Adaptive Route Optimization in Hierarchical Mobile IPv6 Networks

Sangheon Pack, *Member*, *IEEE*, Xuemin (Sherman) Shen, *Senior Member*, *IEEE*, Jon W. Mark, *Life Fellow*, *IEEE*, and Jianping Pan, *Member*, *IEEE*

Abstract—By introducing a mobility anchor point (MAP), Hierarchical Mobile IPv6 (HMIPv6) reduces the signaling overhead and handoff latency associated with Mobile IPv6. However, if a mobile node (MN)'s session activity is high and its mobility is relatively low, HMIPv6 may degrade end-to-end data throughput due to the additional packet tunneling at the MAP. In this paper, we propose an adaptive route optimization (ARO) scheme to improve the throughput performance in HMIPv6 networks. Depending on the measured session-to-mobility ratio (SMR), ARO chooses one of the two different route optimization algorithms adaptively. Specifically, an MN informs a correspondent node (CN) of its on-link care-of address (LCoA) if the CN's SMR is greater than a predefined threshold. If the SMR is equal to or lower than the threshold, the CN is informed with the MN's regional CoA (RCoA). We analyze the performance of ARO in terms of balancing the signaling overhead reduction and the data throughput improvement. We also derive the optimal SMR threshold explicitly to achieve such a balance. Analytical and simulation results demonstrate that ARO is a viable scheme for deployment in HMIPv6 networks.

Index Terms—Hierarchical Mobile IPv6, all-IP networks, adaptive route optimization, performance analysis.

1 Introduction

In wireless/mobile networks, mobile nodes (MNs) can change their attachment points freely while being connected. Therefore, mobility management is essential for tracking the MNs' current locations so that their data can be delivered correctly. Since the next-generation wireless/mobile networks are anticipated to be unified networks based on IP technology, i.e., all-IP networks, IP-based mobility management is critical. Many IP mobility protocols have been proposed in the literature [1]. Among them, Mobile IPv6 (MIPv6) [2] from the Internet Engineering Task Force (IETF) is the de facto protocol for mobility management in IPv6 wireless/mobile networks. However, MIPv6 incurs a high signaling overhead when handoff is too frequent. To overcome this drawback, Hierarchical Mobile IPv6 (HMIPv6), which employs a mobility anchor point (MAP) to handle binding update (BU) for MNs within the MAP domain, has recently been introduced by the IETF [3]. In this way, network-wide signaling is only required when the MN roams outside of its current MAP domain and, thus, signaling traffic and handoff latency can be reduced. Hierarchical mobility management is also widely adopted in other mobility management schemes such as Cellular IP [4], HAWAII [5], IDMP [6], etc.

MIPv6 supports route optimization (RO) for efficient packet delivery. Packets sent by a correspondent node (CN) are first routed to the home agent (HA) of the MN, then the HA forwards the packets to the MN's registered temporary address, i.e., care-of address (CoA). Once the MN receives the packets tunneled from the HA, the MN sends a BU message to the CN. If the MN moves to a new subnet and its CoA is changed, the MN advertises its new CoA by sending BU messages to all CNs listed in its binding update list [2]. The binding update list is maintained by each MN and consists of entries for the CNs having active sessions with the MN. After receiving the BU message, the CN updates its binding cache and sends a binding acknowledgment (BACK) message to the MN. When the MN receives the BACK message, the MN updates its binding update list. After the BU procedures are finished, the CN sends subsequent packets directly to the MN, bypassing the HA. Since HMIPv6 is based on Mobile IPv6, the HMIPv6 specification [3] also defines route optimization procedures. However, in HMIPv6, the MN's regional CoA (RCoA), i.e., an address in the MAP subnet, rather than the MN's on-link CoA (LCoA), is used for route optimization.

RO enables direct packet transmission (bypassing the HA) between the CN and MN. Hence, MIPv6 with RO normally achieves a higher throughput than MIPv6 without RO. However, RO does not always guarantee a better performance for MNs with different session activities and mobility patterns [7]. For instance, if an MN hands off frequently while it has low session activity, the throughput improvement due to RO becomes negligible and RO results in a large amount of signaling traffic due to the frequent execution of the BU/BACK procedures. Similar problems can also be observed in HMIPv6 networks. Notifying the CN of the MN's RCoA is efficient in reducing the binding

- *S. Pack is with the School of Electrical Engineering, Korea University, Anam-dong, Seongbuk-Gu, Seoul, 136-701, Korea. E-mail: shpack@korea.ac.kr.*
- *X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, EIT 4155, Waterloo, Ontario, N2L 3G1, Canada. E-mail: xshen@bbcr.uwaterloo.ca.*
- *J.W. Mark is with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, EIT 4156, Waterloo, Ontario, N2L 3G1, Canada. E-mail: jwmark@bbcr.uwaterloo.ca.*
- *J. Pan is with the Department of Computer Science, University of Victoria, PO Box 3055, STN CSC, ECS 566, Victoria, BC, V8W 3P6, Canada. E-mail: pan@uvic.ca.*

Manuscript received 30 Jan. 2006; revised 20 Oct. 2006; accepted 13 Nov. 2006; published online 7 Feb. 2007.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0034-0106.
Digital Object Identifier no. 10.1109/TMC.2007.1010.

1536-1233/07/$25.00 © 2007 IEEE Published by the IEEE CS, CASS, ComSoc, IES, & SPS

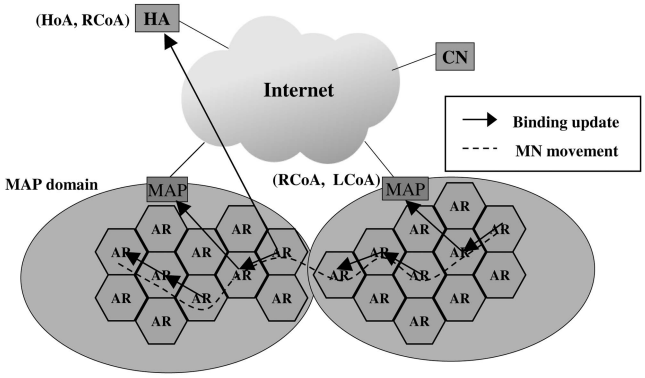


Fig. 1. Binding update procedures in HMIPv6 networks.

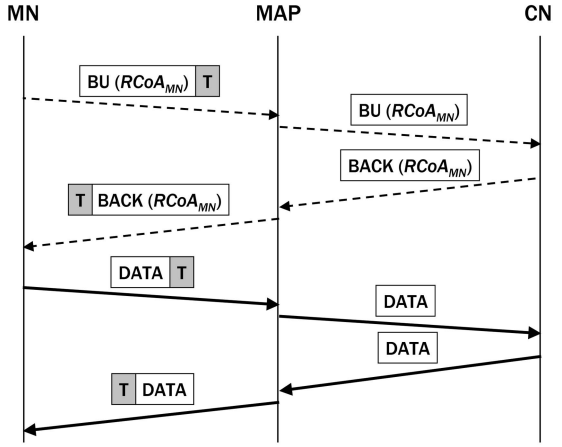


Fig. 2. Packet delivery with the RCoA binding update.

update traffic. This is because the RCoA is not changed while the MN resides in the MAP domain. However, it causes additional tunneling procedures at the MAP and affects the packet delivery latency and throughput. On the other hand, if the MN's LCoA is advertised, the tunneling overhead at the MAP is eliminated, but more binding update traffic is introduced.

To strike a balance between the signaling overhead reduction and the data throughput improvement, we propose an adaptive route optimization (ARO) scheme for HMIPv6 networks. Specifically, ARO determines how to perform RO based on the estimated session-to-mobility ratio (SMR), which is defined as the ratio of the session arrival rate to the handoff rate. If the SMR is greater than a predefined threshold, the session activity has a greater impact on the overall performance than the node mobility. Hence, it needs to reduce the packet delivery overhead rather than the binding update traffic. Therefore, the MN informs the CN of its LCoA for a high SMR (i.e., *LCoA binding update (LBU)*). On the other hand, if the SMR is low, it is better to reduce the binding update traffic so that the MN informs the CN of its RCoA (i.e., *RCoA binding update (RBU)*). Since the SMR threshold value, at which the MN determines which BU scheme (i.e., LBU or RBU) is used, affects both the packet delivery overhead and the binding update overhead, we also determine the optimal SMR threshold to minimize the overall traffic overhead.

Our major contributions in this paper are as follows: 1) We design ARO in which an MN's session activity as well as its mobility are considered. By taking these two factors into account, ARO achieves truly adaptive route optimization in all-IP networks. 2) We develop analytical models for estimating signaling overhead and data throughput, and explicitly derive the optimal SMR threshold for ARO. 3) We justify the feasibility of ARO and validate the analytical results through extensive simulations.

The remainder of this paper is organized as follows: Section 2 compares RCoA and LCoA binding updates. An adaptive route optimization scheme is proposed in Section 3. An analytical model is presented and the optimal SMR threshold is derived in Section 4. Section 5 presents the numerical results. Related work is summarized in Section 6, followed by the concluding remarks in Section 7. The acronyms used throughout the paper are summarized in Appendix A.

2 RCoA Binding Update versus LCoA Binding Update

Fig. 1 illustrates the basic binding update procedures in HMIPv6 networks. An MN is configured with two CoAs: a regional care-of-address (RCoA) and an on-link care-of-address (LCoA). The RCoA is an address in the MAP's subnet. An MN obtains its RCoA when it receives a Router Advertisement (RA) message with the MAP option. On the other hand, the LCoA is an on-link CoA attributed to the MN's interface and is based on the prefix information advertised by an access router (AR). After address configuration, the MN sends a BU message to the MAP in order to bind its current location (i.e., LCoA) with an address in the MAP's subnet (i.e., RCoA). The MAP performs duplicate address detection (DAD) for the MN's RCoA on its link and returns a BACK message to the MN. To register its new RCoA with the HA, the MN sends a BU message which specifies the binding of Home Address (HoA) and RCoA. The HoA is recorded in the home address option (HAO) field and the RCoA can be found in the source address field. The MN also sends BU messages that specify the binding information between the HoA and the RCoA to its CNs, which achieves route optimization. If the MN changes its current address within a MAP domain, it only needs to register a new address (i.e., LCoA) with the MAP. The RCoA is not changed as long as the MN remains in the same MAP domain. This design makes the MN's mobility transparent to the HA and CNs. Since the MAP acts as a local HA, it receives all packets destined to an MN within its domain and tunnels the received packets to the MN's current address. At the same time, all packets originated from MNs within the MAP domain are routed through the MAP for the purpose of symmetric routing.

There are two binding update approaches to achieve RO in HMIPv6 networks: One is *RCoA binding update (RBU)* and the other is *LCoA binding update (LBU)*. Fig. 2 illustrates the binding update and packet delivery procedures in HMIPv6 networks when RBU is employed, where $RCoA_{MN}$ denotes the MN's RCoA and T is the tunneling header. In addition, the solid line represents the data packet flow, whereas the dotted line refers to the signaling packet flow (e.g., BU and BACK messages). In RBU, the packets sent by the CN are first delivered to the MAP and then tunneled to the MN. Consequently, the route between the CN and the MN

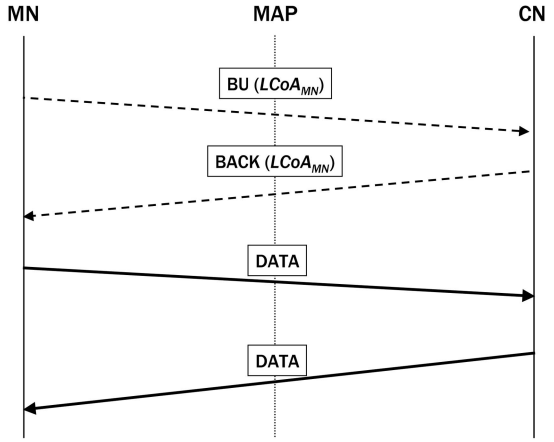


Fig. 3. Packet delivery with the LCoA binding update.

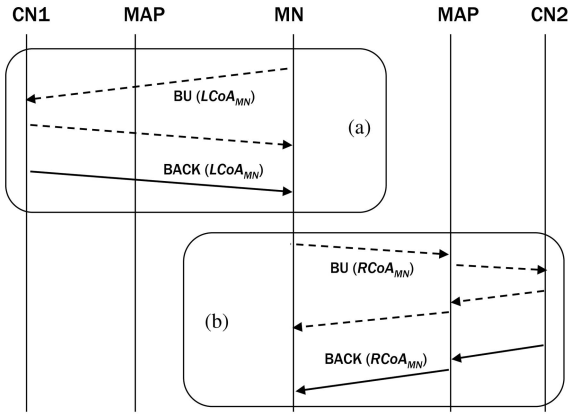


Fig. 4. Packet delivery in ARO.

involves additional tunneling at the MAP, which results in a significant packet delivery latency.

In RBU, the MN sends an encapsulated BU message to the MAP. This is because all packet transmissions in HMIPv6 networks are accomplished through a bidirectional tunnel between the MAP and MN. In the outer header of an encapsulated packet, the MN's LCoA (i.e., $LCoA_{MN}$) and the MAP's address (i.e., $Addr_{MAP}$) are used as the source and destination addresses, respectively. On the other hand, the original BU message, which is decapsulated at the MAP and sent to the CN, uses the MN's RCoA[1] (i.e., $RCoA_{MN}$) and the CN's HoA (i.e., HoA_{CN}) as the source and destination addresses, respectively. The MN's HoA (i.e., HoA_{MN}) is included in the HAO field to identify the MN. In the original BACK message from the CN, HoA_{CN} and $RCoA_{MN}$ are specified in the source and destination address fields, respectively. After receiving the BACK message, the MAP adds a tunneling header with the source address field of $Addr_{MAP}$ and the destination address field of $LCoA_{MN}$.

The binding update and packet delivery procedures of LBU are shown in Fig. 3. Since the LCoA is the same as the CoA in MIPv6, LBU follows the same procedures as those in MIPv6. In LBU, the MN, when changing its LCoA or receiving the tunneled packet from the HA, sends a BU message to the CN. The BU message in LBU includes $LCoA_{MN}$ and HoA_{CN} in the source and destination address fields, respectively. At the same time, HoA_{MN} is specified in the HAO field. In LBU, the MN notifies the CN of its LCoA. Therefore, the BU message is directly delivered to the CN, bypassing the MAP, and, hence, there is no tunneling procedure at the MAP. If the CN receives the BU message, the CN responds with a BACK message and delivers subsequent packets directly to the MN without any tunnelings at the HA or the MAP. To this end, the destination address field of the BACK format is $LCoA_{MN}$.

In LBU, whenever the MN changes its LCoA within the same MAP domain, it notifies the MAP and the CNs recorded in the binding update list of its new LCoA. If the MN moves from one subnet to another frequently, more BU procedures have to be performed and this introduces more signaling traffic over the entire network. Consequently, LBU may result in more binding update traffic than RBU while reducing the tunneling overhead at the MAP.

1. In MIPv6 networks, the source address field of the BU message includes the MN's CoA that is valid in the foreign network in order to prevent ingress filtering. Similarly, HMIPv6 uses the MN's RCoA in the source address field of the BU message sent to its HA/CNs.

3 Adaptive Route Optimization

The adaptive route optimization (ARO) scheme handles the binding updates to the CNs. Therefore, the binding updates to the HA and the MAP are the same as in the HMIPv6 specification [3]. If the session activity is higher than the mobility rate, LBU is more appropriate because it eliminates the MAP tunneling and reduces the packet delivery time. On the contrary, RBU is better than LBU when the mobility rate is relatively high. This is because RBU reduces the number of binding updates to the HA/CNs. Accordingly, ARO should balance the conflicting features of LBU and RBU.

The binding update and packet delivery procedures in ARO are illustrated in Fig. 4. Let δ be the predefined SMR threshold for ARO. In addition, let the SMR of the CN1 be greater than δ and the SMR of the CN2 be smaller than δ. Then, the MN sends a BU message with its LCoA to the CN1 (Fig. 4a) and a BU message with its RCoA to the CN2 (Fig. 4b). Consequently, the packets from the CN1 are directly routed to the MN and the packets from the CN2 are first routed to the MAP and then tunneled to the MN.

In ARO (also in RBU and LBU), the binding update latency to the CN causes packet loss during handoff [8]. To reduce packet loss, the MN can send a BU message to the previous AR or MAP. If LBU is triggered at the previous location, the MN sends a BU message with its LCoA to the previous AR, and then the packets arriving at the previous AR are forwarded to the new AR. This forwarding can be achieved by establishing a tunnel between two ARs, similar to Fast Handover for Mobile IPv6 (FMIPv6) [9]. On the other hand, if RBU is performed at the previous location, the MN sends a BU message with its LCoA to the previous MAP when it moves to a new AR or MAP domain. Then, the packets delivered to the previous MAP can be forwarded by a bidirectional tunnel according to the HMIPv6 standard [3].

To implement ARO, a simple extension to the binding update list is required. Originally, an MN's binding update list maintains binding information, e.g., the address and remaining lifetime for each binding [2], [3]. In ARO, RBU

No	Address	Lifetime	R/L	SMR	δ
1	FF04::4301:1434	1000	R	0.2	0.3
2	3410::3034	500	R	1.0	1.2
3	4200:0923::FF00	320	L	1.3	0.8
4	FF01::6201:1234	180	L	2.0	1.0
...	...	...	...		...

Fig. 5. Modified binding update list.

and LBU are adaptively selected so that the binding update list includes an R/L flag, as shown in Fig. 5. If the flag is set to R, it means that the binding is resulted from RBU. Otherwise, the binding can be made by LBU. In addition, each entry in the binding update list has two additional fields, SMR and δ, to record the estimated SMR and SMR threshold, respectively.

Fig. 6 shows a procedure performed when an MN receives a packet from a CN. The MN determines whether the packet belongs to a current ongoing session or to a new session. To this end, an active state timer with length T_A is maintained at the MN [10], [11]. If the time duration between the last received packet and the current received packet is greater than T_A, the current packet is considered as the first packet of a new session. Otherwise, the packet is a subsequent packet of an ongoing session. T_A has a significant impact on session activity, which has been investigated in [12]. How to determine T_A properly is application-specific and implementation-dependent, and it is beyond the primary focus of this paper.

For an arrived packet of the ongoing session, no binding update to the CN is triggered. Even though applying an adaptive binding update to the CN on a per-packet basis may yield better performance, it may result in out-of-order packet delivery [13]. Consequently, ARO is performed only when a new session is established or a handoff occurs.[2] However, since the packet arrival of the ongoing session affects several parameters, e.g., session length and hop distance, the session information is updated on the packet arrival.

On the other hand, for a new session arrival, the MN checks whether the CN is the registered one in the binding update list. If the CN is a new one (i.e., the session is the first established session from the CN), the MN creates a new entry for the CN in the binding update list. After that, the SMR of the CN should be determined. If the session is the first one from the CN, the SMR of the CN is set to a default value (i.e., 1.0). Otherwise, the MN estimates the CN's SMR by counting the total number of sessions established with the CN and the number of handoffs performed by the MN. To avoid oscillation of the estimated SMR, an exponentially weighted moving average (EWMA) scheme is employed [14]. In addition, the MN should determine the optimal δ, which is affected by several parameters, i.e., the session length, hop distances among network entities (e.g., CN, MN, HA, and MAP), MAP domain size, and BU/BACK packet lengths. The BU/BACK packet lengths and tunneling header length have been specified in [3]. The hop distance can be estimated by examining the packet header, i.e., it is obtained by subtracting the final time-to-live (TTL) (in the IP header) from the initial TTL to be inferred. According to [2], a return routability (RR) test should be performed before a binding update procedure to avoid security attacks on RO. Therefore, the hop distance values can be obtained by monitoring the packets exchanged during the RR test. The MAP domain size can be obtained from the RA message broadcasted by the MAP.[3] In addition, the average session length is estimated based on the EWMA scheme. Using these values, the optimal δ can be calculated. After updating SMR and δ, the MN performs a binding update procedure to the CN, i.e., either LBU or RBU is performed, and records the result in the binding update list with the R/L flag.

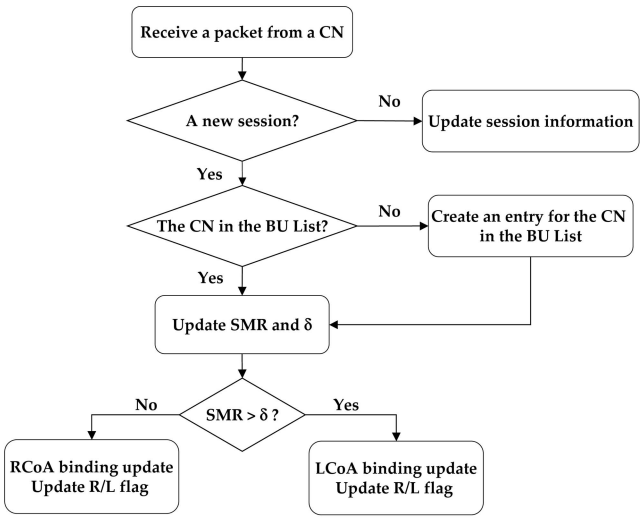


Fig. 6. ARO procedure on receiving a packet.

When an MN moves to another AR subnet, its LCoA is changed. In HMIPv6 networks, the MN does not inform CNs of the new LCoA because only the RCoA is visible to the CNs. However, in ARO, since the SMR and δ are changed by a handoff event, a new RO policy should be applied. Consequently, the procedure illustrated in Fig. 7 is performed whenever a handoff event occurs. This procedure can be performed for all CNs registered in the binding update list. In Fig. 7, i and N denote the index for a CN and the number of CNs registered in the binding update list, respectively. For each CN, the MN updates its SMR and δ. After that, the MN performs RBU or LBU based on the updated SMR and δ, and the result of RO is recorded with the R/L flag.

4 Performance Analysis

In this section, we analyze the performance of ARO. Without loss of generality, we make the following assumptions and notations:

2. The session-based binding update imposes higher computation overhead than the existing MIPv6 and HMIPv6. However, the advances of computation technologies enable the session-based binding update to be implemented in a practical manner [15].

3. ARO works well even when the MAP domain size is unknown. This is because the sensitivity of the optimal δ to different MAP domain sizes is not noticeable (see Section 5).

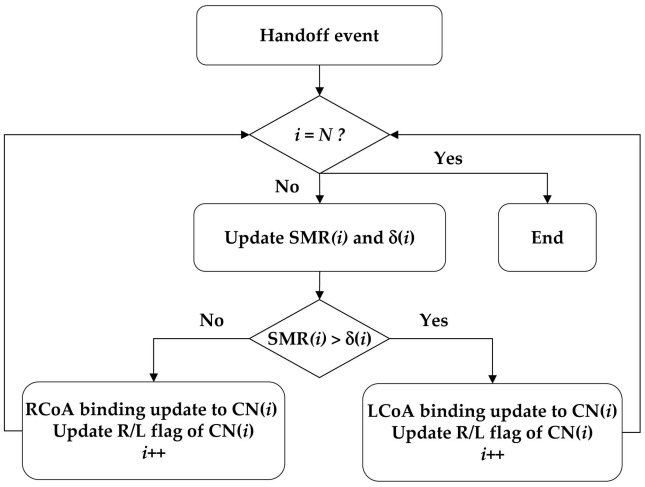


Fig. 7. ARO procedure on a handoff event.

- The session interarrival time follows an exponential distribution with rate λ_S and average session length, in packets, is $E(S)$.
- The AR subnet residence time follows a general distribution with mean $1/\mu_S$. Its probability density function (pdf) is $f_S(t)$. Let $F_S(t)$ and $f_S^*(s)$ denote the cumulative distribution function and the Laplace transform of $f_S(t)$, respectively.
- The MAP domain residence time follows a general distribution with mean $1/\mu_D$ and its pdf is $f_D(t)$. Let $F_D(t)$ and $f_D^*(s)$ denote the cumulative distribution function and Laplace transform of $f_D(t)$, respectively.
- Let t_I and t_S be the random variables for intersession arrival time and AR subnet residence time, respectively. Then, an SMR random variable (r) is given by t_S/t_I and its mean (ρ_S) is λ_S/μ_S. The SMR cumulative distribution can be derived as (see Appendix B)

$$\Pr(r < \delta) = \Pr(t_S/t_I < \delta) = f_S^*(s)\big|_{s=\lambda_S/\delta}.$$

4.1 Total Cost Model

To quantify the overall traffic overhead in ARO, we develop an analytical cost model consisting of the binding update (BU) cost (C_{BU}) to the CN and the packet tunneling (PT) cost (C_{PT}) from the CN to the MN. The BU and PT costs are considered as the additional traffic loads due to exchanging BU/BACK messages and IP tunneling headers[4] of data packets, respectively [16]. Since route optimization for a CN is performed until the lifetime (T_{BU}) of a binding entry expires [2], the BU and PT costs during T_{BU} are considered. The following notations are used for the cost model:

- L_T: Tunneling header length (40 bytes).
- $L_{BU}(L_{BACK})$: BU (BACK) message length.
- d_{A-B}: Hop distance between A and B.
- C_{A-B}: Unit packet tunneling cost from A to B.
- BU_{CN}: Unit binding update cost to the CN.

4. In terms of packet tunneling cost, only the IP tunneling header is considered because the original IP packet does not introduce extra overhead.

As a reference scheme, we first derive the BU and PT costs of the nonroute optimization (NRO) scheme in HMIPv6 networks. Since no BU procedure to the CN is performed in NRO, the BU cost of NRO is simply

$$C_{BU}^{NRO} = 0. \tag{1}$$

Regarding packet delivery in NRO, all packets transmitted by the CN are first routed to the HA, and then the HA tunnels them to the MAP. By Little's law [17], the number of packets delivered during T_{BU} can be approximated by $\lambda_S \cdot T_{BU} \cdot E(S)$. Therefore, the PT cost of NRO is obtained from

$$C_{PT}^{NRO} = \lambda_S \cdot T_{BU} \cdot E(S) \cdot (C_{HA-MAP} + C_{MAP-MN}), \tag{2}$$

where C_{HA-MAP} is given by $L_T \cdot d_{HA-MAP}$. Since there are no tunneling headers for data packets from the CN to the HA, the packet tunneling cost between them is zero. On the other hand, the MAP retunnels the packets to the MN. Therefore, two IP tunneling headers are added to the packets from the MAP to the MN and, hence, C_{MAP-MN} is equal to $2L_T \cdot d_{MAP-MN}$.

When RBU is used, the BU cost is given by

$$C_{BU}^{RBU} = \mu_D \cdot T_{BU} \cdot BU_{CN}, \tag{3}$$

where the product, $\mu_D \cdot T_{BU}$, represents the average number of domain crossings during T_{BU}. BU_{CN} can be calculated as $d_{MAP-MN} \cdot (L_T + L_{BU}) + d_{CN-MAP} \cdot L_{BU} + d_{MAP-MN} \cdot (L_T + L_{BACK}) + d_{CN-MAP} \cdot L_{BACK}$.[5]

In RBU, packets sent by a CN are first delivered to the MAP. In this portion of the delivery, no tunneling is employed, so the PT cost is zero. However, an IP tunneling header is augmented in the delivery from the MAP to the MN. Hence, the PT cost of RBU can be computed as

$$C_{PT}^{RBU} = \lambda_S \cdot T_{BU} \cdot E(S) \cdot C_{MAP-MN}, \tag{4}$$

where C_{MAP-MN} is $L_T \cdot d_{MAP-MN}$.

For LBU, the MN sends a BU message whenever it moves away from a subnet area. Then, the average number of subnet crossings during T_{BU} is equal to $\mu_S \cdot T_{BU}$, where μ_S is the subnet crossing rate. Therefore, the BU cost of LBU is given by

$$C_{BU}^{LBU} = \mu_S \cdot T_{BU} \cdot BU_{CN}, \tag{5}$$

where BU_{CN} is $(d_{MAP-MN} + d_{CN-MAP}) \cdot (L_{BU} + L_{BACK})$.

Although LBU results in a higher BU cost due to frequent binding updates, it can reduce the PT cost by eliminating the MAP tunneling. After route optimization, the HA tunneling is also eliminated. Therefore, the PT cost of LBU is

$$C_{PT}^{LBU} = 0. \tag{6}$$

In ARO, the MN uses RBU if the SMR of a CN is equal to or lower than δ; otherwise, the MN uses LBU. Hence, ARO includes the total costs of RBU and LBU depending on the

5. In HMIPv6 networks, the packets destined to the MN are tunneled at the MAP. At the same time, the packets originated from the MN are also reversely tunneled to the MAP and then they are transmitted to the CN by the MAP. More details can be found in Section 2.

TABLE 1
Summary of the Packet Delivery Time

Term	Expression
$S(CN, HA)$	$d_{CN-HA} \times (\frac{L_P}{BW} + l) + (d_{CN-HA} - 1) \times P_R + P_{HA}$
$S(HA, MAP)$	$d_{HA-MAP} \times (\frac{L_P+L_T}{BW} + l) + (d_{HA-MAP} - 1) \times P_R + P_{MAP}$
$S^{NRO}(MAP, MN)$	$(d_{MAP-MN} - 1) \times (\frac{L_P+2L_T}{BW} + l) + (\frac{L_P+2L_T}{BW_W} + l_W) + (d_{MAP-MN} - 1) \times P_R$
$S^{RO}(MAP, MN)$	$(d_{MAP-MN} - 1) \times (\frac{L_P+L_T}{BW} + l) + (\frac{L_P+L_T}{BW_W} + l_W) + (d_{MAP-MN} - 1) \times P_R$
$S(CN, MAP)$	$d_{CN-MAP} \times (\frac{L_P}{BW} + l) + (d_{CN-MAP} - 1) \times P_R + P_{MAP}$
$S(CN, MN)$	$(d_{CN-MN} - 1) \times (\frac{L_P}{BW} + l) + (\frac{L_P}{BW_W} + l_W) + (d_{CN-MN} - 1) \times P_R$

SMR, i.e., $\Pr(r > \delta) \cdot C_T^{LBU} + \Pr(r \leq \delta) \cdot C_T^{RBU}$, where r is the SMR random variable. Also, in ARO, a session arrival can trigger a binding update because it affects the value of the SMR. If the previous binding update was an LCoA binding update, no binding update is performed on the session arrival. This is because a session arrival increases the SMR and, hence, an RCoA binding update cannot be triggered after an LCoA binding update. On the other hand, if there is a number of session arrivals after an RCoA binding update, an LCoA binding update can be performed on the session arrival. Let θ be the probability that an LCoA binding update is triggered after an RCoA binding update on session arrivals and α and β be the SMRs when the RCoA and LCoA binding updates are performed, respectively. If there are K session arrivals between the RCoA and LCoA binding updates, the following relationship between α and β can be established by the SMR definition.

$$\beta = \begin{cases} \alpha + K, & \alpha \geq 1 \\ \alpha(1 + K), & \alpha < 1, \end{cases} \tag{7}$$

where the average K is given by λ_S/μ_S [18]. Then, θ is expressed as a conditional probability,

$$\theta = \Pr(\beta > \delta | \alpha \leq \delta) = \frac{\Pr(\beta > \delta, \alpha \leq \delta)}{\Pr(\alpha \leq \delta)}, \tag{8}$$

where $\Pr(\alpha \leq \delta)$ and $\Pr(\beta > \delta)$ refer to the probabilities of RCoA and LCoA binding updates being performed, respectively. The additional BU procedures on session arrivals can be initiated only after an RCoA binding update and the average number of RCoA binding updates during T_{BU} is $\mu_S T_{BU} \Pr(r \leq \delta)$. From the SMR distribution, θ can be calculated numerically. Consequently, the total cost of ARO can be represented by

$$\begin{aligned} C_T^{ARO} = {} & \Pr(r > \delta) \cdot C_T^{LBU} + \Pr(r \leq \delta) \cdot C_T^{RBU} \\ & + \mu_S T_{BU} \Pr(r \leq \delta) \cdot \theta \cdot BU_{CN}. \end{aligned} \tag{9}$$

Based on the cost model, the optimal SMR threshold, δ^*, can be obtained from Theorem 1.

Theorem 1. *If λ_S is greater than $\frac{1}{T_{BU} \cdot E(S) \cdot C_{MAP-MN}}$, there exists an SMR optimal threshold in $(0, \infty)$ and it is given by*

$$\delta^* = \frac{(d_{MAP-MN} \cdot (L_T + L_{BU}) + d_{CN-MAP} \cdot L_{BU} + d_{CN-MN} \cdot L_{BACK}) \cdot (1 - 1/\sqrt{N})}{E(S) \cdot L_T \cdot d_{MAP-MN}}. \tag{10}$$

Proof. See Appendix C. □

From Theorem 1, δ^* depends on the session length $(E(S))$, domain size (N), and hop distances. The effects of these parameters will be investigated in Section 5.

4.2 Throughput Model

For the throughput model, we assume a session model where $E(S)$ packets are exchanged between the MN and the CN and the interpacket time is σ. That is, $E(S)$ packets with interval σ are sent by the CN at the beginning of the session. A session transmission can be disrupted by handoff events and packets cannot be transmitted during the disruption period. Therefore, the total transmission time is the sum of handoff disruption time and packet delivery time. In addition, throughput is defined as the ratio of transmitted data volume over total transmission time. The following is a list of the notations and values used in the throughput model [16]:

- $S(i, j)$: Packet delivery time between i and j.
- BW: Wired link bandwidth (100 Mbps).
- BW_W: Wireless link bandwidth (11 Mbps).
- D^X: End-to-end packet delivery time in X ($X \in \{NRO, RBU, LBU\}$).
- P_R: Processing time in the router: routing table lookup and packet processing time (0.001 msec).
- P_{HA}: Processing time in the HA (0.005 msec).
- P_{MAP}: Processing time in the MAP (0.003 msec).
- l: Wired link latency: propagation delay and link layer delay (0.5 msec).
- l_W: Wireless link latency: propagation delay and link layer delay (2 msec).

The end-to-end packet delivery time is defined as the total time elapsed when a packet of a session is delivered from the CN to the MN. Hence, the end-to-end packet delivery time is the sum of packet delivery time among network entities (e.g., CN, MN, HA, and MAP). To obtain the packet delivery time, the packet transmission delay and the link latency should be considered [16], [19], [20]. For instance, if a packet of size L_P is delivered over a wired link, the packet delivery time is equal to $L_P/BW + l$. Table 1 shows the packet delivery time among network entities. $S^{NRO}(MAP, MN)$ is the packet delivery time from the MAP to the MN before route optimization is performed. In other words, the packet is tunneled both at the HA and the MAP and, hence, the tunneling header size between the MAP and the MN is $2L_T$. On the other hand, since $S^{RO}(MAP, MN)$ is the packet delivery time when the RCoA binding update is

performed, the tunneling header size between the MAP and the MN is L_T.

In NRO, all packets are routed to the MN via the HA. The end-to-end packet delivery time of NRO is given by

$$D^{NRO} = S(CN, HA) + S(HA, MAP) + S^{NRO}(MAP, MN). \tag{11}$$

For RBU, packets are delivered to the MN through the MAP, bypassing the HA. Therefore, the end-to-end packet delivery time can be represented by

$$D^{RBU} = S(CN, MAP) + S^{RO}(MAP, MN). \tag{12}$$

For LBU, packets are directly delivered to the MN, bypassing the HA and the MAP. Hence, the packet delivery time can be expressed as

$$D^{LBU} = S(CN, MN). \tag{13}$$

The handoff latency is defined as the completion time of the BU procedures from the MN to the CN. Let H^{RBU} and H^{LBU} be the handoff latencies in RBU and LBU, respectively. Then, H^{RBU} can be computed as in (14), where the first and second terms of the right-hand side represent the delivery time of the BU message from the MN to the CN via the MAP, and the third and fourth terms refer to the delivery time of the BACK message from the CN to the MN via the MAP. For LBU, no tunneling header is used. Therefore, H^{LBU} can be calculated as in (15).

$$\begin{aligned} H^{RBU} = {} & (d_{MAP-MN} + d_{CN-MAP} - 1) \\ & \times \left(\frac{L_{BU} + L_T}{BW} + l + P_R\right) + \left(\frac{L_{BU} + L_T}{BW_W} + l_W\right) \\ & + \left(\frac{L_{BACK} + L_T}{BW_W} + l_W\right) \\ & + (d_{CN-MN} - 1) \times \left(\frac{L_{BACK} + L_T}{BW} + l + P_R\right), \end{aligned} \tag{14}$$

$$\begin{aligned} H^{LBU} = {} & (d_{MAP-MN} + d_{CN-MAP} - 1) \times \left(\frac{L_{BU}}{BW} + l + P_R\right) \\ & + \left(\frac{L_{BU}}{BW_W} + l_W\right) + \left(\frac{L_{BACK}}{BW_W} + l_W\right) \\ & + (d_{CN-MN} - 1) \times \left(\frac{L_{BACK}}{BW} + l + P_R\right). \end{aligned} \tag{15}$$

By (11), (12), and (13), the throughput for each scheme can be obtained. In NRO, there is no binding update and thus handoff latency is 0. Therefore, the session delivery time, which is defined as the total delivery time for $E(S)$ packets, is $D^{NRO} + (E(S) - 1) \cdot \sigma$ and the throughput of NRO can be computed as

$$T^{NRO} = \frac{E(S) \cdot L_P}{D^{NRO} + (E(S) - 1) \cdot \sigma}. \tag{16}$$

The average numbers of intrahandoffs (i.e., subnet crossing) and interhandoffs (i.e., domain crossing) per session are μ_S/λ_S and μ_D/λ_S, respectively [18]. For RBU and LBU, the handoff latency should be considered to calculate the session delivery time. Then, the throughputs of RBU and LBU are respectively obtained from

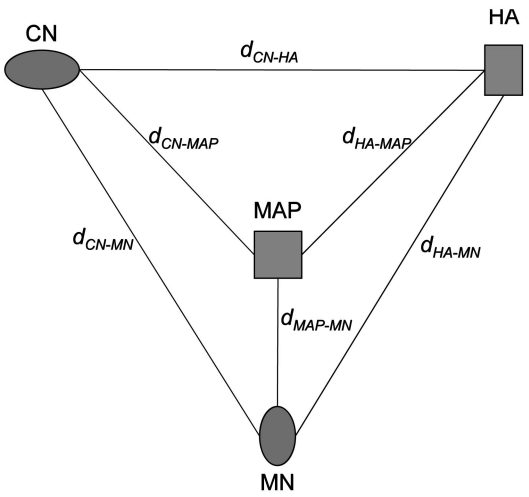


Fig. 8. Network topology under consideration.

$$T^{RBU} = \frac{E(S) \cdot L_P}{D^{RBU} + (E(S) - 1) \cdot \sigma + H^{RBU} \cdot \mu_D/\lambda_S} \tag{17}$$

and

$$T^{LBU} = \frac{E(S) \cdot L_P}{D^{LBU} + (E(S) - 1) \cdot \sigma + H^{LBU} \cdot \mu_S/\lambda_S}. \tag{18}$$

By combining (17) and (18), depending on the SMR distribution, the throughput of ARO can be obtained from

$$T^{ARO} = \frac{E(S) \cdot L_P}{\Pr(r > \delta)(D^{LBU} + (E(S) - 1)\sigma + H^{LBU} \cdot \mu_S/\lambda_S) + \Pr(r \le \delta)(D^{RBU} + (E(S) - 1)\sigma + H^{RBU} \cdot \mu_D/\lambda_S)}. \tag{19}$$

5 Numerical Result

To calculate the total costs of these four RO schemes, the unit costs of PT and BU should be determined in advance. The unit cost is calculated as the product of message length and hop distance, and its unit is $Kbytes * hops$ [16], [19]. From [2], [3], the default lengths of BU and BACK messages in the HMIPv6 specification are used. The average data packet size (L_P) is assumed to be 1,000 bytes. The network topology under consideration is depicted in Fig. 8. We focus on the tunneling overhead at the MAP and, therefore, the hop distance from the CN to the MN is the same no matter whether the MAP is visited or not, i.e., $d_{CN-MN} = d_{CN-MAP} + d_{MAP-MN}$. Similarly, the hop distance from the HA to the MN is constant no matter whether the MAP is visited or not, i.e., $d_{HA-MN} = d_{HA-MAP} + d_{MAP-MN}$. Important parameter values for numerical analysis are summarized in Table 2.

5.1 Optimal Threshold

The optimal SMR threshold as a function of average session length $(E(S))$ is shown in Fig. 9. It can be seen that the optimal SMR threshold is inversely proportional to $E(S)$. When $E(S)$ is small, i.e., for short-lived sessions, it is more efficient to reduce the BU cost rather than the PT cost. This is because the BU cost is a more dominant factor than the PT cost. In order to reduce the BU cost, the probability of using

TABLE 2
Parameters Used in Numerical Analysis

$d_{CN-HA}(hops)$	$d_{HA-MN}(hops)$	$d_{CN-MAP}(hops)$	$d_{MAP-MN}(hops)$	$d_{CN-MN}(hops)$
4	8	6	4	10
$d_{HA-MAP}(hops)$	$T_{BU}(sec)$	N	$E(S)(packets)$	λ_S
4	1000	49	10	1

RBU should be increased by adopting a large SMR threshold. On the contrary, if $E(S)$ is large, the PT cost occupies a larger portion of the total cost when compared with the BU cost. Accordingly, the SMR threshold should be decreased in order to use LBU more frequently.

Fig. 9 also demonstrates the effect of MAP domain size N. In a large MAP domain, most BU/BACK messages are handled by the MAP, not the HA. As a result, a large MAP domain can significantly reduce the BU cost compared to a small MAP domain. Hence, if N is large, RBU is more appropriate than LBU because RBU can significantly reduce the BU cost. However, the variation of δ^* for different MAP domain sizes is not apparent. Therefore, even though the MAP domain size is not given a priori and an SMR threshold obtained by a default MAP domain size is used, performance similar to that with the optimal threshold can still be achieved.

5.2 Total Cost Analysis

Fig. 10 shows the total cost for different average SMRs, i.e., $\rho_S = \lambda_S/\mu_S$. When ρ_S is low, the mobility rate is relatively higher than the session arrival rate. Therefore, RBU is better because it provides a reduced BU cost by not notifying the CN of the LCoA changes. However, as ρ_S increases, the total cost of LBU is drastically reduced. This is because LBU does not involve any additional processing at the MAP. Consequently, LBU is better in the reduction of the PT cost, which is more important when ρ_S is high. Compared to RBU and LBU, ARO adapts to changes induced by mobility and, thus, exhibits total cost reduction. In other words, when ρ_S is low, the total cost of ARO is similar to that of RBU. When ρ_S exceeds a certain point (i.e., the optimal SMR threshold), the ARO's total cost approaches that of LBU. Consequently, ARO is highly effective for MNs that change their mobility patterns from time to time and have diverse session activities.

In ARO, the average session size is estimated by the EWMA mechanism and it is used to determine the optimal SMR threshold. However, since the session length is highly variable, we investigate the effect of session length distribution on the performance of ARO via simulations. In the simulation, the subnet and MAP domain residence times are assumed to have Gamma distributions. In addition, the session length (in units of Kbytes) follows a Lognormal distribution with mean 10 Kbytes and variance 625 Kbytes [21], [22]. Since a fixed packet size is assumed (i.e., 1 Kbytes), the average session length in units of packets is 10 and the variance is 625. The simulation results are plotted in Fig. 10. Even though some discrepancies between analytical and simulation results can be observed due to high variance, they do not have significant effects on the performance of ARO. In other words, simulation results indicate that the analytical results based on the average session length are sufficiently accurate for evaluating the performance of ARO against RBU and LBU.

The session size affects the PT cost as well as the optimal SMR threshold. Fig. 11 shows the total cost gains of RBU, LBU, and NRO for different $E(S)$. The total cost gain is defined as the relative total cost of ARO over the total cost of a scheme, i.e., RBU, LBU, or NRO. If the total cost gain of a scheme is larger than 1.0, it means the corresponding scheme is more efficient than ARO. For RBU and NRO, as $E(S)$ becomes large, the total cost gain is reduced. In other

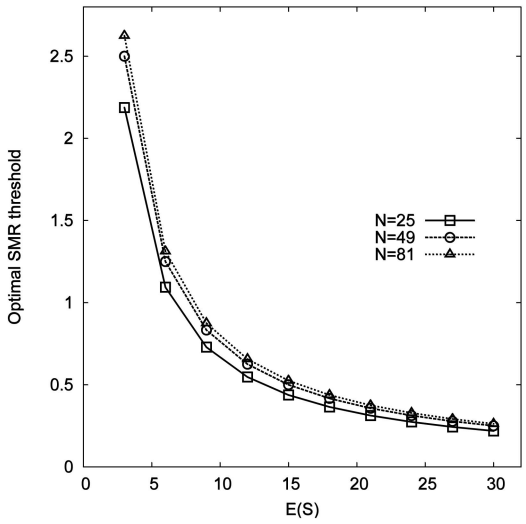


Fig. 9. Optimal SMR threshold versus $E(S)$.

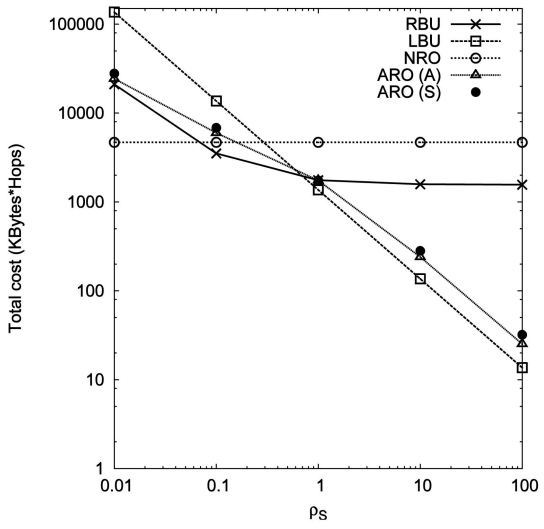


Fig. 10. Total cost as a function of ρ_S (A: Analysis, S: Simulation).

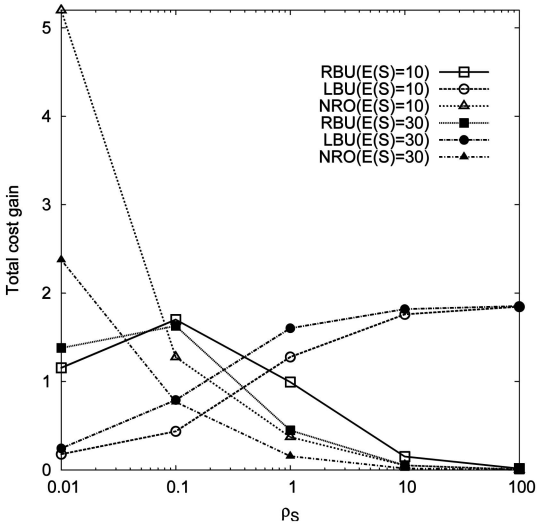


Fig. 11. Total cost gain for different $E(S)$.

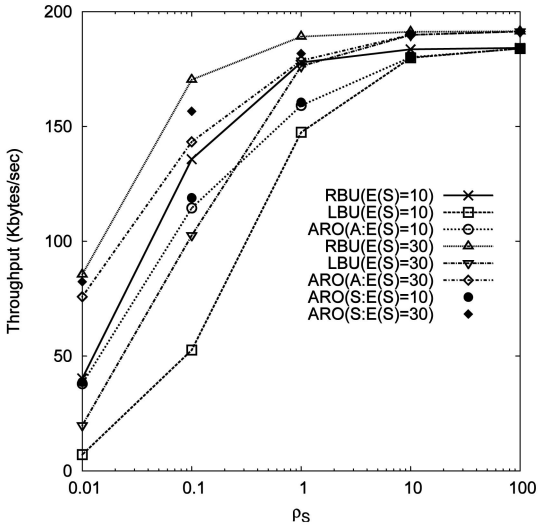


Fig. 12. Throughput as a function of ρ_S (A: Analysis, S: Simulation).

words, RBU and NRO are more effective choices for small-size session. This is because the main advantage of RBU and NRO is to mitigate the BU cost, not the PT cost. In contrast, the total cost gain of LBU, which provides a reduced PT cost, is higher when $E(S)$ is large. However, for both RBU and LBU, the effect of $E(S)$ on the total cost gain diminishes as ρ_S becomes extremely high or low.

Another interesting observation is that the total cost gains of RBU and LBU are bounded to about 1.62 and 1.85, respectively, regardless of $E(S)$. It implies that the total cost of ARO does not exceed 62 percent of the RBU's total cost and 85 percent of the LBU's total cost even in the worst case. This is because ARO chooses RBU or LBU adaptively based on the SMR. On the contrary, if RBU is always used, although the RBU's total cost gain against LBU is higher than 1.0 when ρ_S is low, the total cost gain becomes negligible for a high ρ_S. For instance, the RBU's cost gain against LBU is 6.48 when ρ_S is 0.01, but it is reduced to 0.01 when ρ_S becomes 100. The opposite result is observed in the LBU's total cost gain against RBU. These results reveal that ARO should be employed to minimize the total cost adaptively.

5.3 Throughput Analysis

The throughput variation as a function of ρ_S is illustrated in Fig. 12. When ρ_S is low, the mobility rate is relatively higher than the session arrival rate. Hence, reducing the handoff latency can improve the throughput. Since RBU can mitigate the average handoff latency compared with LBU, the throughput of RBU is the highest when ρ_S is low. On the contrary, the LBU's throughput is the lowest, whereas the throughput of ARO is about 84.1 percent to 94.4 percent of the RBU's throughput when ρ_S is lower than 1.0.

Since λ_S is fixed at 1.0 in the numerical analysis, the number of handoffs decreases with the increase of ρ_S. Consequently, as ρ_S increases, the average session delivery time decreases and, hence, the throughput of each scheme increases. However, the increased throughput converges to a maximum throughput, i.e., about 184.2 Kbytes/sec and 191.5 Kbytes/sec when $E(S)$ is 10 and 30, respectively. This is because our throughput model considers only one session. Although it is not clearly seen in Fig. 12, the throughput of LBU is the highest when ρ_S is 100. In other words, if the mobility rate is not high, LBU is better with respect to maximizing throughput. Similar to the total cost, the throughput of ARO approaches that of LBU as ρ_S increases so that it shows a higher throughput than RBU with a high ρ_S. In short, it can be seen that ARO improves throughput in an adaptive manner with regard to the SMR in Fig. 12.

For all schemes with a large $E(S)$ (i.e., $E(S) = 30$), they achieve a higher throughput. However, the improvement becomes insignificant at large values of ρ_S. For instance, the improvements of the ARO's throughput by a large $E(S)$ are 99.9 percent and 0.04 percent when ρ_S is 0.01 and 100, respectively. This implies that the handoff latency is a key factor to improve the throughput. When ρ_S is low, more handoffs are performed. It means that a session is disrupted by frequent handoffs and, thus, the session delivery time is highly dependent on the handoff latency. Then, as shown in (17), (18), and (19), $E(S)$ affects the throughput directly. On the other hand, if ρ_S is high, the effect of handoff latency on the session delivery time becomes negligible and the throughput of the LBU (and RBU and ARO) is not improved significantly although $E(S)$ increases. Fig. 12 also reveals that the simulation results coincide with the analytical results, especially when $E(S)$ is 10.

Fig. 13 shows the effect of L_P on the throughput. As described before, RBU exhibits the highest throughput with a low ρ_S (Fig. 13a) while LBU maximizes the throughput with a high ρ_S (Fig. 13b). Also, similar to the effect of $E(S)$, the throughput increases as L_P increases. Especially, the effect of L_P is more obvious with a high ρ_S and RBU. If ρ_S is high, the mobility rate is low, so that the effect of handoff latency on the throughput (or the session deliver time) is trivial. Consequently, the throughputs for RBU and LBU approach $E(S) \cdot L_P/(D^{RBU} + (E(S)-1)\cdot\sigma)$ and $E(S) \cdot L_P/(D^{LBU} + (E(S)-1)\cdot\sigma)$, respectively. It means that the throughput with a high ρ_S is

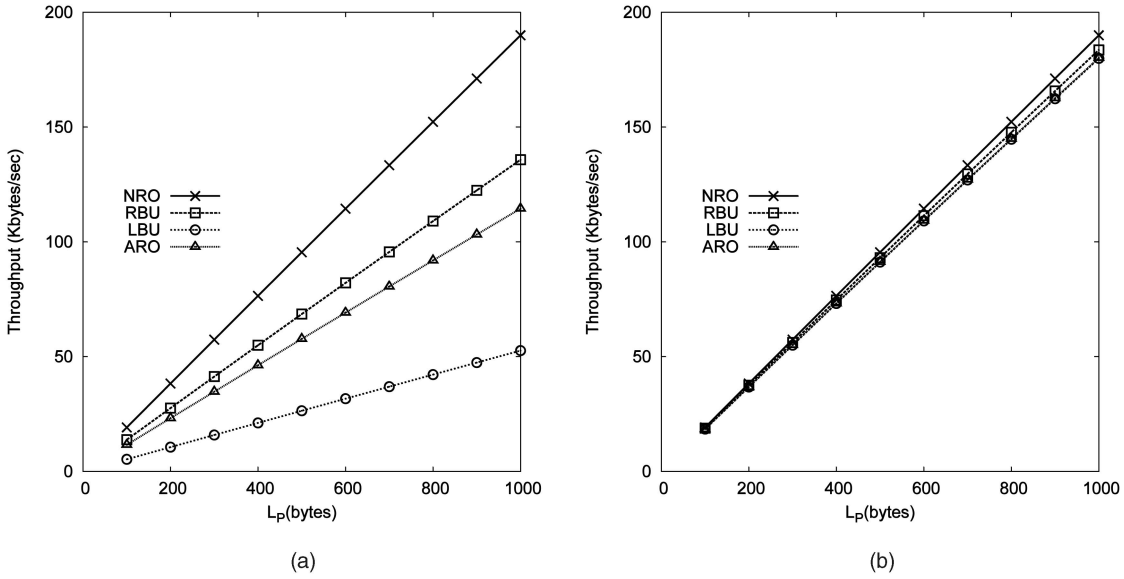


Fig. 13. Throughput as a function of L_P. (a) $\rho_S = 0.1$. (b) $\rho_S = 10$.

proportional to L_P. On the other hand, if ρ_S is low, the session delivery time of RBU is the lowest, so that the RBU's throughput increases more significantly as L_P increases. Namely, for the throughput of RBU,

$$E(S) \cdot L_P/(D^{RBU} + (E(S) - 1) \cdot \sigma + H^{RBU} \cdot \mu_D/\lambda_S),$$

since the denominator is the smallest, the throughput is more sensitive to an increase in the numerator.

6 Related Work

In [7], Lee and Akyildiz propose a scheme to reduce the costs caused by RO in Mobile IPv4 [23]. The link and signaling cost functions are developed in order to capture the trade-off between the network resources consumed by the packet routing, signaling, and processing load incurred by RO in Mobile IPv4 networks. With the cost functions, RO is performed only when it can minimize the total cost, which provides the optimal result from the viewpoint of link and signaling costs. Simulation results demonstrate that the proposed scheme provides the lowest total costs.

In [24], Rajagopalan and Badrinath point out that Mobile IPv4 with RO is not always efficient over all possible call-to-mobility ratio (CMR) scenarios. Based on this observation, they propose a new RO scheme in Mobile IPv4 networks to reduce the total cost regardless of the CMR. In the proposed scheme, an MN is allowed to dynamically choose the packet routing policy depending on the CMR. The packet routing policy consists of two schemes: a triangle routing scheme in the basic Mobile IPv4 and a static update scheme. The static update scheme refers to the RO scheme [25] where the MN actively notifies CNs of its current CoA whenever it moves. The basic triangle routing scheme is used when the CMR is low, whereas the static update scheme is chosen when the CMR is high. Simulation results show that the proposed scheme performs well regardless of the CMR. This scheme is also theoretically evaluated in [26].

However, unlike Mobile IPv4, RO is a mandatory function in the Mobile IPv6 specification. Therefore, the schemes in [7], [24] are not feasible solutions in Mobile IPv6 networks.

Recently, Hwang et al. [27] propose an adaptive scheme using mobility profile in HMIPv6 networks, where some packets are directly delivered to an MN if it seems to reside for a long time in the current subnet. The residence time is maintained in the profile database. However, this scheme relies on only the mobility rate and the session activity is not considered. Therefore, it is not adaptive to the change of session activity. Moreover, the threshold residence time is yet to be determined, which has a significant effect on the performance of the proposed scheme.

In [28], a nonoptimal route problem is investigated when a CN and an MN are located in the same MAP domain. Although there exist shorter routes between two nodes, all packets are inefficiently routed through the MAP, which increases packet delivery time. By checking the source address of the received packets, it is possible to detect that these two nodes are colocated in the same MAP domain. Then, the packet delivery path between these two nodes is adaptively optimized. However, this scheme can be used only for two nodes in the same MAP domain.

7 Conclusion

We have proposed an adaptive route optimization (ARO) scheme, which integrates the best of RBU and LBU depending on the estimated SMR. By employing adaptive binding updates, ARO optimizes binding update traffic and minimizes tunneling overhead at the same time. Extensive analysis results reveal that the performance of ARO is not severely degraded even though an MN's traffic and mobility characteristics are drastically changed. We have also demonstrated that ARO is feasible in practical environments through comprehensive simulations. It is conjectured that ARO is a suitable route optimization scheme for diverse mobile environments where the MNs have different session arrival rates and mobility patterns.

APPENDIX A
ACRONYMS

The acronyms used in this paper are listed as follows:

- ARO: Adaptive Route Optimization.
- CN: Correspondent Node.
- HAO: Home Address Option.
- HMIPv6: Hierarchical Mobile IPv6.
- HoA: Home Address.
- LBU: LCoA Binding Update.
- LCoA: On-link Care of Address.
- MAP: Mobility Anchor Point.
- MIPv6: Mobile IPv6.
- MN: Mobile Node.
- NRO: Non-route Optimization.
- RBU: RCoA Binding Update.
- RCoA: Regional Care of Address.
- SMR: Session-to-Mobility Ratio.

APPENDIX B
SMR DISTRIBUTION

The SMR is defined as t_S/t_I, where t_I follows an exponential distribution with rate λ_S and t_S follows a general distribution. Then, the SMR distribution function can be derived as

$$\begin{aligned}\Pr(r<\delta) &= \Pr(t_S/t_I<\delta)=\Pr(t_S<\delta\cdot t_I)\\ &= \int_0^\infty \Pr(t_S<\delta\tau)\cdot\lambda_S e^{-\lambda_S\tau}\cdot d\tau.\end{aligned} \tag{20}$$

Letting t equal to $\delta\tau$, the right-hand side of (20) becomes

$$\frac{\lambda_S}{\delta}\cdot\int_0^\infty \Pr(t_S<t)\cdot e^{\frac{-\lambda_S t}{\delta}}\cdot dt. \tag{21}$$

By the definition of the Laplace transform [17], $\int_0^\infty \Pr(t_S<t)\cdot e^{-st}\cdot dt=\frac{f_S^*(s)}{s}$, where $f_S^*(s)$ is the Laplace transform of $f_S(t)$ and (21) becomes

$$\frac{\lambda_S}{\delta}\cdot\left.\frac{f_S^*(s)}{s}\right|_{s=\lambda_S/\delta}=f_S^*(s)|_{s=\lambda_S/\delta}. \tag{22}$$

APPENDIX C
PROOF OF THEOREM 1

To obtain the optimal SMR threshold value (δ^*) in ARO, we define a difference function as

$$\Delta C_T=C_T^{LBU}-C_T^{RBU}. \tag{23}$$

At the optimal SMR threshold, the condition $\Delta C_T=0$ is met. To find δ^* satisfying this condition, we substitute the terms C_T^{LBU} and C_T^{RBU} from (3), (4), (5), and (6). Then, C_T^{LBU} and C_T^{RBU} can be shown as functions of μ_S:

$$C_T^{RBU}(\mu_S)=\frac{\mu_S}{\sqrt{N}}\cdot T_{BU}\cdot BU_{CN}+\lambda_S\cdot T_{BU}\cdot E(S)\cdot C_{MAP-MN}, \tag{24}$$

$$C_T^{LBU}(\mu_S)=\mu_S\cdot T_{BU}\cdot BU_{CN}, \tag{25}$$

where N is the number of subnets in a MAP domain and μ_D is equal to $\mu_S/\sqrt{N}$ [29].

If μ_S becomes infinite, the following is met:

$$\lim_{\mu_S\to\infty}\frac{C_T^{RBU}}{C_T^{LBU}}=\frac{1}{\sqrt{N}}<1. \tag{26}$$

On the other hand, when μ_S approaches 0, $\lambda_S\cdot T_{BU}\cdot E(S)\cdot C_{MAP-MN}$ is greater than 1 by assumption.

$$\lim_{\mu_S\to 0}\frac{C_T^{RBU}}{C_T^{LBU}}=\lambda_S\cdot T_{BU}\cdot E(S)\cdot C_{MAP-MN}>1. \tag{27}$$

Therefore, by (26) and (27), there exists an optimal threshold which can be derived from the condition

$$\begin{aligned}\Delta C_T(\delta^*)=&\frac{\mu_S}{\sqrt{N}}\cdot T_{BU}\cdot BU_{CN}\\ &+\lambda_S\cdot T_{BU}\cdot E(S)\cdot C_{MAP-MN}\\ &-\mu_S\cdot T_{BU}\cdot BU_{CN}=0.\end{aligned} \tag{28}$$

After some manipulations, we have (29) and, hence, Theorem 1 follows.

$$\delta^*=\frac{\lambda_S}{\mu_S}=\frac{(d_{MAP-MN}\cdot(L_T+L_{BU})+d_{CN-MAP}\cdot L_{BU}+d_{CN-MN}\cdot L_{BACK})\cdot(1-1/\sqrt{N})}{E(S)\cdot L_T\cdot d_{MAP-MN}}. \tag{29}$$

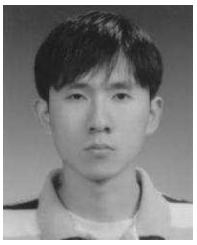
ACKNOWLEDGMENTS

This work was supported in part by a Strategic Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada under Grant No. STPGP 257682 and in part by the Korea Research Foundation Grant No. M01-2005-000-10073-0.

REFERENCES

[1] I. Akyildiz, J. Xie, and S. Mohanty, "A Survey on Mobility Management in Next Generation All-IP Based Wireless Systems," *IEEE Wireless Comm. Magazine*, vol. 11, no. 4, pp. 16-28, Aug. 2004.
[2] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," IETF RFC 3775, June 2003.
[3] H. Soliman, C. Castelluccia, K.E. Malki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)," IETF RFC 4140, Aug. 2005.
[4] A. Campbell, J. Gomez, S. Kim, A. Valko, C. Wan, and Z. Turanyi, "Design, Implementation, and Evaluation of Cellular IP," *IEEE Personal Comm. Magazine*, vol. 7, no. 4, pp. 42-49, Aug. 2000.
[5] R. Ramjee, T. Porta, S. Thuel, K. Varadhan, and S. Wang, "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks," *IEEE/ACM Trans. Networking*, vol. 6, no. 2, pp. 396-410, June 2002.
[6] A. Misra, S. Das, A. Dutta, A. McAuley, and S.K. Das, "IDMP-Based Fast Handoffs and Paging in IP-Based 4G Mobile Networks," *IEEE Comm. Magazine*, vol. 40, no. 3, pp. 138-145, Mar. 2002.
[7] Y. Lee and I. Akyildiz, "A New Scheme for Reducing Link and Signaling Costs in Mobile IP," *IEEE Trans. Computers*, vol. 52, no. 6, pp. 706-712, June 2003.
[8] R. Hsieh and A. Seneviratne, "A Comparison of Mechanisms for Improving Mobile IP Handoff Latency for End-to-End TCP," *Proc. MobiCom*, Sept. 2003.
[9] R. Koodli, "Fast Handovers for Mobile IPv6," IETF RFC 4068, July 2005.
[10] X. Zhang, J. Castellanos, and A. Capbell, "P-MIP: Paging Extensions for Mobile IP," *ACM Mobile Networks and Applications*, vol. 7, no. 2, pp. 127-141, Mar. 2002.

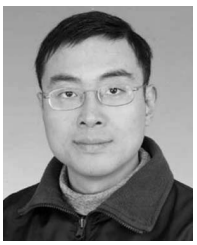
[11] Y. Lin, A. Pang, and H. Rao, "Impact of Mobility on Mobile Telecommunications Networks," *Wiley Wireless Comm. and Mobile Computing,* vol. 5, no. 8, pp. 713-732, Nov. 2005.
[12] Y. Chung, D. Sung, and A. Aghvami, "Steady State Analysis of P-MIP Mobility Management," *IEEE Comm. Letters,* vol. 7, no. 6, pp. 278-280, June 2003.
[13] Y. Wong, T. Wang, and Y. Lin, "Effects of Route Optimization on Out-of-Order Packet Delivery in Mobile IP Networks," *Information Sciences Informatics and Computer Science: An Int'l J.,* vol. 169, nos. 3-4, pp. 263-278, Feb. 2005.
[14] S. Pack, M. Nam, T. Kwon, and Y. Choi, "An Adaptive Mobility Anchor Point Selection Scheme in Hierarchical Mobile IPv6 Networks," *Elsevier Computer Comm.,* vol. 29, no. 16, pp. 3066-3078, Oct. 2006.
[15] A. Kortebi, L. Muscariello, S. Oueslati, and J. Roberts, "Minimizing the Overhead in Implementing Flow-Aware Networking," *Proc. ACM Symp. Architectures for Networking and Comm. Systems (ANCS '05),* Oct. 2005.
[16] S. Lo, G. Lee, W. Chen, and J. Liu, "Architecture for Mobility and QoS Support in All-IP Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 22, no. 4, pp. 691-705, May 2004.
[17] L. Kleinrock, *Queuing Systems Volume 1: Theory.* John Wiley & Sons, 1975.
[18] Y. Xiao, Y. Pan, and J. Li, "Design and Analysis of Location Management for 3G Cellular Networks," *IEEE Trans. Parallel and Distributed Systems,* vol. 15, no. 4, pp. 339-349, Apr. 2004.
[19] A. Stephane and A. Aghvami, "Fast Handover Schemes for Future Wireless IP Networks: A Proposal and Analysis," *Proc. IEEE Vehicular Technoloy Conf. (VTC '01),* 2001.
[20] N. Banerjee, S. Das, S. Dawkins, and J. Pathak, "Mobility Support in Wireless Internet," *IEEE Wireless Comm. Magazine,* vol. 10, no. 5, pp. 54-61, Oct. 2003.
[21] D. Staehle, K. Leibnitz, and K. Tsipotis, "QoS of Internet Access with GPRS," *ACM Wireless Networks,* vol. 9, no. 3, pp. 213-222, May 2003.
[22] A. Downey, "The Structural Cause of File Size Distributions," *Proc. IEEE MASCOT '01,* Aug. 2001.
[23] C. Perkins, "IP Mobility Support in IPv4," IETF RFC 3344, Aug. 2002.
[24] S. Rajagopalan and B. Badrinath, "Adaptive Location Management for Mobile IP," *Proc. MobiCom,* Nov. 1995.
[25] C. Perkins and D. Johnson, "Route Optimization in Mobile IP," IETF Internet draft, Nov. 2000.
[26] R. Yates, C. Rose, S. Rajagopalan, and B. Badrinath, "Analysis of a Mobile-Assisted Adaptive Location Management Strategy," *ACM Mobile Networks and Applications,* vol. 1, no. 2, pp. 105-112, Oct. 1996.
[27] S. Hwang, B. Lee, Y. Han, and C. Hwang, "An Adaptive Hierarchical Mobile IPv6 Using Mobility Profile," *Wiley Wireless Comm. and Mobile Computing,* vol. 4, no. 2, pp. 233-245, Mar. 2004.
[28] S. Pack, T. Kwon, and Y. Choi, "Adaptive Local Route Optimization in Hierarchical Mobile IPv6 Networks," *Proc. IEEE Wireless Comm. and Networking Conf. (WCNC '05),* Mar. 2005.
[29] K. Wang and J. Huey, "A Cost Effective Distributed Location Management Strategy for Wireless Networks," *ACM Wireless Networks,* vol. 5, no. 4, pp. 287-297, July 1999.



Sangheon Pack received the BS (2000, magna cum laude) and PhD (2005) degrees from Seoul National University, both in computer engineering. Since March 2007, he has been an assistant professor in the School of Electrical Engineering, Korea University, Korea. From July 2006 to February 2007, he was a postdoctoral fellow at Seoul National University. From 2005 to 2006, he was a postdoctoral fellow in the Broadband Communications Research (BBCR) Group at University of Waterloo, Canada. From 2002 to 2005, he was a recipient of the Korea Foundation for Advanced Studies (KFAS) Computer Science and Information Technology Scholarship. He has also been a member of Samsung Frontier Membership (SFM) since 1999. He received a student travel grant award for the IFIP Personal Wireless Conference (PWC) 2003. He was a visiting researcher at Fraunhofer FOKUS, Germany, in 2003. His research interests include mobility management, multimedia transmission, and QoS provision issues in next-generation wireless/mobile networks. He is a member of the ACM and the IEEE.

Xuemin (Sherman) Shen received the BSc (1982) degree from Dalian Maritime University, China, and the MSc (1987) and PhD degrees (1990) from Rutgers University, New Jersey, all in electrical engineering. Dr. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where he is a professor and the associate chair for graduate studies. His research focuses on mobility and resource management in interconnected wireless/wireline networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a coauthor of two books and has published more than 200 papers and book chapters in wireless communications and networks, control, and filtering. Dr. Shen serves as the technical program committee chair for IEEE Globecom '07, the IEEE WCNC '07 Network Symposium, Qshine '05, IEEE Broadnet '05, WirelessCom '05, IFIP Networking '05, ISPAN '04, and the IEEE Globecom '03 Symposium on Next Generation Networks and Internet. He also serves as an associate editor for the *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, *ACM Wireless Networks*, *Computer Networks*, *Wireless Communications*, *Mobile Computing* (Wiley), etc., and as a guest editor for the *IEEE Journal on Selected Areas in Communications*, *IEEE Wireless Communications*, and *IEEE Communications Magazine*. He received Premier's Research Excellence Award (PREA) from the Province of Ontario, Canada, for demonstrated excellence of scientific and academic contributions in 2003 and the Outstanding Performance Award from the University of Waterloo for outstanding contributions in teaching, scholarship, and service in 2002. He is a registered professional engineer of Ontario, Canada. He is a senior member of the IEEE and the IEEE Computer Society.

Jon W. Mark received the PhD degree in electrical engineering from McMaster University, Canada, in 1970. Upon graduation, he joined the Department of Electrical Engineering (now Electrical and Computer Engineering) at the University of Waterloo, became a full professor in 1978, and served as Department Chairman from July 1984 to June 1990. In 1996, he established the Centre for Wireless Communications (CWC) at the University of Waterloo and has since been serving as the founding director. He was on sabbatical leave at the IBM Thomas Watson Research Center, Yorktown Heights, New York, as a visiting research scientist (1976-1977); at AT&T Bell Laboratories, Murray Hill, New Jersey, as a resident consultant (1982-1983); at the Laboratoire MASI, Université Pierre et Marie Curie, Paris, France, as an invited professor (1990-1991); and at the Department of Electrical Engineering, National University of Singapore, as a visiting professor (1994-1995). His current research interests are in wireless communications and wireless/wireline interworking, particularly in the areas of resource management, mobility management, and end-to-end information delivery with QoS provisioning. Dr. Mark is a coauthor of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). He is a life fellow of the IEEE and has served as a member of a number of editorial boards, including the *IEEE Transactions on Communications*, and *ACM/Baltzer Wireless Networks*, *Telecommunication Systems*, He was a member of the Inter-Society Steering Committee of the *IEEE/ACM Transactions on Networking* from 1992-2003 and a member of the IEEE COMSOC Awards Committee during the period 1995-1998.



Jianping Pan received the BS and PhD degrees in computer science from Southeast University, Nanjing, China, in 1994 and 1998, respectively. From 1999 to 2001, he was a postdoctoral fellow and then a research associate at the University of Waterloo, Ontario, Canada. From 2001 to 2005, he was a member of the research staff at Fujitsu Labs and then a research scientist at NTT MCL in Silicon Valley, California. He is currently an assistant professor of computer science at the University of Victoria, British Columbia, Canada. His area of specialization is distributed systems and computer networks, and his recent research interests include protocols for advanced networking, performance analysis of networked systems, and applied network security. He is a member of the ACM, the IEEE, and the IEEE Computer Society.