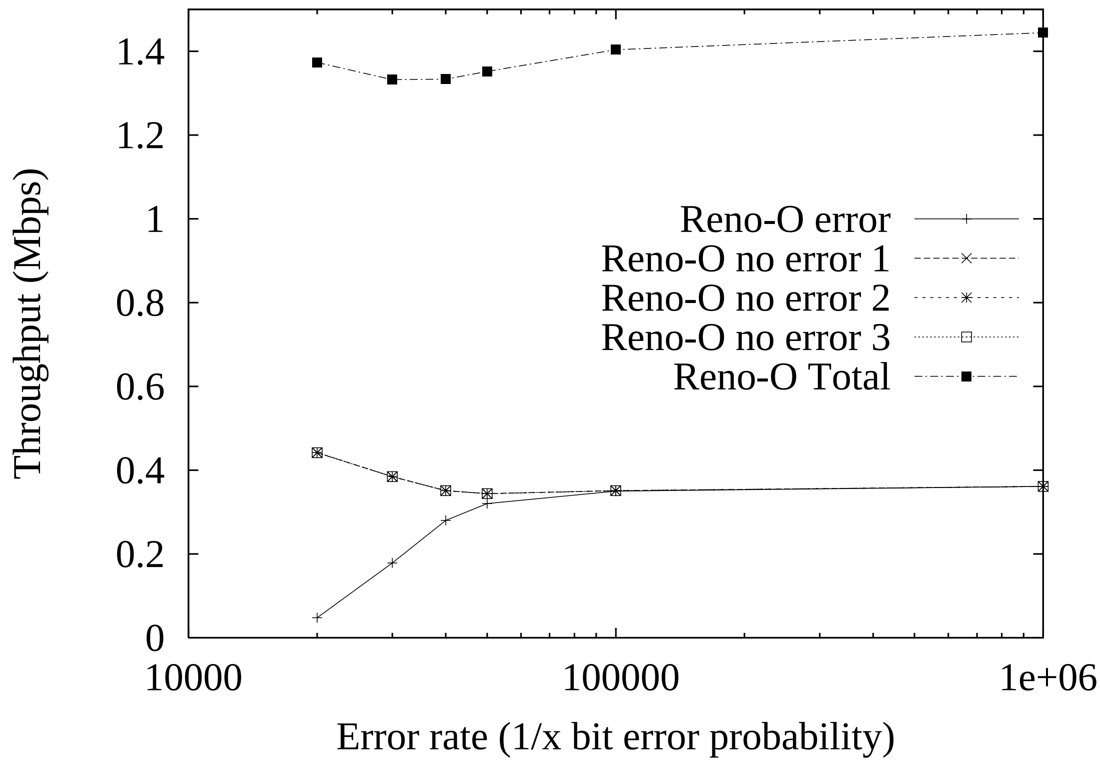

Figure 7.31: Average throughput of the original method (delay 1ms) for four terminals (one error) model.

Figs. 7.35–7.40. All the results in these figures tend to be similar to that in Figs. 7.29–7.34. The important difference between these results are that the average throughput of the original method in Figs. 7.37 and 7.38 are almost same in Figs. 7.31 and 7.32 for the one error terminal model, because the three terminals in the high error condition do not almost communicate and the packets in the queue are only used for the other one terminal.

Finally, the error bit rate of all terminals are changed from 1/20000 to 1/1000000. The results are shown in Figs. 7.41–7.44. From Fig. 7.41, the both methods have the almost same results. However, the status of each queue is different between the two methods. As shown in Fig. 7.42, the average of the total queue length keeps to be long even in the high error condition in the proposed method, then all terminals share the available bandwidth. In the original method, although the average value is reduced, the aggregation of the four

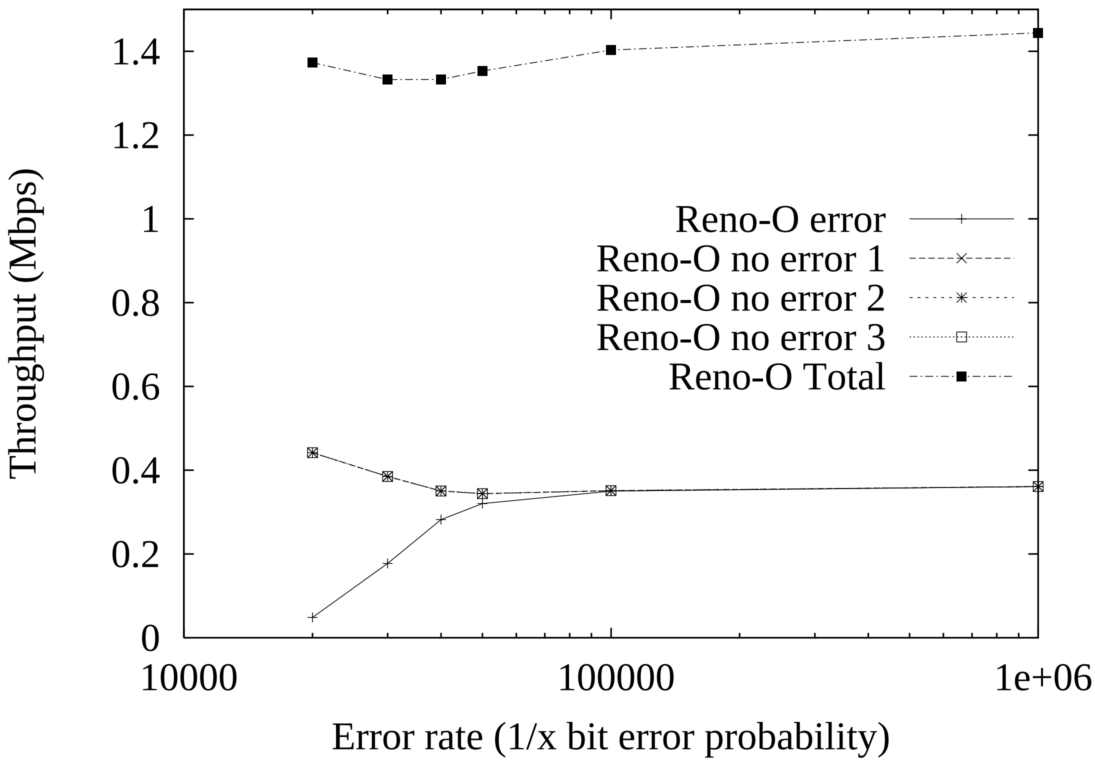

Figure 7.32: Average throughput of the original method (delay 100ms) for four terminals (one error) model.

flows only can achieve to the limitation from the shorter transmission delay. In fact, the original method decreases the average throughput in Fig. 7.43 in spite that the average queue length is also reduced same as in Fig. 7.42.

7.4 Conclusions for Chapter 7

This chapter proposes an improved queue management method for TCP communications in wireless environments. Since the proposed method is designed to help local retransmission in the MAC layer, the packet loss is completely eliminated from a wireless link and all packets are delivered in the order. Moreover, the proposed method is simple to implement to the conventional system. From the simulation experiments, the throughput of TCP is improved especially in the high error bit rate conditions, and the available

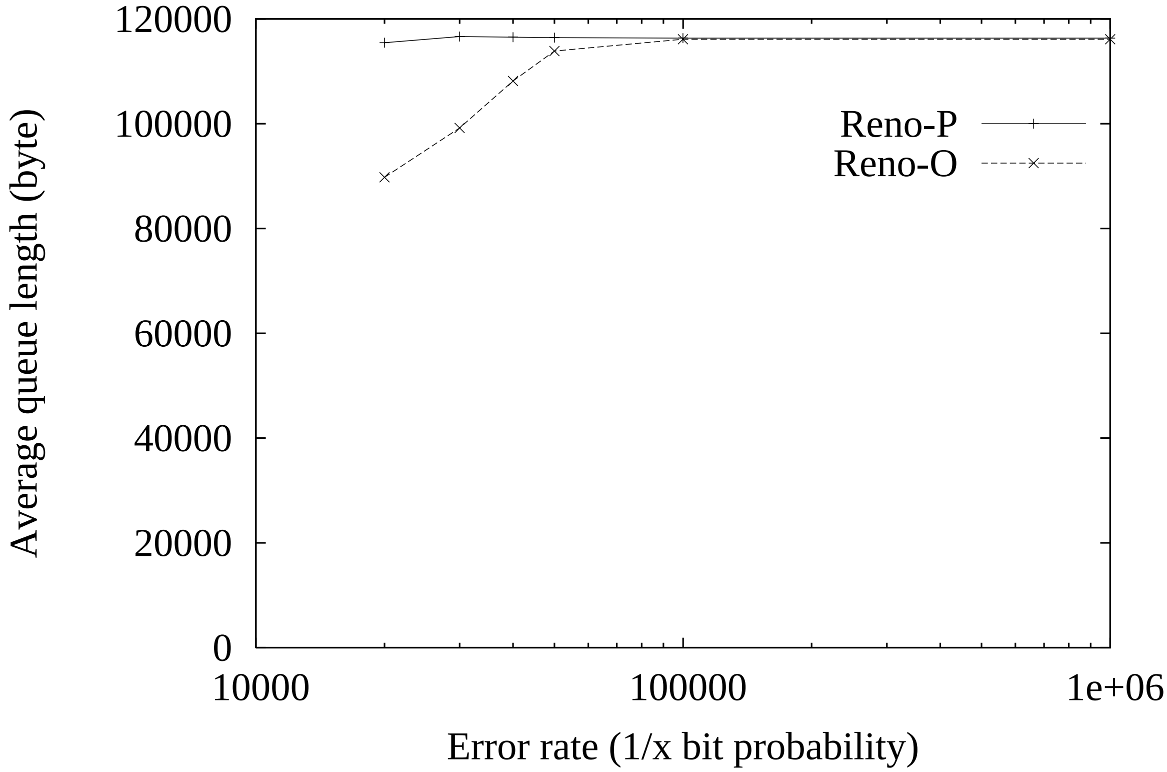

Figure 7.33: Average of total queue length (delay 1ms) for four terminals (one error) model.

bandwidth of the wireless link is tried to maximumly consume. The influences from the limited queue length are also observed. In the simulation, the proposed method is only implemented in BS, but implementing also to terminals would be more effective when the terminals send data through BS.

The proposed method is effective for wireless LAN environments, in the cases that the MAC layer uses "go back 1" algorithm for the reliable communications. When a wireless link delay is large such as wireless WAN environments, the proposed method should adopt "go back N" and so on to the MAC layer control for using the bandwidth effectively.

In high bit error rate condition, the proposed method require many times of the MAC retransmission. At TCP level, the bandwidth of a wireless link is fairly shared among terminals, but not at MAC level. The MAC level fairness should also be considered. Farther sophisticated control for multimedia traffic should be considered to the proposed method.

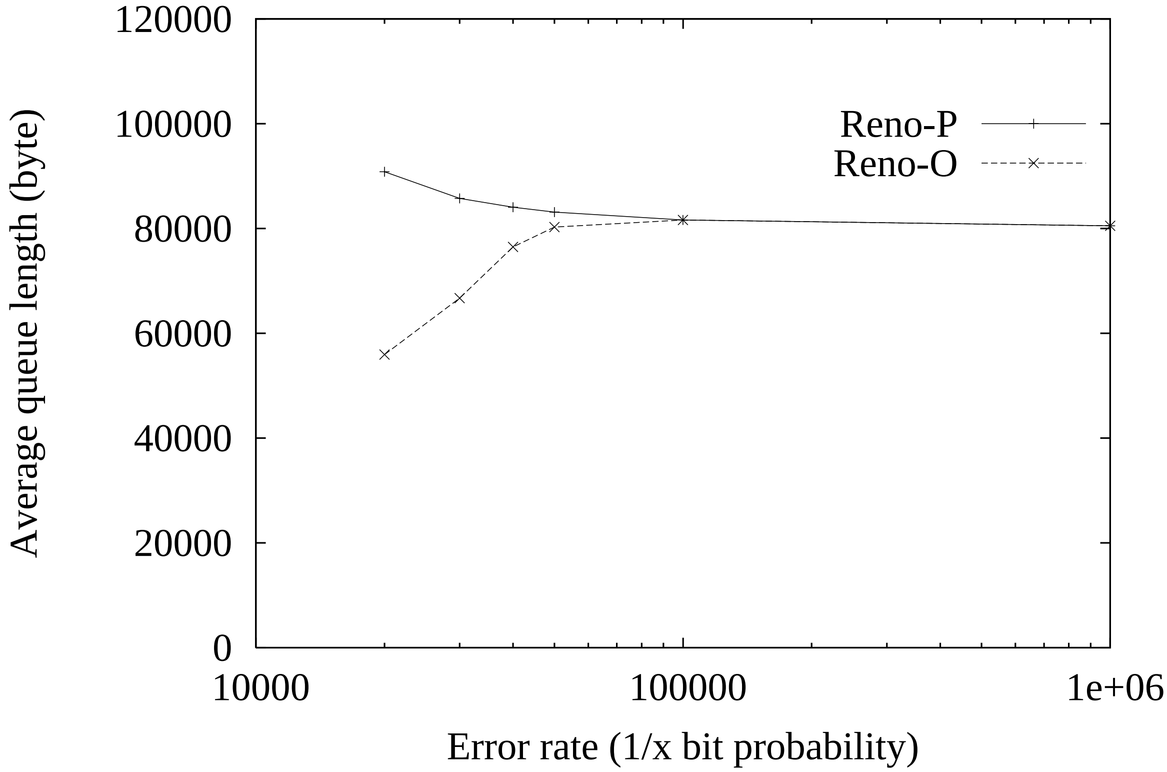

Figure 7.34: Average of total queue length (delay 100ms) for four terminals (one error) model.

For example, the addition of Random Early Discard, Explicit Congestion Notification or Class Based Queuing will improve the performance.

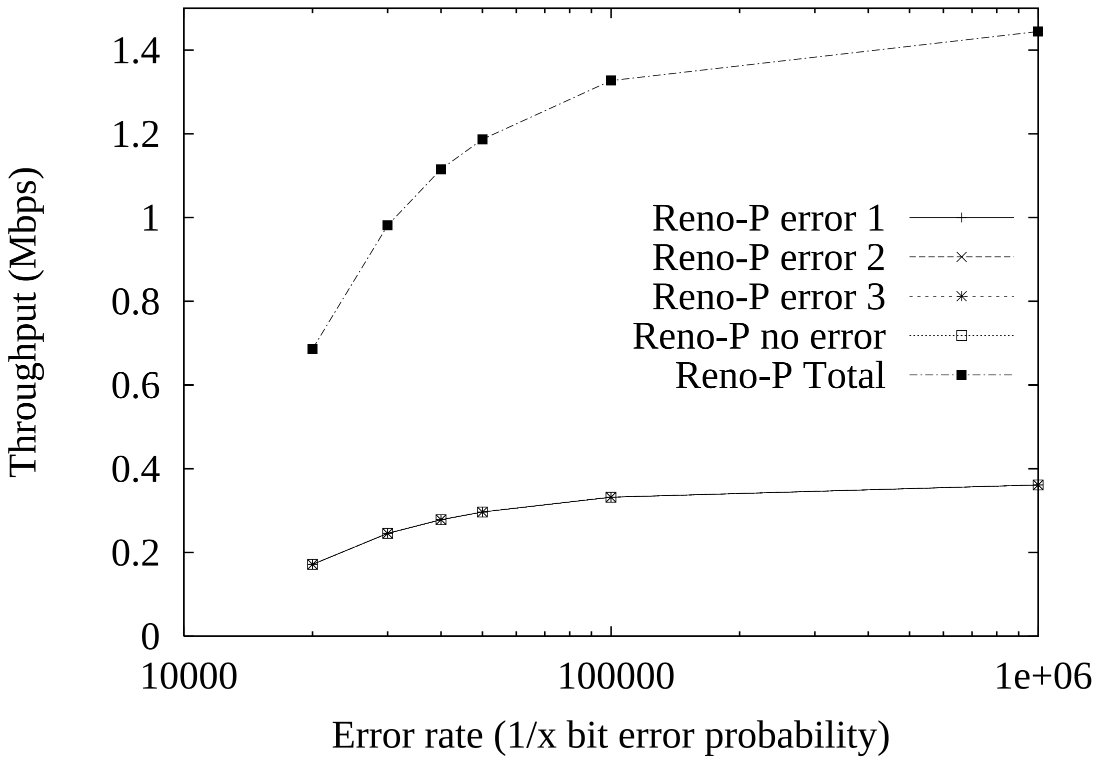

Figure 7.35: Average throughput of the proposed method (delay 1ms) for four terminals (three errors) model.

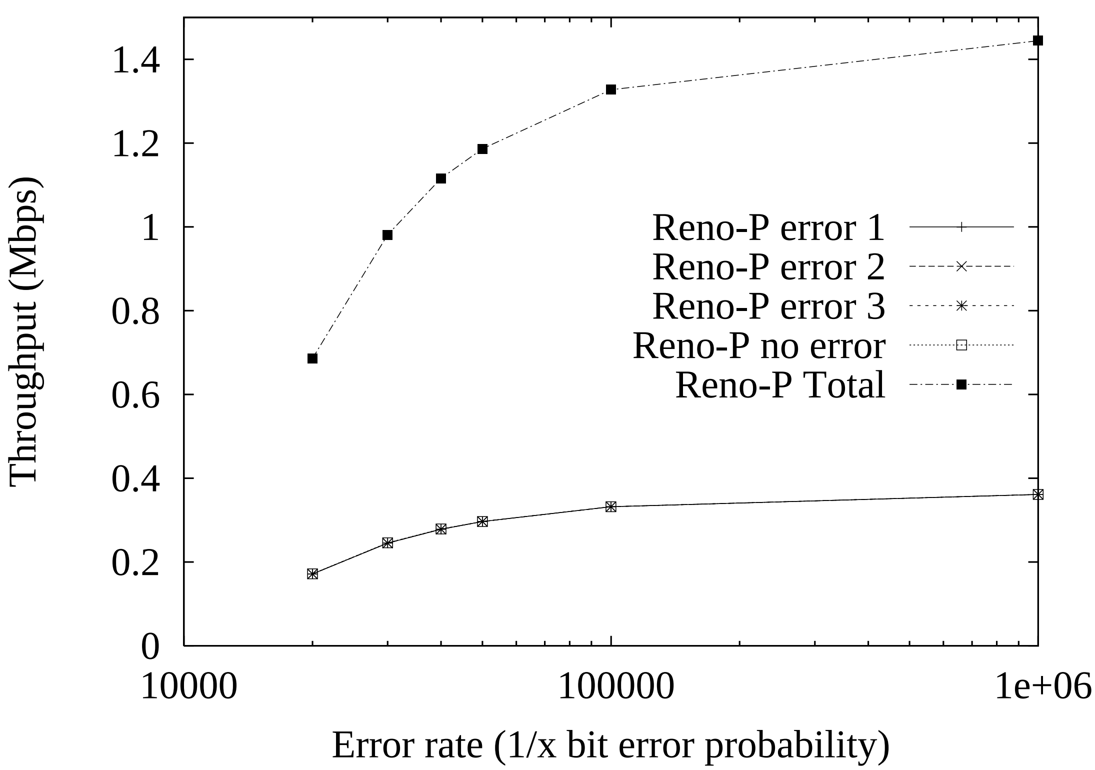

Figure 7.36: Average throughput of the proposed method (delay 100ms) for four terminals (three errors) model.

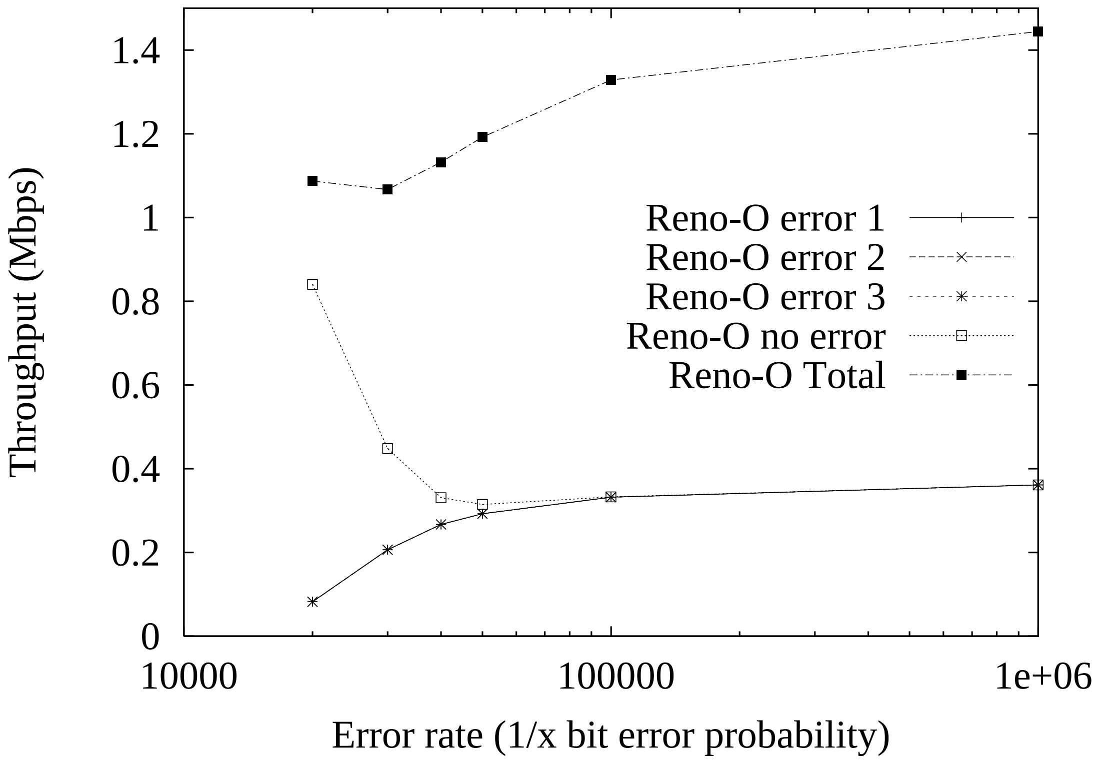

Figure 7.37: Average throughput of the original method (delay 1ms) for four terminals (three errors) model.

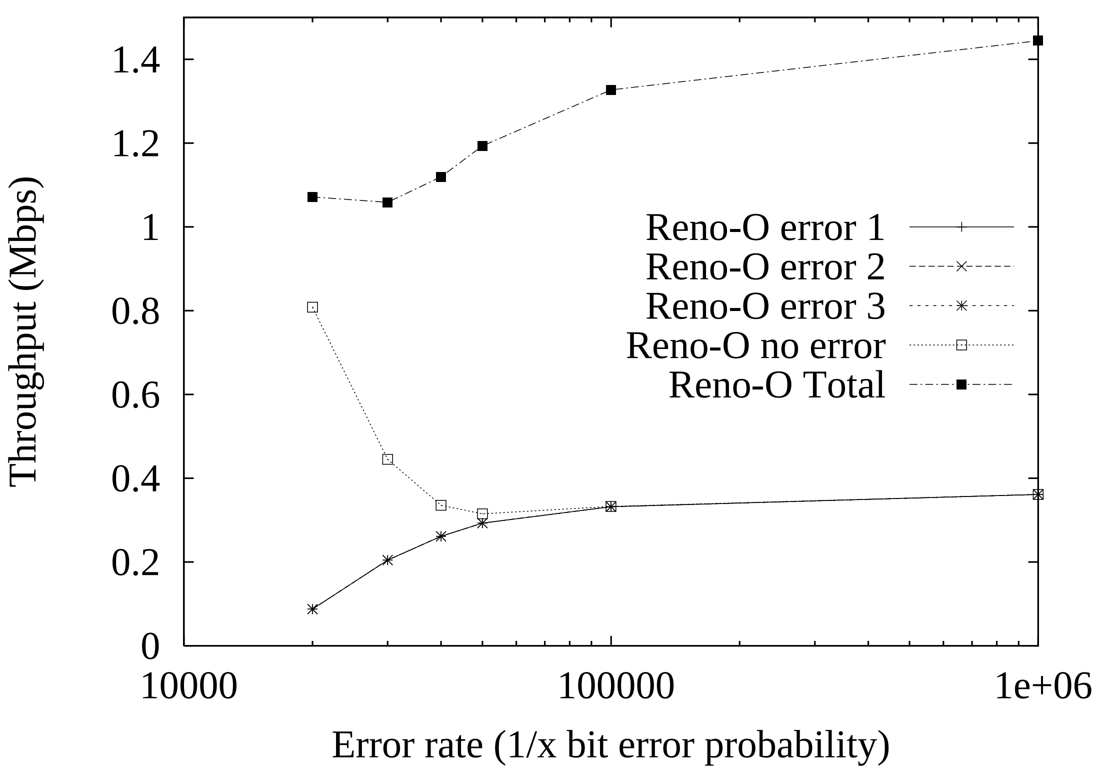

Figure 7.38: Average throughput of the original method (delay 100ms) for four terminals (three errors) model.

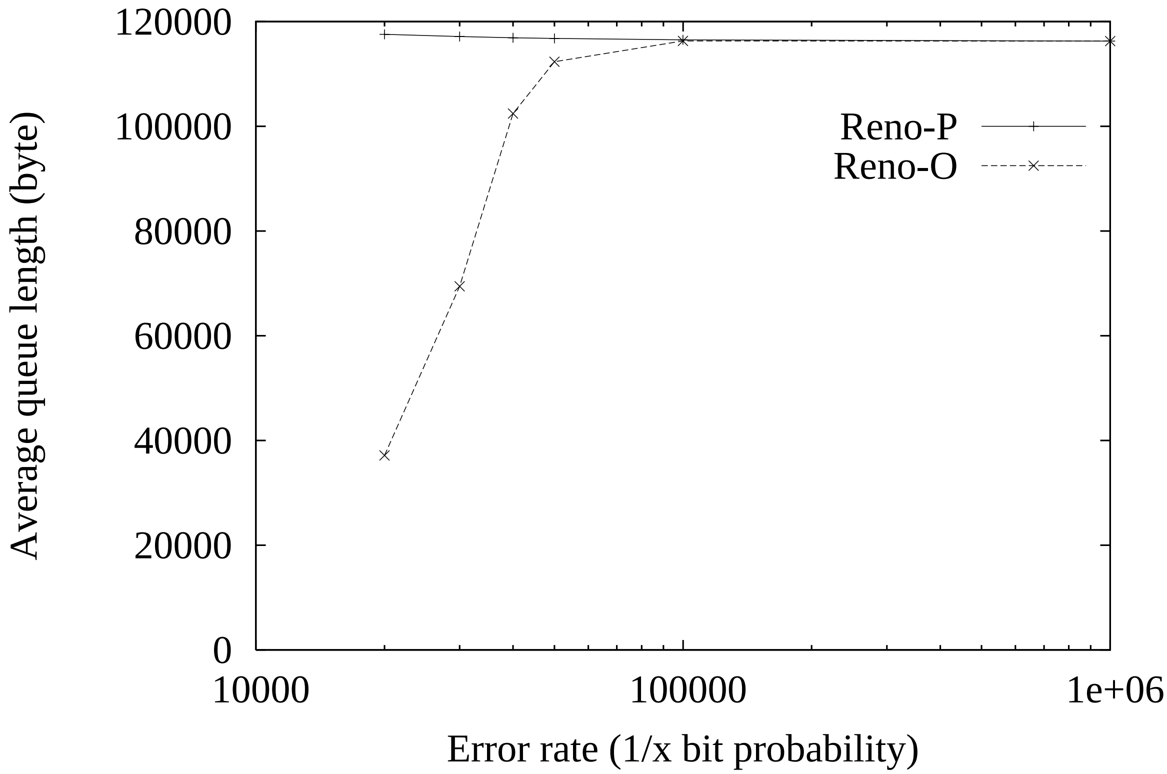

Figure 7.39: Average of total queue length (delay 1ms) for four terminals (three errors) model.

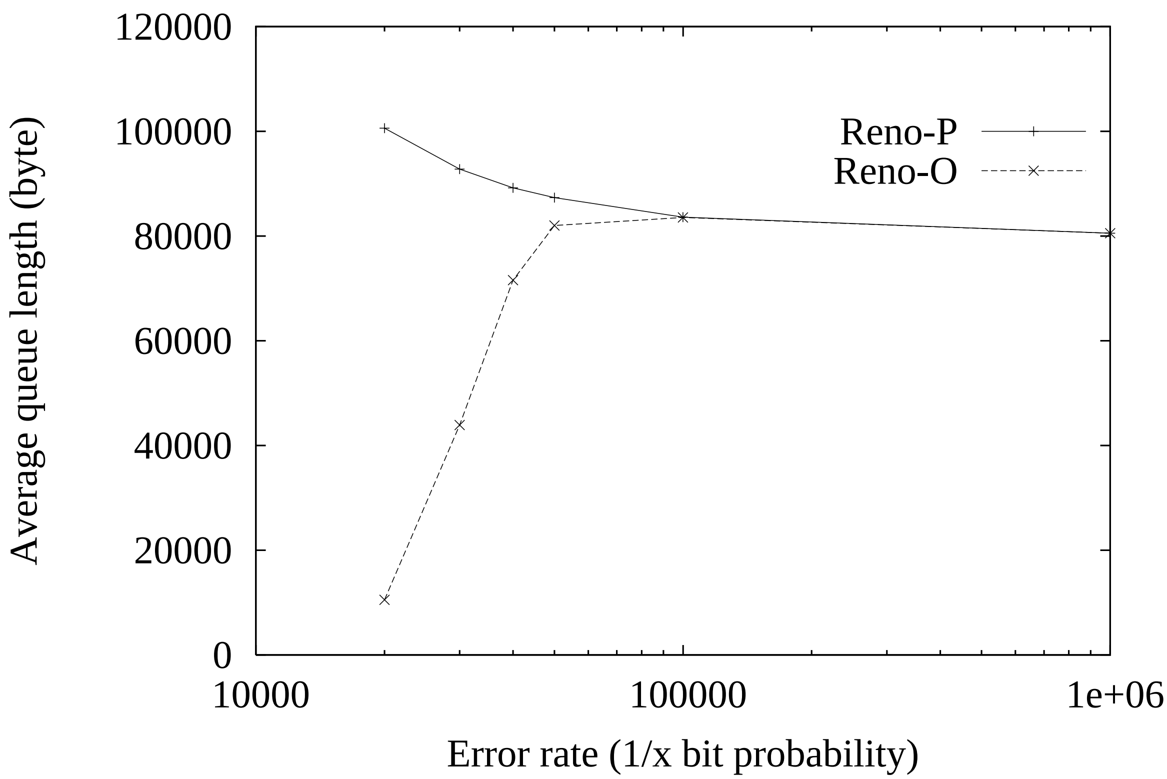

Figure 7.40: Average of total queue length (delay 100ms) for four terminals (three errors) model.

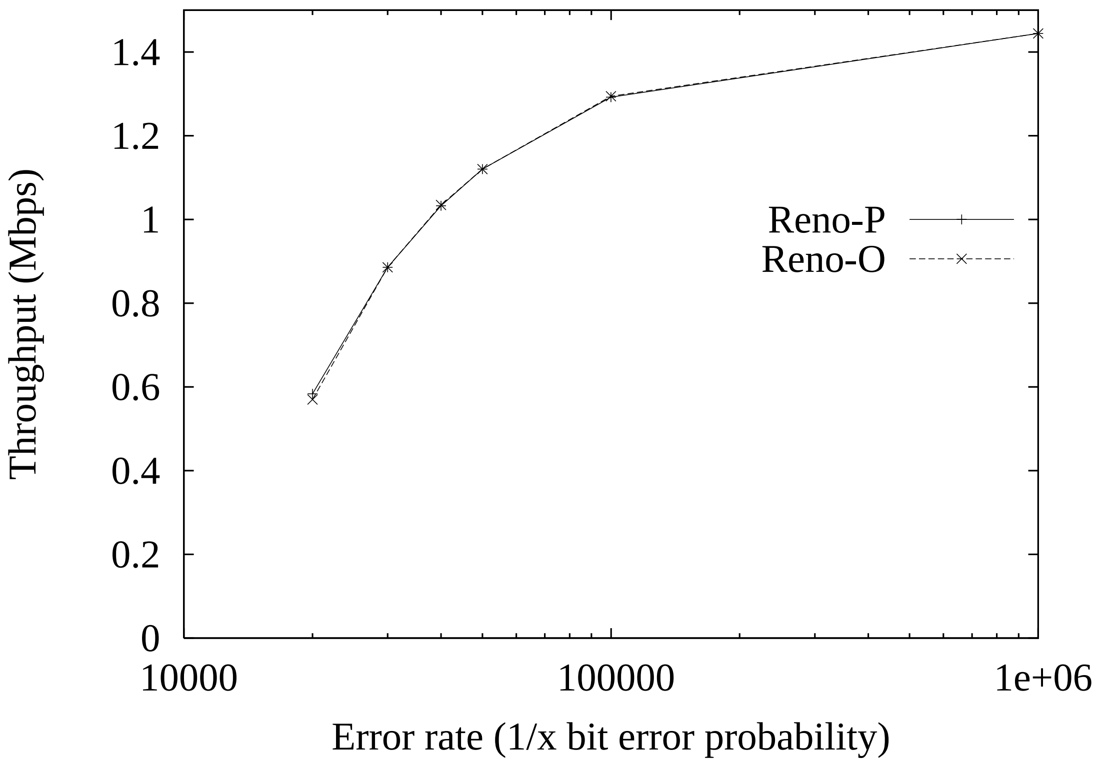

Figure 7.41: Average throughput of the proposed method (delay 1ms) for four terminals (four errors) model.

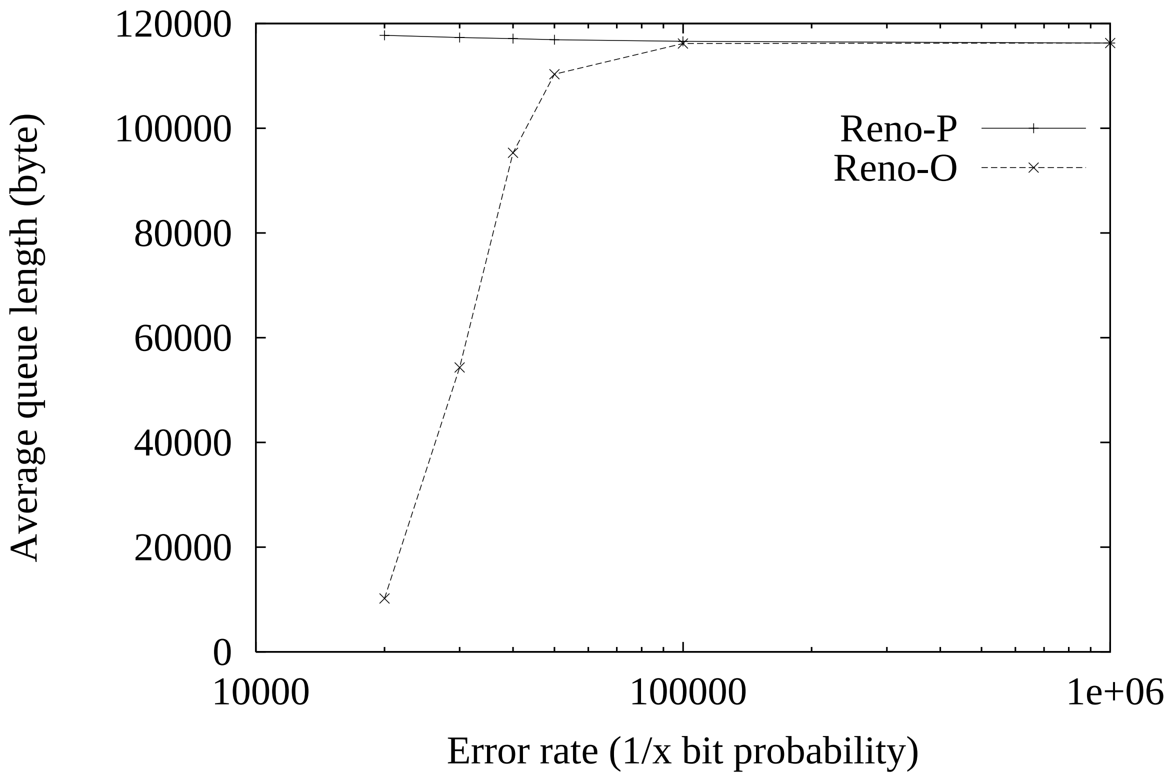

Figure 7.42: Average of total queue length (delay 1ms) for four terminals (four errors) model.

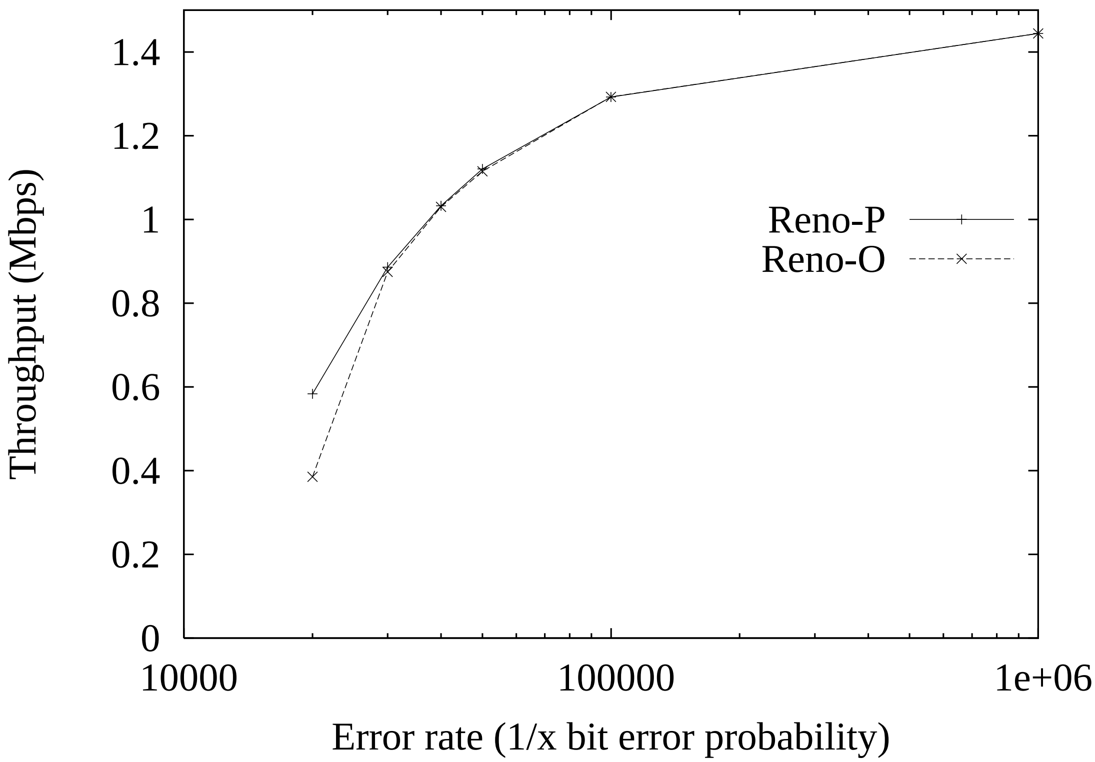

Figure 7.43: Average throughput of the proposed method (delay 100ms) for four terminals (four errors) model.

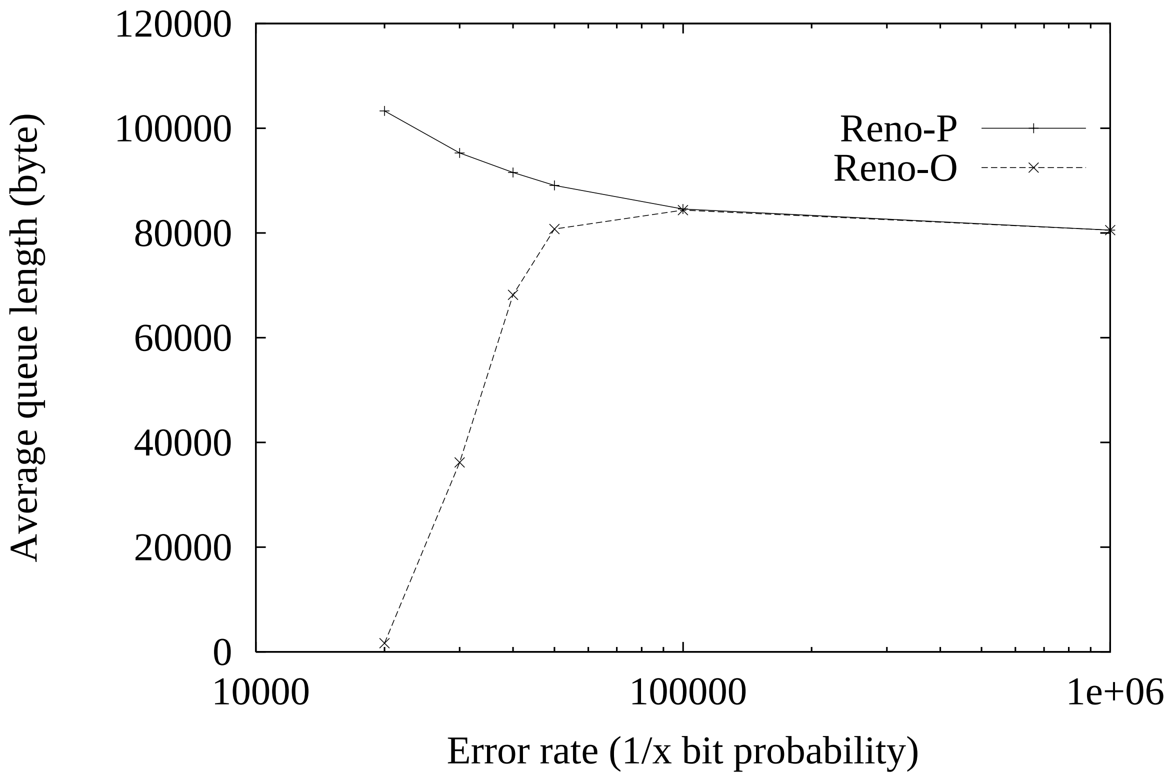

Figure 7.44: Average of total queue length (delay 100ms) for four terminals (four errors) model.

Chapter 8

Conclusions

This thesis proposes control methods for improving communication quality in mobile Internet environments. The seamless handoff method in Chapter 3 serves on disrupted communications in Mobile IP with simple mobility prediction of the MN. The fast authentication method in Chapter 4 serves secure and seamless handoff by extending the seamless handoff method. The adaptive resource reservation and admission control in Chapter 5 are proposed by using the simple mobility prediction of the seamless handoff method. The adaptive handoff algorithm in Chapter 6 is proposed for efficient network resource utilization. Since these methods are prepared for solution of micro mobility problems, the communication quality is improved. The proposed queue management in Chapter 7 also serves high throughput of TCP in wireless LAN environments.

Following topics should be studied in future works.

- This thesis only discusses within Mobile IP version 4. The proposed method should be adapted to the version 6.

- For all methods, simulation models and conditions are still simple, more complex and realistic simulation evaluations are required.

- In Chapter 5, the QoS signaling protocol is for between MH and BS which is user-network signaling but the network-network signaling one is omitted. The latter protocol would also improve the performance by using the information to dynamically change the "static part" of the admission control.

- By collaborating "adaptive resource reservation method" in Chapter 5 and "adaptive handoff method" in Chapter 6, the network resource would be further effectively used.

- Chapter 7 discusses only TCP performance in wireless environments but mobile environments are should be evaluated and improved for the method.

- The proposed methods in the thesis are for one hop mobile or wireless environments, the multi-hop environments should be considered.

Acknowledgments

The author would like to express many thanks to all those who have provided kind and heartful support for the research.

First of all, the author must express special thanks to Professor Yoshihiko Ebihara, and Associate Professor Shigetomo Kimura, of the Institute of Information Sciences and Electronics at the University of Tsukuba, for their sincere supervisions, hospitality and encouragement.

The author also expresses sincere thanks to Professor Hisao Kameda, Professor Nobuo Ohbo, Professor Seiichi Nishihara, and Professor Hiroaki Nishikawa, of the Institute of Information Sciences and Electronics at the University of Tsukuba, for their valuable comments and discussions.

Finally, the author thanks members of Computer Networks Laboratory, at University of Tsukuba, for their kind consideration and hospitality.

Bibliography

[1] J. Postel, "Internet Protocol," *RFC791*, 1981.

[2] C. Perkins, "IP Mobility Support," *RFC2002, 1996.*

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weisss, "An architecture for differentiated service," *RFC 2475, 1998.*

[4] M. Allman, V. Paxson and W. Stevens, "TCP Congestion Control," *RFC2581, 1999.*

[5] S. Floyd and T. Henderson, "The New Reno Modification to TCP's Fast Recovery Algorithm," *RFC2582, 1999.*

[6] S. McCanne and S. Floyd, "Network Simulator ns (version2)," *http://www.isi.edu/nsnam/ns/.*

[7] S. Seshan, H. Balakrishnan and R. H. Katz, "Handoffs Cellular Wireless Networks : The Daekalus Implementation and Experience," *Wireless Personal Communications, Vol.4, No.2, pp.141–162,* 1997.

[8] M. Stemm and R. H. Katz, "Vertical Handoffs in Wireless Overlay Networks," *ACM Mobile Networking and Applications, Vol. 3, No.4, pp.335–350,* 1998.

[9] C. Perkins and K. Wang, "Optimized Smooth Handoffs in Mobile IP," *ICSS'99, pp.340–346,* 1999.

[10] R. Caceres and V. N. Padamanabhan, "Fast and Scalable Wireless Handoffs in Support of Mobile Internet Audio," *ACM Mobile Networking and Applications, Vol. 3, No. 4, pp. 350–363,* 1998.

[11] D. B. Johnson and C. Perkins, "Mobility Support in IPv6," *draft-ietf-mobileip-ipv6-11.txt*, 2000.

[12] P. R. Calhoun, H. Akuhtar, E. Qaddoura and N. Asokan, "Minimal Latency Secure Hand-off," *draft-calhoun-mobileip-min-lat-handoff-01.txt, 2000.*

[13] H. Krawczyk, "SKEME: A Versatile Secure Key Exchange Mechanism for Internet," *Proc. Internet Soc. Symp. Network and Dist. Sys. Security, 1996.*

[14] A. Aziz and M. Patterson, "Design and Implementation of SKIP," *http://skip.incog.com/inet-95.ps, 1995.*

[15] J. Zao, et al., "A Public-Key Based Secure Mobile IP," *MOBICOM '97, pp.173–184 1997.*

[16] V. Gupta and G. Montenegro, "Secure and mobile networking," *Mobile Networks and Applications 3, pp.381–390, 1998.*

[17] C. Park, "On Certificate-Based Security Protocols for Wireless Mobile Communication Systems," *IEEE Network Sep., pp.50–55, 1997.*

[18] R. Rivest, "The MD5 Message-Digest Algorithm," *RFC 1321, 1992.*

[19] T. Camp, J. Lusth and J. Matocha, "Reduced Cell Switching in a Mobile Computing Environment," *MOBICOM '00, pp.143–153, 2000.*

[20] J. Hoon, et al., “An Adaptive Resource Allocation Mechanism Including Fast and Reliable Handoff in IP-Based 3G Wireless Networks,” *IEEE Personal Communications, Dec., pp.42–47, 2000.*

[21] D. Hong and S. S. Rappaport, “Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures,” *IEEE Trans. Veh. Technol.*, Vol. 35, pp. 77–92, 1986.

[22] M. Naghshineh and M. Schwartz, “Distributed Call Admission Control in Mobile/Wireless Networks,” *IEEE J. Select. Areas Commun.*, Vol. 14, pp. 711–717, 1996.

[23] C. Oliveira, J. B. Kim and T. Suda, “An Adaptive Bandwidth Reservation Scheme for High-Speed Multimedia Wireless Networks,” *IEEE J. Select. Areas Commun.*, Vol. 16, No. 6, pp. 858–874, 1998.

[24] D. A. Levine, I. F. Akyildiz and M. Naghshineh, “A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept,” *IEEE/ACM Trans. on Networking*, Vol. 5, No. 1, pp. 1–12, 1997.

[25] M. Chiu and M. A. Bassionuni, “Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning,” *IEEE J. Select. Areas Commun.*, Vol. 18 No. 3, pp. 510–522, 2000.

[26] A. Alijadhai and T. F. Znati, “Predictive Mobility Support for QoS Provisioning in Mobile Wireless Environments,” *IEEE J. Select. Areas Commun.*, Vol. 19, No. 10, pp. 1915–1930, 2001.

[27] K. D. Wong and D. C. Cox, “A Pattern Recognition System for Handoff,” *IEEE J. Select. Areas Commun.*, Vol. 18 No. 7 pp. 1302–1311, 2000.

[28] R. Braden, L. Zhang, S. Berson, S.Herzon, and S. Jamin, “Resource ReSerVation Protocol (RSVP) Version 1, Functional Specification,” *RFC 2210*, 1997.

[29] I. Mahadevan and K. M. Sivalingam, “Architecture and Experimental Results for Quality of Service in Mobile Networks Using RSVP and CBQ,” *Wireless Networks*, No. 6, pp. 221–233, 2000.

[30] I. Mahadevan and K. M. Sivalingam, “Architecture and Experimental Framework for Supporting QoS in Wireless Networks Using Differentiated Service,” *Mobile Networks and Applications*, No. 6, pp. 385–395, 2001.

[31] H. G. Ebersman and O. K. Tonguz, “Handoff Ordering Using Signal Prediction Priority Queuing in Personal Communication Systems,” *IEEE Trans. Veh. Technol.*, Vol. 48, No. 1, pp. 20–35, 1999.

[32] R. Vijayan and J. M. Jack, “A Model for Analyzing Handoff Algorithms,” *IEEE Trans. Veh. Technol.*, Vol. 43, pp. 351–356, 1993.

[33] M. Mathis, J. Mahdavi, S. Floyd and A. Romanow, “Selective acknowledgment opinions,” *RFC2018*, 1996.

[34] A. Chrungoo, V. Gupta, H. Saran and R. Shorey, “TCP k-SACK: A Simple Protocol to Improve Performance over Lossy Links,” *Proc. IEEE Globecom’02*, pp. 1713–1717, 2001.

[35] T. Goff, J. Moronski, D. S. Phatak and V. Gupta, "Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments," *Proc. IEEE Infocom'00*, pp. 1537–1545, 2000.

[36] H. Balakrishnan S. Seshan, and R. H. Katz, "Improving reliable transport and hand-off protocols in cellular wireless networks," *ACM Wireless Networks*, Vol. 1, Issue 4, pp. 469–481, 1996.

[37] H. Balakrishnan and R. H. Katz, "Explicit loss notification and wireless web performance," *IEEE Globecom'98*, http://nms.lcs.mit.edu/papers/globecom98/, 1998.

[38] W. Ding and A. Jamalipour, "Delay performance of the new explicit loss notification TCP technique for wireless networks," *IEEE Globecom'01*, pp. 3483–3487, 2001.

[39] M. Shreedhar and G. Varghese, "Efficient Fair Queuing Deficit Round Robin," *ACM Sigcomm'95*, pp. 231–242, 1995.

List of Papers

Journal

(1) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“A Proposal of Seamless Handoff Method for Cellular Internet Environments”

IEICE Transactions on Communications, Vol. E84-B, No. 4, pp752-759, 2001.

(2) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“A Fast Authentication Method for Secure and Seamless Handoff”

Lecture Note in Computer Science 2344 Information Networking part 2, pp. 243–252, 2002.

International Conference

(3) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“Soft Handoff Method for Wireless Cellular Real-Time Communications in Mobile IP”

14th International Conference on Information Networking, pp. 4C-2.1–4C-2.6, 2000.

(4) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“Adaptive Handoff Algorithm for Efficient Network Resource Utilization”

The International Conference on Advanced Information Networking and Applications (AINA 2003), Accepted.

Technical Note

(5) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“A Proposal of Communication Route Switching Method Preventing Handoff Latency for Mobile IP”

MBL 8, 1999 (in Japanese).

(6) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“A Proposal of Dynamic Bandwidth Reservation and Admission Control Based on Information of Micro Mobility”

MoMuC, March, 2003 (to appear).

(7) Satoshi Ohzahata, Shigetomo Kimura and Yoshihiko Ebihara

“A Proposal of Queue Management Method and Retransmission Mechanism for Wireless TCP Communications” MBL 24, 2003 (to appear)