Quality-of-Service Mechanisms in All-IP Wireless Access Networks

Bongkyo Moon and A. Hamid Aghvami, *Senior Member, IEEE*

***Abstract*—In this paper, we focus on resource reservation protocol (RSVP)-based quality-of-service (QoS) provisioning schemes under Internet protocol (IP) micromobility. We consider QoS provisioning mechanisms for on-going RSVP flows during handoff. First, the rerouting of RSVP branch path at a crossover router (CR) at every handoff event can minimize resource reservation delays and signaling overheads, and in turn the handoff service degradation can be minimized. We show that RSVP branch path rerouting scheme could give a good tradeoff between the resource reservation cost and the link usage. Second, the new RSVP reservation can be made along the branch path toward the CR via a new base station in advance, while the existing reservation path is maintained, and in turn the on-going flow can be kept with the guaranteed QoS. We also show that seamless switching of RSVP branch path could provide the QoS guarantee by adaptively adjusting the pilot signal threshold values. Third, during RSVP resource reservation over wireless link, dynamic resource allocation scheme is used to give a statistical guarantee on the handoff success of on-going flows. We finally obtain the forced termination probability of guaranteed service flows, the average system time of best effort flows by using a transition rate matrix approach.**

***Index Terms*—Advance reservation, dynamic resource allocation, path rerouting, resource reservation protocol (RSVP), seamless quality-of-service (QoS).**

I. INTRODUCTION

IN Internet protocol (IP)-based wireless networks (RANs), base stations (BSs) connect radio system to an IP radio access network (RAN). These BSs [or access routers (AR)] use IP protocols for data transport and signaling in either wireless local area networks (WLANs) or fourth-generation (4G) networks, but their coverage areas (or cells) may be arranged in any arbitrary topology. Recently, as it becomes much easier to access the Internet from a mobile host through wireless networks, the mobile users will be demanding the same real-time service available to fixed hosts. Typically, real-time applications impose very strict quality-of-service (QoS) requirements on the underlying transmission network. IPv6, next-generation Internet protocol, allows DiffServ-style QoS to be applied with a flow-ID field in its IP header. Nevertheless, some applications such as streaming audio and video would be much better served under the IntServ since they have a relatively constant bandwidth requirement for a known period of time; IntServ and resource reservation protocol (RSVP) allow Internet real-time applications to reserve resources before they start transmitting data.

However, there are several problems in considering conventional RSVP under host mobility scenario [1]. First, in standard RSVP operation, since each data source issues PATH messages periodically to automatically reserve resources along a new path after a routing change occurs, a mobile receiver must wait for a PATH message at its new location before it can send a RESV message back along the new path to the source for reservation of resources. This problem can be avoided by the mechanism that the sender triggers the transmission of a PATH message after receiving a binding update message from the mobile node (MN). Second, due to the long resource reservation delay during reestablishment of a flow after handoff under RSVP, service disruptions could occur in providing real-time services. Third, there is also no guarantee that the same level of resources will be available under a new point of attachment to which an MN moves. This implies that there may be service disruption with the mobility of the host. Consequently, the wireless links and host mobility will be the most critical point for QoS delivery and, thus, access networks without any QoS support will not be able to adequately support a wide range of services.

Until now, there have been a number of studies on supporting the QoS under mobile IP and RSVP [2]–[7]. In this paper, however, we focus on RSVP-based QoS provisioning schemes under IP micromobility protocol. We, thus, propose a mechanism for the QoS provisioning of on-going session during handoff under RSVP. First, the rerouting of RSVP branch path at a crossover router (CR) at every handoff event can minimize resource reservation delays and signaling overheads, and in turn the handoff service degradation can be minimized. Second, the new RSVP reservation can be made along the branch path toward the CR via a new BS in advance while the existing reservation path is maintained, and in turn the on-going flow can be kept with the guaranteed QoS. Third, during RSVP resource reservation over wireless link, dynamic resource allocation scheme can be used to give a statistical guarantee on the handoff success of on-going flows. Eventually, the on-going flow can be kept with the guaranteed QoS.

We describe RSVP issues for IP QoS in RANs in Section II. RSVP path management and seamless switching of RSVP branch path are presented in Sections III and IV, respectively. Dynamic resource allocation over wireless link is presented in Section V. In Section VI, examples and discussions are illustrated and discussed. Finally, conclusions are given in Section VI.

Manuscript received June 1, 2003; revised December 1, 2003. This work was supported in part by the Centre for Telecommunications Research (CTR), King's College, University of London, London, U.K.

The authors are with the Centre for Telecommunications Research (CTR), King's College, University of London, Strand, London WC2R 2LS, U.K. (e-mail: bong-kyo.moon@kcl.ac.uk; hamid.aghvami@kcl.ac.uk).

Digital Object Identifier 10.1109/JSAC.2004.826924

II. RSVP ISSUES FOR IP QoS IN RANs

In this section, we present an overview of IP micromobility protocol and its interaction with RSVP, and descriptions of QoS issues such as RSVP signaling loads and long reservation delay, seamless handoff of the flow with RSVP reservation path, and RSVP resource reservation over wireless link. The purpose of this discussion is to set the stage for the following sections.

A. IP Micromobility Protocol

Most IP micromobility protocols such as cellular-IP [19] and handoff-aware wireless access Internet infrastructure (HAWAII) [12] have hierarchical topology in which the network is normally organized in a tree with a single gateway node (GW) as a root [8]. These protocols were suggested as not replacements but enhancements to the mobile IP protocol and, thus, reduce the performance impact of mobility by hiding local migrations from home agents (HAs). That is, an MN's IP address known by an HA represents the GW's address that is common to a potentially large numbers of network access points. Hence, a correspondent node (CN) regards the GW's address as an MN's care-of-address. In order to ensure that packets arriving at the GW are forwarded to the appropriate access point and, thus, delivered to the MN's actual point of attachment, IP micromobility protocols maintain a routing database which maps host identifier, IP address, to actual location information, interface number. More specifically, each entry in a routing database contains a pointer to the next node toward the MN's actual point of attachment. Thus, when forwarding a downlink packet, intermediate nodes must read the original IP address in the packet and then find the corresponding entry in the database. They finally forward the packet to the next node by the interface pointer.

Mobile IP has been optimized mainly for macromobility and relatively slow-moving mobile hosts. This makes high signaling overhead under the frequent notification to an MN's HA, and thereby gives service disruption during handoff. However, IP micromobility protocols aim to handle local movement of mobile hosts without interaction with the mobile IP enabled Internet. This has the benefit of reducing delay and packet loss during handoff and eliminating registration between mobile hosts and possibly distant HAs when MNs remain inside their local coverage areas. Eliminating registration in this manner reduces the signaling load experienced by the core network in support of mobility.

B. Efficient RSVP Resource Reservation

Under IP micromobility protocols, the mobility of the end node will be hidden by the gateway and, thus, there will not be any RSVP signaling messages between a sender in Internet and a gateway in a wireless access network. Instead, RSVP resource reservation path needs to be repaired locally between the GW and an MN whenever an MN changes its access point within a domain. Moreover, RSVP state information needs to be modified because the forwarding mechanism does not depend on the normal IP routing but a routing scheme according to IP micromobility protocol. Typically, RSVP path setup needs to be performed very frequently when serving large number of mobile users (with RSVP sessions) moving very often between small cells. Thereby, this can result in increased reservation restoration latency and unnecessary control traffic. Fig. 1 shows an RSVP-enabled router diagram in IP micromobility protocols. In Section III, we propose an RSVP branch path rerouting scheme at a CR for giving a solution to this problem.

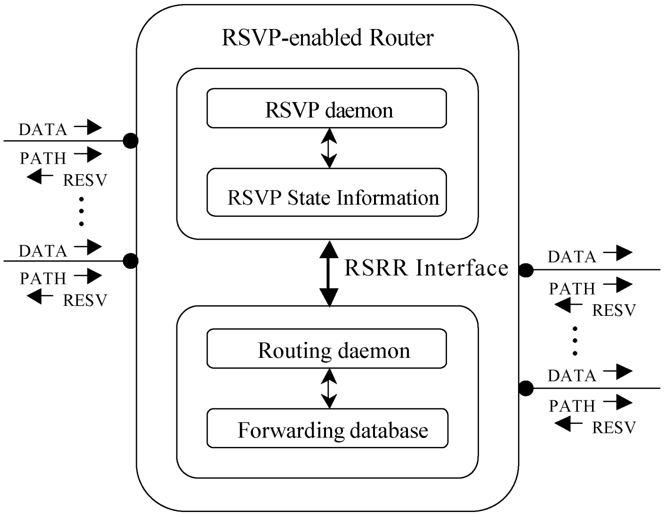

Fig. 1. RSVP-enabled router diagram in IP micromobility protocol.

C. Seamless Rerouting of RSVP Branch Path

Typically, handoff event can cause packet losses when processing of an admission request in a new BS is delayed. In this situation, some packets might be discarded, and thereby overall quality of the on-going flow might be degraded significantly. In the worst case, a delay in handoff may cause the connection to be dropped. Therefore, QoS signaling at handoff should be carried out very quickly. If the time required to restore a flow of traffic, after an MN receives the beacon that triggers a handoff, is very short, it might be possible to provide QoS guarantee to some real-time applications together with proper retransmission buffer size and well-measured beacon period. With RSVP path reservation, however, it may not possible to get enough short total handoff time in order to provide QoS guarantee. This may be caused by resource reservation delay during RSVP branch path reestablishment at every handoff. For example, when the amount of link bandwidth to be reserved for a real-time application is rather large, even short handoff time can cause a significant packet loss, and thereby on-going flow sessions may be disrupted. In Section IV, therefore, we propose a seamless switching of RSVP branch path for soft handoff.

D. Resource Reservation Over Wireless Link

In wireless access networks, channel capacity is typically limited and expensive compared to that in wired network, therefore, the resource reservation scheme must facilitate the network utilization to the highest degrees possible. However, the amount of the resource reservation required by RSVP flows in a cell can differ according to their QoS needs and can be time-variant due to mobility pattern of mobile users. This makes it difficult to reserve the same level of resource (over wireless link) in a new cell during handoff. Meanwhile, resource allocation over wireless link should be as simple as possible in order to reserve resource immediately on processing RSVP RESV message. In Section V, thus, a dynamic resource allocation scheme is proposed to give a statistical guarantee on handoff success of on-going RSVP flows.

III. EFFICIENT RSVP PATH MANAGEMENT UNDER IP MICROMOBILITY

Frequent handoff of RSVP flows between small cells results in increased reservation restoration latency and unnecessary control traffic. In order to reduce RSVP reservation failure due to network congestion, as well as RSVP reservation latency due to control signal overheads, shorter hops and circuit reuse should be taken into account during RSVP path rerouting at handoff. In particular, for the flows with excessive capacity demands such as streaming audio and video, fewer hops can guarantee overall high link utilization.

As a result, RSVP resource reservation signaling needs to be managed efficiently in only branch part of the tree, instead of traveling the entire path between the gateway node and the MN. That is, RSVP signaling messages need to be exchanged between the closest network node common to the former RSVP path and an MN since the significant portion of the overall handoff time under RSVP depends on the partial path connection setup time. Therefore, special mechanisms are needed to minimize the resource reservation delay and the packet loss resulting from handoffs under RSVP. We here propose an RSVP path rerouting scheme in order to be able to efficiently prevent this undesirable situation. That is, by rerouting RSVP branch path at a CR at every handoff event, handoff resource reservation delays and signaling overheads can be minimized, which in turn minimizes the handoff service degradation.

A. Determining the Location of a CR

The RSVP connection setup latency at handoff mainly depends on the number of hops required in rerouting the RSVP branch path, which is CR dependent. In general, even though path rerouting has the potential of creating an optimal path, the time taken to determine the location of a CR and the time taken to reroute an RSVP branch path may exceed delay bounds for the QoS provisioning in mesh topology [14]. However, due to the characteristics of tree topology in All-IP wireless access networks which adopt IP micromobility protocols, the time taken to discover a CR becomes enough short to meet QoS delay bounds. The process to locate a CR can be described as follows.

- When an MN sends a route update message along the new path toward the GW, one of interior routers on the new path detects the changes in forwarding entries. By seeing the same IP address (of the source or the destination) in the new forwarding entry with the interface number different from the one stored in the old forwarding entry, the interior router then decides the location of the CR.
- The interior router, which is a CR, then sends its RSVP daemon a path change notification (PCN) message using routing support for resource reservations (RSRR) interface [15], which is a specification for communication between RSVP and routing daemons. In other words, RSVP resource reservation is accomplished by using the routing interface that allows RSVP to access to forwarding database entries when those entries change.

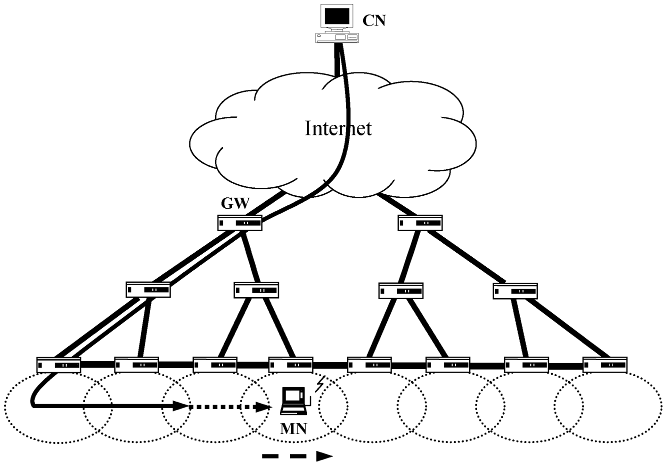

Fig. 2. HMRSVP with PF.

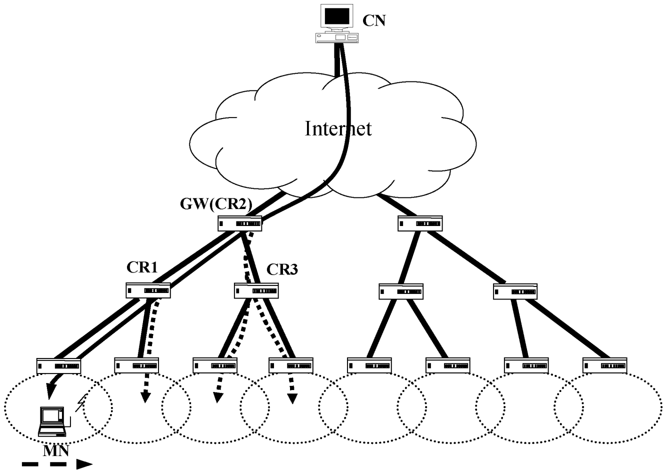

Fig. 3. RSVP branch path rerouting at CR in IP micromobility scenario.

B. Rerouting of RSVP Branch Path at a CR

HAWAII [12] includes the interactions with RSVP in domain root router (e.g., GW), while an MN remains within a domain. That is, the RSVP path between the GW and an MN is restored newly whenever an MN changes its access point within a domain. This may cause unnecessary control traffic overhead and rather long reservation delay. Hierarchical mobile (HM) RSVP [13] integrates RSVP with mobile IP regional registration and makes advance resource reservations only when an interregion movement may possibly happen. The pointer forwarding (PF) scheme makes advance resource reservations only a forward one-step path from an MN along the forwarding pointer chains [13]. This scheme can, thus, reduce the length of links to reserve newly at each handoff, but may cause excessive long RSVP path due to triangle routing or loop routing during several handoffs (see Fig. 2). In this situation, interior nodes located on this unnecessarily long path may become overloaded, since they suddenly might need to support more traffic than they have capacity for. These severe congestion situations will severely affect the RSVP reservation for the real-time traffic passing through such nodes.

Fig. 3 shows an example of the RSVP branch path rerouting at CR in IP micromobility scenario. The rerouting of branch-path

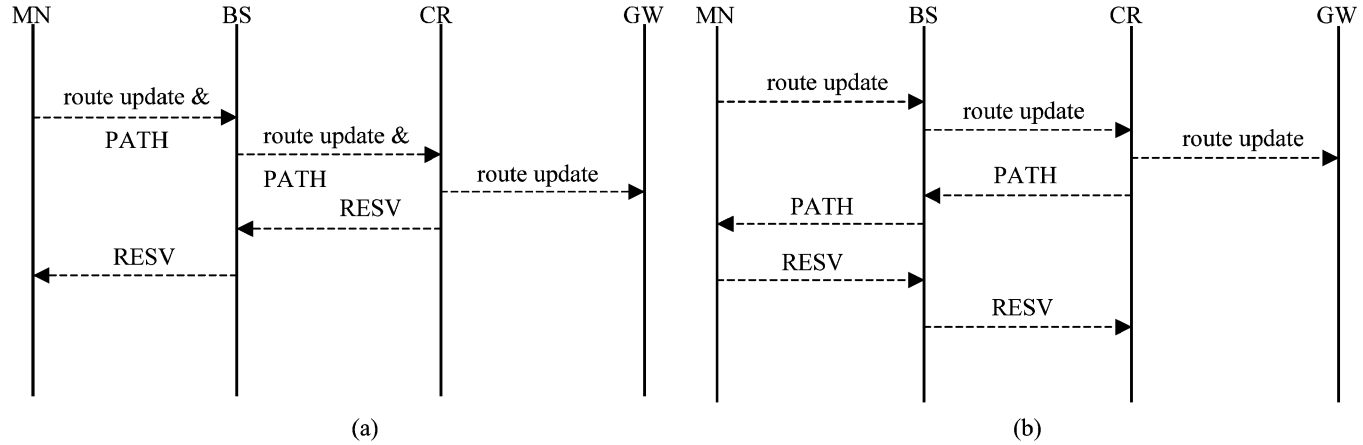

Fig. 4. RSVP signaling messages for path rerouting at CR during handoffs. (a) MN is a sender. (b) MN is a receiver.

at CR can guarantee the minimum path change and the shortest path from the GW to an MN: in the worst case a CR is the GW. In this section, we describe the branch-path rerouting scheme for efficient RSVP path management during handoff. In this scheme, a CR can send a PATH (or RESV) message immediately after updating its forwarding table: only RSVP connection between a CR and an MN is reestablished. Two cases of an MN can be described as follows.

- In the case an MN is a sender, when the RSVP daemon on the CR receives RSVP PATH message, it sees the same PATH message arriving with a previous hop address that differs from the one stored in the original PATH state and, thus, immediately sends an RSVP RESV message to the MN [see Fig. 4(a)].
- In the case an MN is a receiver, the RSVP daemon on the CR can trigger an RSVP PATH message immediately when detecting any changes to the stored PATH state or receiving a PCN message from the underlying routing daemon. When an MN receives this PATH message, it generates an RSVP RESV message after a receiver processing delay on RSVP daemon [see Fig. 4(b)].

C. Performance Evaluation

For analytical simplicity, we assume a full binary tree with a GW as a root for analysis model (see Fig. 3), and an MN moves among only the leaf nodes (terminal BSs) in tree structure. Hence, the depth k of a full binary tree with N nodes is $\lfloor \log_2 N \rfloor + 1$, that is, a full binary tree of depth k has $2^k - 1$ nodes, $k \geq 1$, and the number of nodes (BSs) at level k is 2^{k-1}. We assume that the tree topology is homogeneous, i.e., tree structures in all micromobility regions are topologically identical. Therefore, we can model the mobility behaviors by focusing on only one tree for each micromobility level and together by including the infinite number of the same tree structures for macromobility level. We also assume that macromobility rate depends on the micromobility rate and the number of BSs in micromobility region: the probability an MN moves within a micromobility region is much larger than that within macromobility region.

1) Modeling Mobility Behavior of an MN: For simplicity, we model the mobility behaviors of an MN among leaf nodes in the lowest level of the tree hierarchy as Markov chain, as shown in Fig. 5. To describe this Markov chain, we define the following parameters.

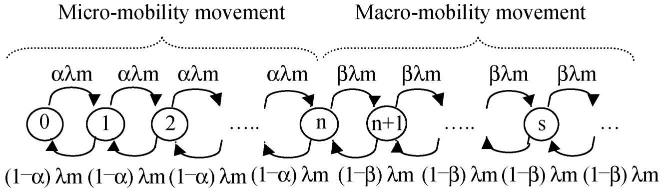

Fig. 5 State-transition-rate diagram for the mobility behavior of an MN in one direction.

i	State that an MN stays in a BS under micromobility, where state (BS) i is numbered sequentially from left-end to right-end among leaf nodes of a tree or an MN stays in a tree region under macromobility, where state (tree region) i is numbered sequentially under the assumption that infinite tree regions exist.
$P(i)$	Steady-state probability that an MN stays in a BS (or a tree region) with the state i.
α	Mean probability that an MN may move right to its neighboring BS and perform micromobility handoff.
β	Mean probability that an MN may move right to its neighboring region and perform macromobility handoff.
λ_m	Mobility rate of an MN.

We can obtain the balance equations from the Markov chain of Fig. 5. Therefore

$$\begin{aligned} P(i)\alpha\lambda_m &= P(i+1)(1-\alpha)\lambda_m, \quad 0 \leq i \leq n-1 \\ P(i)\beta\lambda_m &= P(i+1)(1-\beta)\lambda_m, \quad i \geq n \\ \sum_{i=0}^{\infty} P(i) &= 1 \end{aligned}$$

where

$$\beta = \alpha^{\frac{n+2}{2}}, \beta < \frac{1}{2}, \text{ and } n = 2^{k-1} - 1.$$

From the above, the balance equations, the closed form of each steady-state probability $P(i)$ for all i's can be derived as

$$P(i) = A^i\Phi(n), \quad 0 \le i \le n$$
$$P(i) = A^i B^{i-(n+1)}\Phi(n), \quad i \ge n+1$$

where
$$\Phi(n) = \frac{(A-1)(B-1)}{(A^{n+1}-1)(B-1) - A^n(A-1)}$$
$$A = \frac{\alpha}{1-\alpha} \quad \text{and} \quad B = \frac{\beta}{1-\beta}.$$

2) Analysis of Link Usage and Resource Reservation Cost: We define following parameters to compare the resource reservation cost and the link usage of the CR rerouting scheme with those of the GW rerouting scheme and the PF scheme. ρ represents the resource reservation cost for a link between two neighboring nodes in a tree. τ represents path optimization (or tunneling) cost for fixing the new global path by informing the HA and the CN of the new GW's address as the MN's new care-of-address after macromobility handoff.

In both CR and GW rerouting schemes, the mean length of the path required for holding an existing RSVP connection in a tree is equal to $(k-1)$ since rerouting scheme is based on the shortest path to an MN from the GW node. In the PF scheme [13], meanwhile, since an existing RSVP connection can be kept over more than two trees during its lifetime (see Fig. 2), the mean path length of forwarding pointer chains over trees can be given by

$$(k-1) + \sum_{i=0}^{n} iP(i) + (n+1)\sum_{i=n+1}^{\infty}(i-n+1)P(i)$$

where the first term $(k-1)$ is the number of links to the leaf node from the GW in a depth k tree, the second term is the mean path length forwarded by an MN from the left-end node (BS) among leaf nodes in a depth k tree, and the third term is the mean path length within infinite numbers of depth k trees under macromobility.

On the other hand, in the PF scheme, the mean reservation cost for the path rerouted (added or removed) newly at every handoff is equal to a constant ρ since an MN forward (or backward) by one-step link whenever performing a handoff. However, the number of links rerouted newly by a CR in RSVP reservation path at every handoff depends on the current BS (state i) and MNs moving direction (i.e. forward or backward) in a depth k tree region. In a depth k tree topology, L the number of rerouted RSVP path links at handoff varies from 1 to $k-1$. That is, when an MN moves between the nodes having same parents in a tree, L is only 1, but L becomes 2 when moving between the nodes having only same grandparents in a tree. In a worst case, L is $k-1$ when moving between the nodes having only same root in a tree. Eventually, the same number of links as L is rerouted at every 2^L interval node states whenever an MN moves right with probability α at state i or moves left with probability $(1-\alpha)$ at state $i+1$ among leaf nodes at handoff. Thus, the mean reservation cost for the RSVP branch path rerouted newly by a CR at every handoff can be given by

$$\rho\sum_{L=1}^{k-1}\sum_{i\in S_L} L\left(\alpha P(i) + (1-\alpha)P(i+1)\right) + \tau\sum_{i=n+1}^{\infty}(i-n)P(i)$$

where S_L defines a set of all the node states obtained from increasing by 2^L from $(2^{L-1}-1)$th node state to nth node state when L is the number of links rerouted by a CR in a tree. The second term in the above expression is the mean of extra cost due to macromobility handoffs during the lifetime of an RSVP connection. This is from the fact that RSVP branch-path rerouting schemes by both GW and CR have to pay the extra cost such as path optimization (or tunneling) for fixing the path between the CN and a new GW. In the rerouting scheme by GW, $(k-1)$ links are rerouted at every handoff. Hence, the mean reservation cost for the path rerouted newly by a GW in a tree at a handoff can simply be given by

$$\rho(k-1) + \tau\sum_{i=n+1}^{\infty}(i-n)P(i).$$

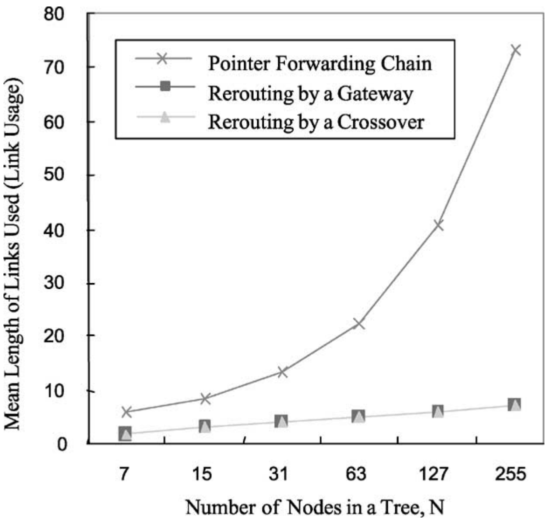

Fig. 6. Mean length of links used to hold an RSVP connection.

D. Examples and Discussions

For an RSVP path rerouting scheme at every handoff event, we assume a full binary tree with a GW as a root. Fig. 6 represents the performance results of the mean length of links required to hold an RSVP connection for the three schemes. First, it shows that the mean path length of PF scheme is significantly longer than those of both CR and GW based path-rerouting schemes as the number of nodes in a tree increases. In the PF scheme, an RSVP connection is maintained over more links due to the unnecessarily long or nonoptimized path. Thus, when the number of users with RSVP sessions increases, the probability of network congestion may become higher or RSVP resource reservation failure may occur more frequently. In both CR and GW based path-rerouting schemes, however, the mean length of the links required to keep a RSVP connection is kept low.

Fig. 7 depicts the mean resource reservation costs for the paths rerouted newly at every handoff. The reservation cost in the PF scheme is same regardless of the number of nodes in a tree. The cost of GW rerouting scheme is linearly proportional to the number of nodes in a tree. However, the mean resource reservation cost for the CR rerouting scheme converges on the value 2 even though the number of nodes increases up to 255. This is because the only partial branch path between a CR and

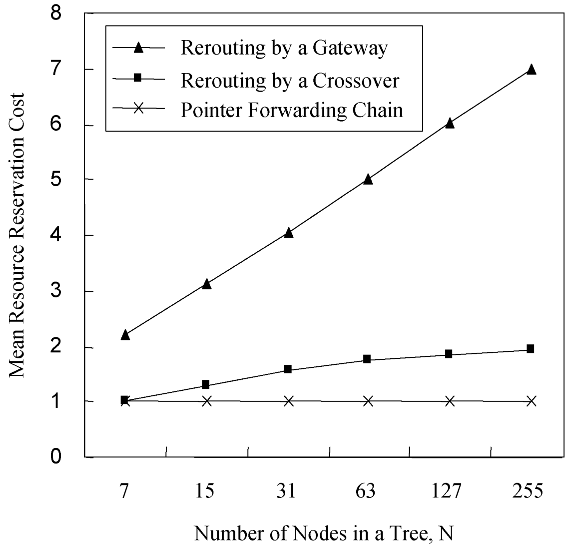

Fig. 7. Mean resource reservation costs for the paths rerouted newly at every handoff.

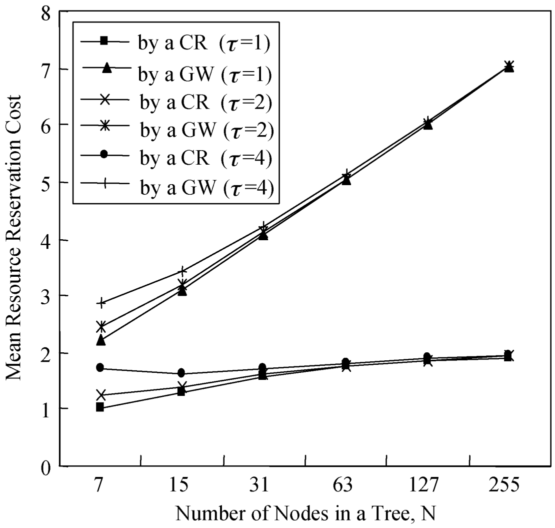

Fig. 8. Comparisons of the mean resource reservation costs in the both CR and GW based rerouting schemes.

an AR is rerouted, and the remaining path is reused without new reservation due to micromobility.

In Fig. 8, we observe that the mean resource reservation costs in the both CR and GW based rerouting schemes with the different values ($\tau = 1, 2, 4$) of the path optimization (or tunneling) cost (τ) due to macromobility. In Fig. 8, the variations in τ makes a little difference on the mean reservation cost when the number of BSs (or nodes) in a tree is not big. However, we can know that the path optimization does not make a notable effect on the mean reservation cost when the number of nodes in a tree is rather large. In other words, the mean path optimization cost due to macromobility handoffs is kept very low since β, the mean probability that an MN perform macromobility handoff, becomes reduced rapidly as the number of nodes in a tree approaches to a limit value, about 63 in the case of Fig. 8.

IV. SEAMLESS SWITCHING OF RSVP BRANCH PATH

In Section III, RSVP branch-path rerouting scheme during handoff may considerably reduce the reservation latency; nevertheless, there is no guarantee for the QoS of on-going RSVP session at handoff event: RSVP flow sessions may normally be disrupted until the new reservation is installed along the path via a new AR after handoff under RSVP. Therefore, we propose a seamless switching scheme of RSVP branch path for the on-going RSVP session not to be broken due to handoff.

A. Handoff

There are three types of handoffs in cellular systems: Intra frequency handoff, interfrequency handoff and intersystem handoff. Intrafrequency handoff is the movement event between BSs using the same carrier frequency, e.g., soft handoff in code-division multiple-access (CDMA) systems. Interfrequency handoff is the movement event between BSs using different carrier frequencies, e.g., hard handoff in time-division multiple-access (TDMA) and frequency-division multiple-access (FDMA) systems. Intersystem handoff is also the movement event between BSs employing different air interfaces: vertical handoff or interface switching, e.g., handoff in heterogeneous network environments such as GPRS/WLAN.

1) Semi-Soft Handoff: For hard handoff, handoff latency is the time that elapses between the handoff and the arrival of the first packet through the new route. For a flow with RSVP path reservation, this equals the path reestablishment time (and propagation delay) between the MN and the CR which is common to both old and new downlink routes. During this time, downlink packets may be lost since an MN has to tune its radio to the new BS (or AR) and wait for the first packet through the new RSVP reservation path. However, at hard handoff, the path via the old BS is not actually cleared until the soft-state timer expires. Therefore, there exists a period when both the old and new downlink routes are valid and packets are delivered through both BSs. This can provide probabilistic guarantees even though packet loss cannot be fully eliminated. This feature is called semi-soft handoff [19].

In order to minimize QoS disruption due to handoff latency, the RSVP path reservation via the new AR must be created before the actual handoff takes place. Therefore, when a mobile sender with an RSVP flow initiates handoff, it sends a route update packet, which includes a PATH message, to a new AR and immediately returns to tuning its radio to the old AR. When a mobile receiver initiates handoff, it also immediately returns to the old AR after sending a route update packet to a new AR. After a CR detects a route update packet, it sends a PATH message toward the new AR, which sends an RESV message back for reserving RSVP path. Thus, while the MN is still in contact with the old AR, RSVP reservation states can be installed along the route between the new AR and a CR.

Even though semi-soft handoff may make it possible for an MN to keep receiving (or sending) packets immediately after handoff, it does not provide fully seamless handoff of a flow with QoS requirements. Depending on the network topology and traffic conditions, the time to transmit packets from the CR to both old and new ARs may be different and the packet streams transmitted through the two ARs will typically not be synchronized at the MN [19].

2) Soft Handoff: Soft handoff is typically used in CDMA systems, but it can also be considered in TDMA/FDMA

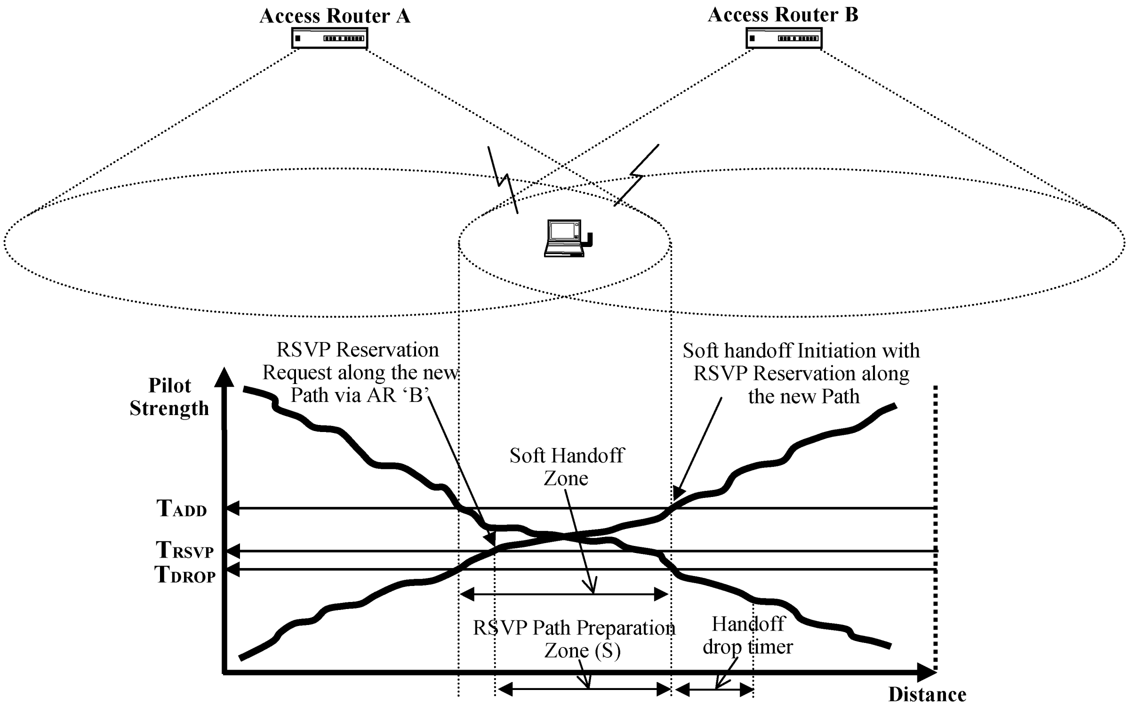

Fig. 9. RSVP path reservation by a pilot strength threshold.

systems that use different carrier frequencies if a mobile terminal has multiple network interfaces. However, it is not easy to implement soft handoff in TDMA/FDMA due to the problems such as synchronization of time slots and multiple receptions and transmissions at different frequencies [17]. Thus, TDMA/FDMA systems typically execute hard handoff, which restricts their utilization of macro diversity by implementing hysteresis. The major advantage of soft handoff in CDMA systems is macroscopic diversity, where a mobile station can be in contact with more than one BS at a time. That is, soft handoff allows a mobile station to communicate with multiple BSs. The soft handoff mechanisms in several CDMA proposals for 3G/4G services are very similar to those that are supported in IS-95, even though a substantial change is required for CDMA2000 1x EV-DV (Data/Voice) and W-CDMA HSDPA (high-speed downlink packet access) because they do not employ symmetric soft handoff [18].

The mobile assisted handoff (MAHO) scheme [17] can be considered, where the mobile station takes the signal strength measurements, and the BS makes decisions. A mobile station must make a conditional decision, which depends on the changes in signal strength from the two or more BSs involved. Actual handoff is performed after making sure that the signal from one BS is considerably stronger than those from the others. Thus, an MN eventually communicates with only one BS. In the interim period, a mobile station communicates with multiple BSs at the same time. Thereby, soft handoff can guarantee that a mobile station is indeed linked at all times to the BS from which it receives the strongest signal, whereas hard handoff cannot guarantee this.

B. Advance Reservation of RSVP Branch Path

Fig. 9 illustrates typical pilot strength variation as an MN moves from one BS area to another area in the proposed scheme. We assume that an MN moves from a BS coverage area to the adjacent BS area. If an MN finds a neighboring BS with a pilot signal higher than a predetermined threshold (T_{ADD}), then a new link to the BS is established while the existing link is maintained. If the pilot signal from either the old BS or the new BS drops below a predetermined threshold T_{DROP}, the corresponding link is released. In [16], authors proposed a mechanism for resource reservation on wireless link by a predetermined threshold in a BS coverage area.

However, we here extend this reservation mechanism to all the links within IP micromobility network in aid of RSVP and CR discovery scheme. Thus, we define a new parameter T_{RSVP}, which is a threshold of RSVP reservation requests. This value is less than T_{ADD}. If an MN finds any neighboring BS with pilot strength exceeding T_{RSVP}, it sends a RSVP reservation request message to the RSVP daemon on the associated BS (or AR). Afterwards, if the pilot strength drops and stays below T_{RSVP} during a predetermined period, the MN asks the RSVP daemon on the associated AR to release the RSVP branch path. On the other hand, if the pilot strength reaches T_{ADD}, then the MN with a RSVP session switches old branch path to the newly reserved branch path by initiating soft handoff and keeps the flow with the guaranteed QoS. In Fig. 9, the threshold T_{RSVP} is not an absolute value but is rather a relative value to T_{ADD}. When T_{ADD} is dynamically determined according to the current wireless-link status and network load condition, T_{RSVP} can also be adaptively modified.

C. Proxy Agent

When an MN initiates an RSVP reservation session, it sends an RSVP flow specification (called SPEC) message to a set of possible ARs (called MSPEC) to which the MN may visit in next step. Typically, SPEC message of a mobile sender contains *Sender_Tspec*, *Adspec* and the destination address of a flow,

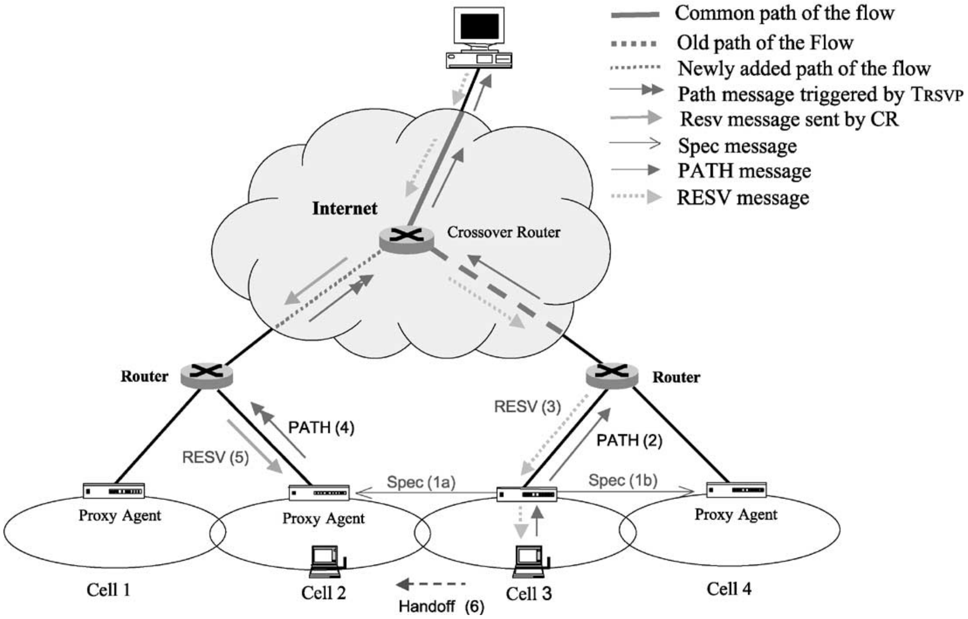

Fig. 10. RSVP branch path reservations by a proxy agent on behalf of an MN.

and SPEC message of a mobile receiver contains *FlowSpec* and *Flow ID*. Thus, RSVP session objects in a SPEC message are installed at the newly discovered AR's in the MSPEC. All ARs actually have proxy agents, which play a significant role in processing RSVP reservation request message triggered by $\mathrm{T_{RSVP}}$. Hence, an MN needs to discover these proxy agents in order to determine the IP addresses of them.

Consequently, when an MN makes a reservation via a current AR, it sends the SPEC of a flow to its remote proxy agents on ARs in its MSPEC. For making advance reservation before actual handoff, a remote proxy agent will setup, using the stored SPEC, the RSVP reservation (PATH or RESV) on behalf of an MN on receiving an RSVP reservation request message triggered by T_{RSVP}. More specifically, when an MN finds any neighboring BS (or AR) with pilot strength exceeding T_{RSVP}, it sends a RSVP reservation request message to the proxy agent on the associated AR. Then, the proxy agent establishes the branch path between the CR and the associated AR so that the path can be reserved in advance. Fig. 10 shows that a proxy agent sets up the RSVP reservation on behalf of an MN by using the SPEC.

In this scheme, the MN's MSPEC can be changed dynamically while the flow session is open. After rerouting RSVP branch path seamlessly during handoff, the SPECs stored in old remote proxy agents except a current proxy agent are cleared, and instead, SPEC message of the flow is sent to remote proxy agents in new MSPEC. No change is necessary in the RSVP message processing and forwarding rules except at the proxy agents and the MNs. Meanwhile, during handoff, RSVP message exchange can be omitted over one-hop wireless link since a proxy agent sets up the RSVP reservation on behalf of an MN that will handoff soon. This can make signaling overhead (over one-hop wireless link) reduced and can also allow radio resource to be managed with special allocation schemes.

D. Seamless Rerouting of RSVP Branch Path

Fig. 11 shows RSVP branch-path rerouting at CR and the switching of the branch path for soft handoff. This scheme is very effective in the case that cells, if hard handoff is considered, do not overlap enough to allow handoffs to complete before an MN loses connectivity with its previous BS. Seamless switching of RSVP branch path during handoff can be described as follows.

- When an MN is a sender, a proxy agent on a new AR sends a PATH message (using the stored SPEC) to the CR on receiving RSVP reservation request message before it performs an actual handoff. The CR then sends a RESV message to the new AR and, thus, new RSVP branch path reservation is accomplished while the existing reservation path is maintained.
- In the case of a mobile receiver, a proxy agent on a new AR first sends a route update message on receiving RSVP reservation request message. Then, a proxy agent responds with a RESV message after receiving the PATH message sent by the CR. Before an MN performs an actual handoff, the packets are still delivered through the path reserved previously via the old AR.
- During handoff, dynamic resource allocation is performed for an MN over one-hop wireless link. As soon as an MN sends (or receives) the packets through the branch path reserved freshly via the new AR, it can explicitly remove the reservations along the old branch path, and eventually tunes its radio to the new AR.
- During this RSVP branch path-switching period, the double wireless resource reservation is kept through both the old and the new branch path. However, this period is very short because the old branch path is released soon.

E. Performance Measures

The signal received by an MN can be taken as a random variable with lognormal distribution

$$f(x) = (1)/(\sqrt{2\pi\sigma})e^{(-(x-\mu)^2)/(2\alpha^2)}$$

where x is the signal level in decibels received by the MN from the neighboring BS, which has the strongest signal among all

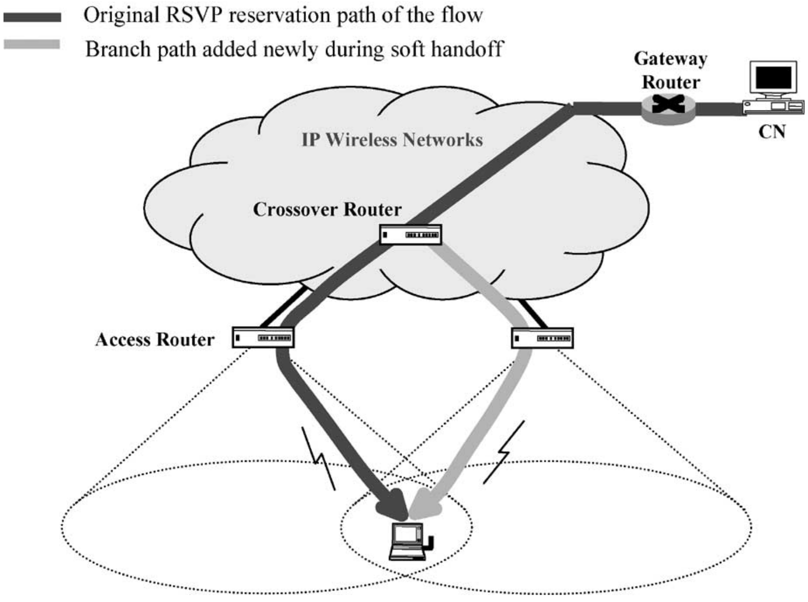

Fig. 11. Switching of RSVP branch path using soft handoff.

other neighboring BSs, assuming that Raleigh fading can be averaged out. μ is average received signal levels due to the path loss from the neighboring BS and can be given as $\mu(d) = K_1 - K_2 \log(d)$, where K_1 depends on the transmitted power in the BS, K_2 is typically a constant due to path loss and d is the distance from the neighboring BS. A handoff is required if the average signal level received from the current BS drops below that of the neighboring BS.

In our simulation, we assume that the threshold values of T_{ADD}, T_{DROP}, and T_{RSVP} are given within the range between -31.5 and 0 dB. Furthermore, we assume that an MN's velocity V is a constant and the power strength of each BS is identical. Thus, the distance that a MN moves within cell boundary area becomes proportional to the difference between two pilot signal strengths. Hence, from the above formula, "*RSVP path preparation zone*" denoted by S in Fig. 9 is equal to $10((K_1 - T_{\mathrm{RSVP}})/(K_2) - 10(K_1 - T_{\mathrm{ADD}})/(K_2))$, where K_1 depends on the transmitted power in the BS and K_2 is typically a constant due to path loss. As the MN moves to the neighboring BS, the time it takes for a pilot signal to be changed from T_{RSVP} to T_{ADD}, ΔT can thus be calculated by S/V. ΔT can actually be adjusted dynamically by changing T_{RSVP}, but too low T_{RSVP} value can cause excessive unnecessary RSVP reservations.

On the other hand, a significant portion of the overall handoff time under RSVP depends on the partial path connection setup time. RSVP connection setup latency is a function of the signaling protocol overhead, the number of hops, the signal propagation delay and the number of connection requests. Since it is not desirable to limit the number of connection requests made by real-time applications, the signaling protocol overhead and the number of hops required to setup the partial path during handoff become the target for reducing the connection setup time [11]. Now, we need to define the terminology for the signaling protocol overhead and the number of hops. The RSVP packet processing delays, which are ΔP and ΔR for PATH and RESV messages, respectively, on a router are defined as the difference between the time stamps at which the packet appears on the input and output links [21]. The setup time for RSVP branch path rerouting $\Delta\mathrm{RSVP}$ is defined as the delay between the time when the first PATH message is appeared on the MN's (or CR's) interface and the time when the first RESV message is detected on the same interface. The receiver processing delay ΔPR during this setup is the latency between the time the first PATH message is seen on the CR's (or MN's) interface and the time at which the first RESV message is seen on the same interface.

TABLE I
AVERAGE CONTROL PACKET LATENCY (IN ms) UNDER THREE TYPES OF LOADS IN THE NETWORK

Type of load	ΔP	ΔR	ΔPR
Low load	2.00 (0.20)	3.07 (0.30)	1.72 (0.17)
Medium load	3.90 (0.40)	5.41 (0.54)	2.14 (0.20)
High load	5.44 (0.54)	6.40 (0.64)	2.98 (0.30)

These delays depend on the load of routers and the number of the existing RSVP connections; we can assume that they go by normal distributions with the parameters in Table I, which is obtained from the results in [21]. It shows the mean latency of RSVP control packets in the three types of loads when a certain number of real-time sessions already exist. The numbers in parentheses show the standard deviations. The setup time, $\Delta\mathrm{RSVP}$ is, thus, given as $L \times \Delta P + L \times \Delta R + \Delta PR$, where L is the number of links (routers) between the CR and an MN. Total handoff delay ΔT_RSVP of a real-time flow with RSVP resource reservation is the time it takes for an MN to receive (or send) the first packet through the new RSVP reservation path when performing an actual handoff. Therefore, the time difference between ΔT and ΔT_RSVP then becomes an important parameter for measuring the QoS of a real-time flow with RSVP reservation path. When an MN moves into a new BS (or AR) area under RSVP, the packets transmitted on the RSVP path reserved freshly via the new BS may be lost if ΔT_RSVP is

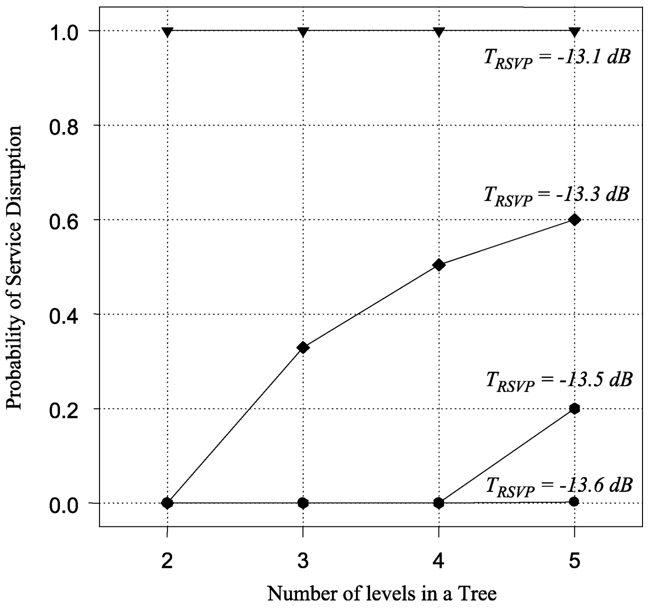

Fig. 12. Probability of service disruption with various T_{RSVP}.

greater than ΔT. Hence, the total handoff signaling process via new BS should be completed during the period ΔT.

When an MN is a sender, ΔT_RSVP is $\Delta RSVP$ if a PATH message is sent inserted in a route update message in order to reduce the signaling overhead. When an MN is a receiver, it has to wait for ΔU to send the route update packet to the CR, and then also wait for $\Delta RSVP$ for RSVP branch path rerouting. In order to keep the flow with the same QoS after a handoff, it has to wait for the additional propagation delay ΔG from the CR to an MN. That is, an MN has to perform an actual soft handoff after receiving the first packet through the new RSVP branch path. However, the propagation delay ΔG can be ignored under micromobility network. Thus, for a mobile receiver, the total handoff delay ΔT_RSVP is $\Delta U + \Delta RSVP$. Hence, the total number of lost packets of a flow with RSVP resource reservation at soft handoff is $B(\Delta T\_\mathrm{RSVP} - \Delta T)$, where B is the amount of link bandwidth reserved for the flow. Consequently, if the value of $\Delta T\_\mathrm{RSVP} - \Delta T$ is less than zero or equal to zero, service disruption does not occur. This means that this scheme can provide the QoS guarantee by dynamically adjusting the ΔT through threshold $\mathrm{T}_{\mathrm{RSVP}}$.

F. Examples and Discussions

For a seamless switching scheme of RSVP branch path, Fig. 12 shows the probability of service disruption with various threshold values of T_{RSVP}. For the simulation, the threshold values (T_{ADD}) is -13.0 dB, an MN's velocity V is 9.0 m/s, and constants $K1$ and $K2$ in received signal level as 35.3 and 40.9, respectively. In this figure, service disruption always occurs with $T_{\mathrm{RSVP}} = -13.1$ dB, but there is no service disruption after $T_{\mathrm{RSVP}} = -13.6$ dB. That is, as the pilot signal difference between T_{ADD} and T_{RSVP} is bigger, the time that an MN spends in the RSVP path preparation zone in cell boundary area becomes longer. On the other hand, the mean number of links required to set up branch path at every handoff depends on the number of levels in a tree. Hence, the smaller the number of levels in a tree is, the less RSVP branch path setup time is. In this figure, the reasonable number of levels in a tree is within five. Also, well-defined T_{RSVP} value results in getting enough time to set up new RSVP branch path when an MN moves to the neighboring BS.

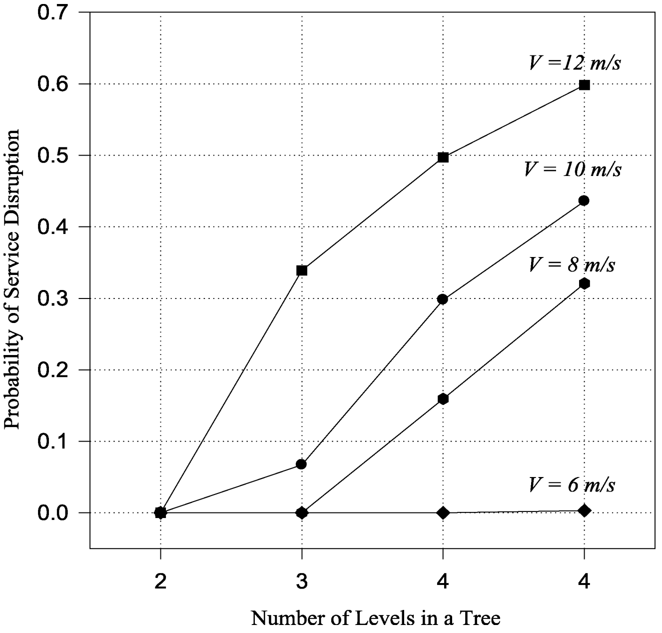

Fig. 13. Probability of service disruption with MNs various velocity.

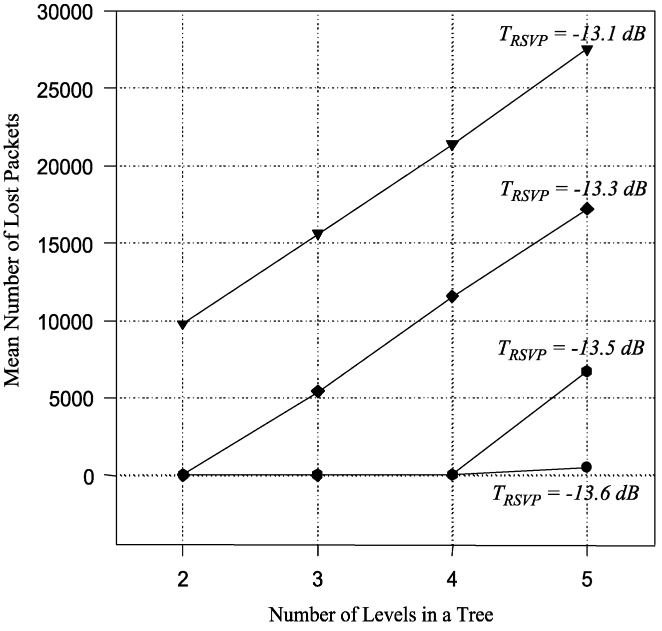

Fig. 14. Mean number of lost packets with various T_{RSVP} under $V = 9.0$ m/s, high load, and 1-Mb/s data rate.

Fig. 13 shows the probability of service disruption with an MN's various velocity under $T_{\mathrm{RSVP}} = -13.4$ dB. In this figure, the faster MN's velocity is, the shorter the time that an MN spends in cell boundary area is. Hence, seamless switching of RSVP branch path can be obtained when the velocity is less than or equal to 6 m/s. Fig. 14 shows the mean number of lost packets with various threshold values of T_{RSVP} under $V = 9.0$ m/s, high load of routers and application with 1 Mb/s reserved bandwidth. In this figure, when T_{RSVP} is -13.5 dB and the number

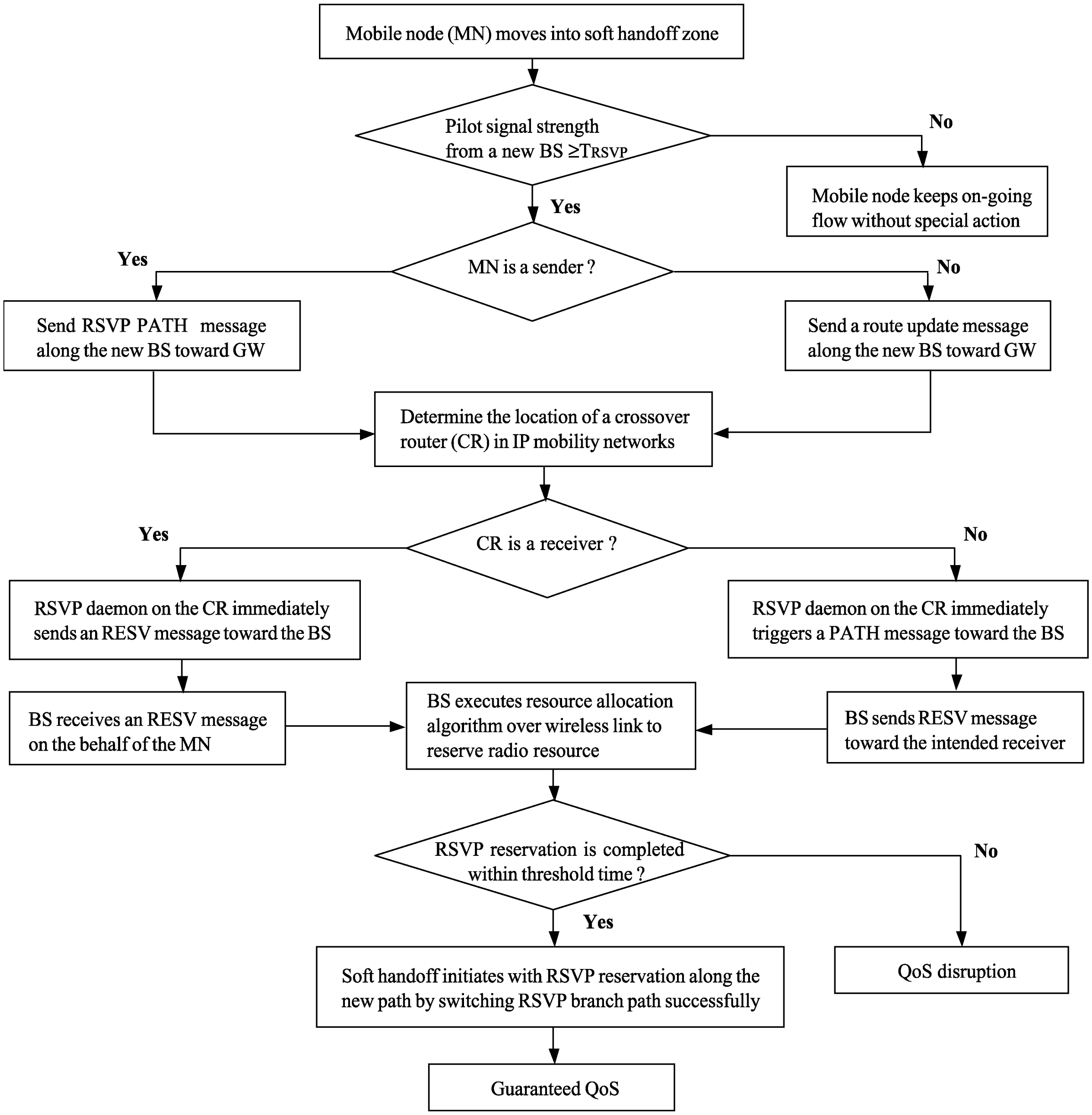

Fig. 15. QoS mechanisms for on-going flow under RSVP and IP micromobility.

of levels in a tree is less than 4, the on-going RSVP session can obtain seamless service despite the branch path rerouting at handoff. We can, thus, control the mean packet loss time by adjusting the threshold T_{RSVP} in addition to the number of levels in a tree. In this case, the total number of lost packets of a flow with RSVP resource reservation depends on the amount of link bandwidth reserved for the flow.

V. DYNAMIC RESOURCE ALLOCATION OVER WIRELESS LINK

Fig. 15 shows overall QoS mechanisms for on-going flow during handoff under RSVP and IP mobility. If an MN enters soft handoff zone, it checks the pilot signal strength from a new BS. If the pilot signal is greater than or equal to the threshold value T_{RSVP}, it attempts to reserve resources along the path via a new AR (or BS) in advance in order to keep an on-going RSVP flow with guaranteed QoS. Hence, it is necessary to find a CR for determining an intended receiver or an intended sender in RSVP branch path rerouting process.

When an MN moves into a neighboring AR coverage area at soft handoff, it attempts to reserve resources in advance in order to keep an on-going RSVP flow with guaranteed QoS. When an MN is a sender, a new AR intercepts an RESV message sent by the intended receiver (CR), and in turn completes resource allocation over wireless link. Thus, an end-to-end reservation is successfully established. When an MN is a receiver, a new AR (on behalf of an MN) responds with a RESV message after receiving a PATH message. Next, a new AR also has to perform the resource allocation over wireless kink to aid RSVP reservation process. During this resource reservation phase, the packets are still delivered through the path reserved previously via the old AR.

Meanwhile, there exist three possible scenarios in radio resource allocation. First, radio resource reservation request for an MN is accepted without any effect on the QoS of other MNs in the same AR coverage area. Second, the reservation request is rejected due to the sharing of the wireless link reaching its full capacity. Third, the reservation request is accepted with poorer QoS for other MNs due to heavy sharing of the channel's bandwidth. Typically, when a network resource is overloaded, it is preferable to keep reservations of the previously established flows, while blocking new RSVP requests. Therefore, the reservation requests in connection with RSVP handoff flows should be given higher priority than new requests when performing resource reservation. In this section, we propose a dynamic resource allocation scheme during RSVP resource reservation over wireless link. This scheme is eventually able to give a statistical guarantee on the handoff success of on-going flows.

A. Model Description

In IntServ-enabled RANs with service classes such as guaranteed service (GS) and controlled-load service (CLS), a cell might accept GS, CLS, and best effort (BE) flows. Moreover, a BS (or AR) typically has to perform explicit radio resource allocation and exercise admission control between these service classes, and eventually provide end-to-end QoS provisioning. In order to provide QoS guarantees to the GS class, it is necessary to reserve an amount of bandwidth. It is also required to effectively balance bandwidth sharing between the various service classes in order to minimize service degradation as much as possible.

The proposed model has infinite waiting queue for both CLS and BE handoff flows and guard channels for GS handoff flows. We also allow BE flows to overflow over the region assigned for GS flows with the risk of being preempted by the newly arriving GS flows. We assume the presence of control channels to manage the queue for both CLS and BE flows between the control unit (located in the AR) and MN. Basically, BE service should provide bandwidth fairness to TCP flows so that each TCP flow can get fair amount of bandwidth during periods of congestion. Typically, it is much natural that BE flows are considered at packet level.

For the simplicity of analysis, however, we here assume a BE flow has unit bandwidth and just see the flows at call level. We can estimate the congestion of each BE flow with the mean system (waiting) time of flows. If BE flows are preempted and then queued at call level, some BE flows may not receive any bandwidth for large time periods. From the real TCP point of view over wireless link, this kind of disconnection periods can be of the order of several seconds causing packet loss or delay in the transmission of acknowledgment of received packets. This disconnection results in the TCP sender timing-out and closing its congestion window and, thus greatly reducing the efficiency of the connection. In order to ensure that the TCP connection to a mobile is efficient, it is necessary to prevent the sender from shrinking its congestion window when packets are lost due to disconnection. Furthermore, it is important to ensure that, when the mobile is reconnected, it begins receiving data immediately.

The proposed dynamic allocation scheme for efficient bandwidth utilization can also be described as follows.

- When an MN with GS flows attempts to reserve resource via a new AR at soft handoff, the GS new (handoff) flows can use up to the given wireless channel bandwidth limits with the preemptive priority over BE flows in any coverage area.

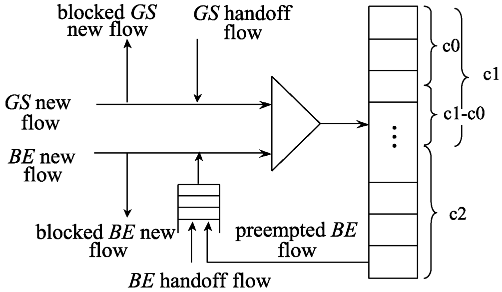

Fig. 16. Resource allocation diagram for numerical model.

- Moreover, the GS handoff flows can use exclusive guard channels in the GS bandwidth region for the purpose of minimizing the forced termination. If there is no idle bandwidth in the region reserved for GS flows in heavily crowded AR coverage area during a soft handoff, RSVP resource reservation via this AR fails and, thus, the GS handoff request is just dropped. Hence, it is natural to give a statistical guarantee on the GS handoff success in wireless environments.
- The remaining bandwidth channels except for the regions reserved for GS flows in a cell are dedicated to both CLS and BE flows. Furthermore, CLS handoff flows also have preemptive priority over BE flows existing already in dedicated region, and BE flows preempted by the arriving GS flows and CLS handoff flows in a cell are queued in the infinite queue to wait for service instead of being cleared.
- However, CLS handoff flows are also queued if there is no existing BE flow to be preempted and queued. If there is any flow waiting for service in the queue, both originating CLS flows and BE flows are blocked.

B. Numerical Model

In the numerical model, we assume that service classes are classified into GS and BE for simplicity of mathematical analysis. We assume that a total capacity of cell is divided by m basic bandwidth units (BBUs), where BBU corresponds to a logical channel. We also assume that a BE flow has one BBU for the convenience of the comparison, the GS flow requires n BBUs $(n < m)$, and all BBUs assigned to a GS flow are occupied and released together. $c0$ is the number of BBUs reserved for the new GS flows and $c1$-$c0$ is the number of BBUs reserved for the GS handoff flows (see Fig. 16). We assume that the overall wireless cellular system is homogeneous, i.e., all cells are statistically identical. Therefore, we can analyze the overall system by focusing on only one cell and consider the statistical behavior of this focused cell under the condition that all the neighboring cells exhibit typical random independent behavior.

We assume that the GS class and the BE class flows arrive according to a Poisson process with mean arrival rate λ_k for a type k flow ($k = 1, 2$), and that service time is exponentially distributed with mean service time of $1/\mu_k$ for a type k ($k = 1, 2$) flow. Then, the system can be modeled as a two-dimensional Markov process, characterized by $\{n2(t), n1(t)\}$, where $n2(t)$ and $n1(t)$ are the numbers of BE and GS flows in the system at time t, respectively. In addition, the state space is represented by the set $\{s(n2, n1) | 0 \leq n2, 0 \leq n1 \leq \lfloor \mathrm{c}1/n \rfloor\}$, where $\lfloor x \rfloor$ denotes the greatest integer smaller than or equal to x. Also, the steady-state probability that the system is in state $s(n2, n1)$ be $p(n2, n1)$. The steady-state probability vector P (ordered lexicographically) is then partitioned as $P = (p_0, p_1, \ldots)$, where for $i = 0, 1, 2, \ldots, p_i = (p(i, 0), p(i, 1), \ldots, p(i, \lfloor \mathrm{c}1/n \rfloor))$. The vector P is the solution of equations

$$\begin{aligned} PQ &= 0 \\ Pe &= 1 \end{aligned} \tag{1}$$

where e and 0 are vectors of all ones and zeros, respectively, and Q is the transition rate matrix of the Markov process. We can obtain the transition rate matrix Q of the Markov process (whose transition matrix is of the block-partitioned form) for the proposed model, which can be formulated as quasi-birth-and-death processes (QBDs) [22]

$$Q = \begin{bmatrix} A_0 & D_0 & & & & & & \\ B_1 & A_1 & D_1 & & & & & \\ & \cdot & \cdot & \cdot & & & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & B_r & A_r & D_r & & \\ & & & & B_r & A_r & D_r & \\ & & & & & \cdot & \cdot & \cdot \end{bmatrix} \tag{2}$$

where the submatrices are defined for $i, j = 0, 1, \ldots, \lfloor \mathrm{c}1/n \rfloor$ and $l = 0, 1, \ldots, r(n \cdot \lfloor \mathrm{c}1/n \rfloor + r \geq m)$ by

$$\begin{aligned} \mathrm{A}_l(i,j) &= \begin{cases} \lambda_1, & \text{if } i = j-1 \text{ and } j \leq \left\lfloor \frac{\mathrm{c}0}{n} \right\rfloor \\ \lambda_{1h}, & \text{if } i = j-1 \text{ and } \left\lfloor \frac{\mathrm{c}0}{n} \right\rfloor < j \leq \left\lfloor \frac{\mathrm{c}1}{n} \right\rfloor \\ (j+1)\mu_1, & \text{if } i = j+1 \\ a_1(i), & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases} \\ \mathrm{B}_l(i,j) &= \begin{cases} \min(l, m - n \cdot j)\mu_2, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \\ \mathrm{D}_l(i,j) &= \begin{cases} \lambda_2, & \text{if } i = j \text{ and } l < m - n \cdot j \\ \lambda_{2h}, & \text{if } i = j \text{ and } l \geq m - n \cdot j \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

where $al(i)$ is the value that makes the sum of the row elements of Q to be equal to zero.

To solve (1) with this transition rate matrix Q, we apply the Neuts' theorem [20] to the matrix Q. Let us first consider the stability of the system. The system is stable when the Markov process Q is positive recurrent, i.e., $\pi B_\mathrm{r} e > \pi D_\mathrm{r} e$, where π is the stationary probability vector of $A_\mathrm{r} + B_\mathrm{r} + D_\mathrm{r}$. Intuitively, the system is stable under the condition $\lambda_2 < r\mu_2$. Next, we determine the minimal nonnegative matrix R of the matrix equation

$$R^2 B_r + R A_r + D_r = 0 \tag{4}$$

by iteration. The steady-state probability vector $Pk(k \geq r)$ is given by

$$p_k = p_{r-1} R^{k-r+1}, \qquad k \geq r. \tag{5}$$

Next, the probability vector $\widetilde{p} = (p_0, p_1, \ldots, p_{r-1})$ is solved from the equations

$$\begin{aligned} &\widetilde{p} T = 0 \\ &\widetilde{p} e + p_{r-1} R (I - R)^{-1} e = 1 \end{aligned} \tag{6}$$

where the matrix T is a generator $(Te = 0)$ given by

$$T = \begin{bmatrix} A_0 & D_0 & & & & \\ B_1 & A_1 & D_1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & B_{r-2} & A_{r-2} & D_{r-2} \\ & & & & B_{r-1} & A_{r-1} + RB_r \end{bmatrix}. \tag{7}$$

The solution is then obtained by a two-step process: First, the rate matrix R is determined by iterative substitution in (4). Second, once R is known, the boundary probabilities are determined by solving a system of linear equations. If the dimension of the matrix T for the boundary probabilities is large, the direct solution of the linear system (6) is difficult to provide the desired accuracy. An alternative technique using decomposition can be applied to reduce the dimension. Once the matrix R and the boundary probabilities have been computed, the performance measures are easily obtained.

C. *Performance Measures*

The probability that bandwidth reservation request for a originating GS flow fails is

$$P_{B_1} = \sum_{n2=0}^{\infty} \sum_{n1=k(c0)}^{k(c1)} p(n2, n1) \tag{8}$$

which, using (5), is easily reduced to

$$P_{B_1} = \sum_{n2=0}^{r-1} \sum_{n1=k(c0)}^{k(c1)} p(n2, n1) + \left[P_{r-1} R (I - R)^{-1} \right]_{(k(c0)..k(c1))} \tag{9}$$

where $k(l)$ denotes the function of $\lfloor l/n \rfloor$, and $[]_{(a..b)}$ denotes the sum to the bth element from the ath element of the vector in the bracket.

The handoff failure probability of a GS flow is the probability that an arriving GS handoff request sees all the bandwidth channels assigned for GS handoff flows in a cell busy, and thereby the BS fails to reserve an amount of wireless bandwidth required to keep going with QoS guarantee. This metric should be minimized to give statistic QoS guarantee. It is given by

$$\begin{aligned} P_{O_1} &= \sum_{n_2=0}^{\infty} p(n_2, k(c_1)) \\ &= \sum_{n_2=0}^{r-1} p(n_2, k(c_1)) + \left[P_{r-1} R (I - R)^{-1} \right]_{(k(c_1))}. \end{aligned} \tag{10}$$

The forced termination probability of a GS flow, PFT is the probability that a GS flow initially accepted is interrupted due to bandwidth reservation failure over wireless link by handoff during its lifetime. We just consider GS new flows. A GS new flow is initially accommodated by a cell and then managed with priority so as not to be terminated due to bandwidth reservation failure over wireless link during handoff. Let α_1 be the probability that a GS flow currently served by a cell requires another handoff request which is accommodated by neighboring cells.

Similarly, let α_2 denote the probability that the call requires another handoff request which fails. Then, using the Markovian properties of the model, we determine

$$\alpha_1 = (1 - P_{O_1})\frac{U_1}{\mu_1 + U_1} \quad (11)$$

$$\alpha_2 = P_{O_1}\frac{U_1}{\mu_1 + U_1} \quad (12)$$

where μ_1 is average session duration time of GS flow and U_1 is average bandwidth channel holding time of GS flow.

Now, let us focus on a GS new flow that is initially accommodated by a cell. When this flow leaves the service of a cell, it can be forced into termination by bandwidth reservation failure over wireless link. Thus, the probability that a GS flow currently accommodated by a cell is forced into termination is

$$P_{FT} = \sum_{i=0}^{\infty} \alpha_1^i \alpha_2. \quad (13)$$

Now, we can consider the average system time (i.e., the sum of queueing time and service time) of BE flows. The mean number of a BE flow in a cell is

$$N_2 = \sum_{n_1=0}^{k(c_1)} \sum_{n_2=0}^{\infty} n_2 p(n_2, n_1) \quad (14)$$

which is, also using (5), reduced to

$$N_2 = P_{r-1}R^2(I-R)^{-2}e + rP_{r-1}R(I-R)^{-1}e + \sum_{n_1=0}^{k(c_1)} \sum_{n_2=0}^{r-1} n_2 p(n_2, n_1). \quad (15)$$

In addition, the blocking probability of BE flows in a cell is

$$P_{B_2} = \sum_{n_1=0}^{k(c_1)} \sum_{n_2=(m-n_1\cdot n)}^{\infty} p(n_2, n_1) = \sum_{n_1=0}^{k(c_1)} \sum_{n_2=(m-n_1\cdot n)}^{r-1} p(n_2, n_1) + P_{r-1}R(I-R)^{-1}e. \quad (16)$$

Therefore, the mean system time for the BE flows from Little's formula is

$$W_2 = \frac{N_2}{\lambda_{2\mathrm{h}} + \lambda_{2\mathrm{o}}(1 - P_{\mathrm{B}_2})}. \quad (17)$$

Another important system performance is carried traffic. For a given number of channels, a large carried traffic value implies efficient use of bandwidth. Carried traffic per GS flow in a cell E_{GS} and carried traffic per BE flow E_{BE} can be easily calculated once the state probabilities are determined. They are simply the average number of occupied channels per flow, can be as given by

$$E_{\mathrm{GS}} = \sum_{n_1=0}^{k(c_1)} \sum_{n_2=0}^{\infty} n_1 p(n_2, n_1) = \sum_{n_1=0}^{k(c_1)} \sum_{n_2=0}^{r-1} n_1 p(n_2, n_1) + 2^{-1}k(c_1)(k(c_1)+1)P_{r-1}R(I-R)^{-1}e \quad (18)$$

$$E_{\mathrm{BE}} = \sum_{n_1=0}^{k(c_1)} \sum_{n_2=0}^{m-n_1\cdot n} n_2 p(n_2, n_1). \quad (19)$$

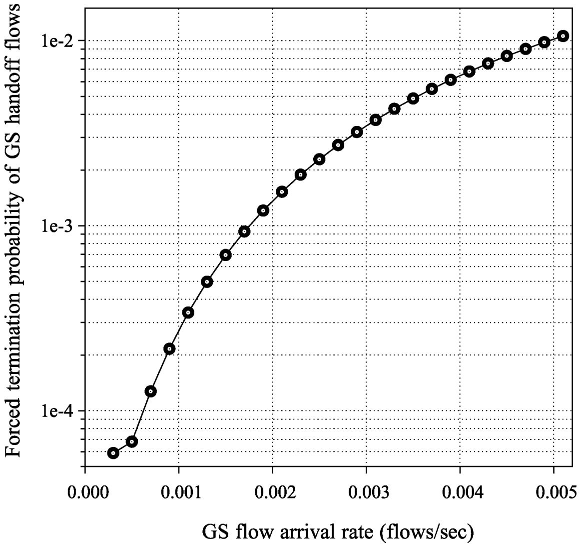

Fig. 17. Forced termination probability for GS flow.

Because a GS flow requires n BBUs, the total carried traffic in a cell is given by

$$E = nE_{\mathrm{GS}} + E_{\mathrm{BE}}. \quad (20)$$

To measure the combined QoS of the BE and GS flows, we can get the quality factor QT for BE and GS flows. Maximization of QT amounts to minimizing of the forced termination probability of GS flow and the waiting delay of BE flow. QT factor can be defined by

$$QT = \frac{W_{\mathrm{GT}}(1 - P_{\mathrm{FT}})}{\mu_2 W_2}, \quad (0 < QT < 1) \quad (21)$$

where W_{GT} is an weighted factor to adjust the difference between P_{FT} and W_2.

D. Examples and Discussions

For dynamic resource allocation scheme for RSVP reservation over wireless link, in this section, we present analytical results for an example of a system with both GS and BE flows. The total bandwidth units of each cell are 6. In each cell, the number of bandwidth units assigned for GS handoff flows and GS new flows are 2 and 4, respectively. That is, $n = 2$, $m = 6$, c0 $= 2$, and c1 $= 4$. In this situation, a BE flow can overflow to the region assigned for GS flows if the number of GS flows in use is less than 4. A GS new flow request will be rejected if the number of GS flows in use is equal to 2. The arrival rate of GS flows varies from 0.0003 to 0.005 (flows/s). The arrival rate of BE flow is kept constant at 11 Erlangs. The average channel holding time of GS flows in each cell is 200 s.

When the arrival rate of GS new flows in a cell increases to the limit, they might be early blocked for the purpose of improving the quality of service by reducing the input traffic load. Fig. 17 shows the forced termination probability of GS flows. The exclusive guard bandwidth channels can control the forced termination of GS handoff flows at reasonable level. Hence, the statistical guarantee of handoff success of GS flows can be obtained and, thus, this can provides reasonable QoS guarantee of GS handoff flows.

Fig. 18 shows the mean service time of BE flows. In this figure, the arrival rate of BE flows is kept constant at 11 Erlang,

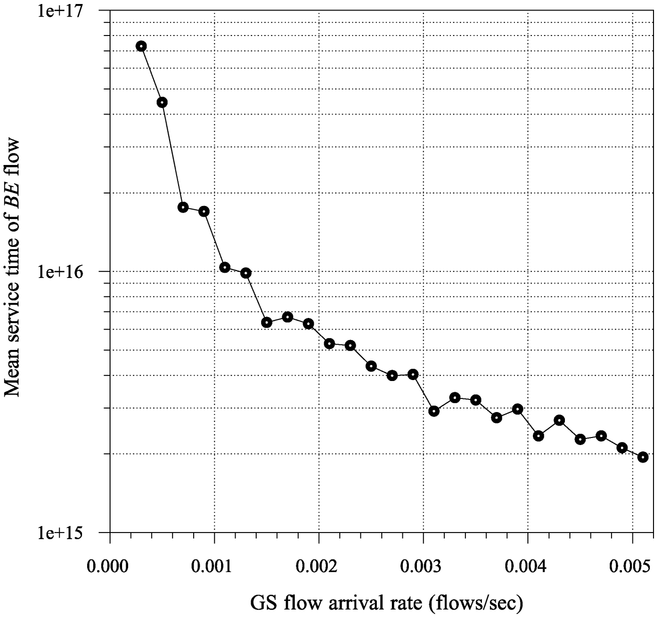

Fig. 18. Mean service time of BE flows.

which is very high traffic load. In the case that the arrival rate of GS flows is less than 0.0003, BE flows could be overflowed into the region reserved for GS flows. However, as the arrival rate of GS flow increases (e.g., 0.0003), most of BE flows are suddenly preempted and queued. This situation makes the queue length of BE flows longer suddenly and, thus, the mean service time also becomes prolonged. In addition, newly arriving BE flows are earlier blocked for preventing severe congestion and later, at some stable traffic load, BE flows will be accepted again. Due to this kind of phenomenon, there exist some fluctuations in mean service time of BE flows. Consequently, the probability that BE flows are preempted and queued goes up gradually as the arrival rate of GS flows increases. Thus, the total number of BE flows in a system decreases continuously as the arrival rate of GS flows increases.

VI. CONCLUSION

In this paper, we considered QoS mechanisms for maintaining real-time flows to roaming hosts while reducing excessive RSVP signaling overhead due to frequent host mobility. First, we showed that a CR-based rerouting scheme has less mean path length than those of other schemes and pays very reasonable mean resource reservation cost even though the number of nodes increases highly. Second, we also showed that an advance reservation of RSVP branch path for soft handoff can provide the seamless service by adaptively adjusting the pilot signal threshold and by defining well the number of tree levels in a RAN. Third, we developed bandwidth-efficient resource allocation model to aid the RSVP reservation over wireless link.

ACKNOWLEDGMENT

The work reported in this paper has formed part of the Networks and Services work area of the Core 2 Research Program of the Virtual Centre of Excellence in Mobile and Personal Communications, Mobile VCE (www.mobilevce.com), whose funding support, including that of EPSRC, is gratefully acknowledged. More detailed technical reports on this research are available to Industrial Members of Mobile VCE.

REFERENCES

[1] B. K. Moon and A. H. Aghvami, "RSVP extensions for real-time services in wireless mobile networks," *IEEE Commun. Mag.*, vol. 39, pp. 52–59, Dec. 2001.

[2] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "MRSVP: A reservation protocol for an integrated services packet network with mobile hosts," Dept. Comput. Sci., Rutgers Univ., Piscataway, NJ, Tech. Rep. TR-337.

[3] W. T. Chen and L. C. Huang, "RSVP mobility support: A signaling protocol for integrated services Internet with mobile hosts," in *Proc. INFOCOM*, vol. 3, Mar. 2000, pp. 1283–1292.

[4] A. Terzis, M. Srivastava, and L. Zhang, "A simple QoS signaling protocol for mobile hosts in the integrated services Internet," in *Proc. INFOCOM*, vol. 3, Mar. 1999, pp. 1011–1018.

[5] G. Chiruvolu, A. Agrawal, and M. Vandenhoute, "Mobility and QoS support for IPv6 based real-time wireless Internet traffic," in *Proc. IEEE Int. Conf. Communications*, vol. 1, June 1999, pp. 334–338.

[6] Q. Shen, A. Lo, W. Seah, and C. C. Ko, "On providing flow transparent mobility support for IPv6-based real-time services," presented at the MoMuC 2000, Tokyo, Japan, Oct. 2000.

[7] R. Jain, R. Burns, M. Bereschinsky, and C. Graff, Mobile IP with location registers (MIP-LR), Internet-Draft, draft-jain-miplr-01.txt, July 2001.

[8] A. T. Campbell and G. C. Javier, "IP micro mobility protocols," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 4, no. 4, pp. 45–53, Oct. 2000.

[9] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, Resource reservation protocol (RSVP), version 1 functional specification, RFC 2205.

[10] T. Zhang, P. Agrawal, and J. C. Chen, "IP-based base stations and soft handoff in all-IP wireless networks," *IEEE Pers. Commun. Mag.*, vol. 8, pp. 24–30, Oct. 2001.

[11] C. K. Toh, "Performance evaluation of crossover switch discovery algorithms for wireless ATM LANs," in *Proc. INFOCOM*, San Francisco, CA, 1996, pp. 1380–1387.

[12] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, and L. Salgarelli, IP micro-mobility support using HAWAII, Internet Draft, draft-ietf-mobileip-hawaii-01.txt, July 2000.

[13] G. C. Lee, T. P. Wang, and C. C. Tseng, "Resource reservation with pointer forwarding schemes for the mobile RSVP," *IEEE Commun. Lett.*, vol. 5, pp. 298–300, July 2001.

[14] W. S. Wong, C. B. Chan, and C. M. Leung, "Path optimization for inter-switch handoff in wireless ATM networks," in *Proc. ICUPC*, Florence, Italy, Oct. 1998, pp. 615–619.

[15] IETF Internet Draft, RSRR: A routing Interface for RSVP, draft-ietf-rsvp-routing-02.txt, D. Zappala, J. Kann, July 1998.

[16] J. W. Chang and D. K. Sung, "Adaptive channel reservation scheme for soft handoff in DS-CDMA cellular systems," *IEEE Trans. Veh. Technol.*, vol. 50, pp. 341–353, Mar. 2001.

[17] D. Wong and T. J. Lim, "Soft handoffs in CDMA mobile systems," *IEEE Pers. Commun.*, vol. 4, pp. 6–17, Dec. 1997.

[18] H. Harri and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. New York: Wiley, Spring 2001.

[19] A. T. Campbell, G. Javier, and A. G. Valko, "An overview of cellular IP," in *Proc. IEEE Wireless Communications Networking Conf. (WCNC'99)*, New Orleans, LA, Sept. 1999, pp. 606–610.

[20] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. New York: Dover, 1994.

[21] A. Neogi and T. C. Chiueh, "Performance analysis of an RSVP-Capable router," *IEEE Network*, vol. 13, pp. 56–63, Sept. 1999.

[22] Y. H. Kim and C. K. Un, "Analysis of bandwidth allocation strategies with access restrictions in broadband ISDN," *IEEE Trans. Commun.*, vol. 41, pp. 771–781, May 1993.

Bongkyo Moon received the B.S. degree in computer science from Sogang University, Seoul, Korea, in 1992, the M.S. degree in information and communications from Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, in 1998, and the Ph.D. degree from the Centre for Telecommunications Research, King's College, University of London, London, U.K., where he took part in Mobile VCE Project for IP QoS and mobility area.

He worked as an Engineer in the software and telecommunication areas at the Electronics and Telecommunications Research Institute (ETRI), Taejon, Korea, from 1998 to 1999 and at Inex Technologies, Inc., Santa Clara, CA, from 1992 to 1996. He is currently working as a Senior Engineer in the Telecommunication R&D Centre, Samsung Electronics, Suwon, Korea.

Hamid Aghvami (M'87–SM'91) received the M.Sc. and Ph.D. degrees from King's College, University of London, London, U.K., in 1978 and 1981, respectively.

In April 1981, he joined the Department of Electronic and Electrical Engineering, King's College, as a Postdoctoral Research Associate and worked on digital communications and microwave techniques projects sponsored by EPSRC. He joined the Academic Staff at King's College in 1984. In 1989, he was promoted to Reader and Professor of Telecommunications Engineering in 1992. He is presently the Director of the Centre for Telecommunications Research, King's College. He carries out consulting work on digital radio communications systems for both British and International companies. He was a Visiting Professor at NTT Radio Communication Systems Laboratories, Japan, in 1990 and a Senior Research Fellow at BT Laboratories, U.K., in 1998–1999. He is currently an Executive Advisor to Wireless Facilities, Inc., and Managing Director of Wireless Multimedia Communications, Ltd. (his own consultancy company). He has published over 250 technical papers and given invited talks all over the world on various aspects of personal and mobile radio communications, as well as giving courses on the subject worldwide. He leads an active research team working on numerous mobile and personal communications projects for fourth-generation systems; these projects are supported by the government and by industry.

Dr. Aghvami is a Distinguished Lecturer of the IEEE Communications Society. He was an Elect Member of the Board of Governors of the IEEE Communications Society in 2002–2003. He has been a member, Chairman, and Vice-Chairman of the technical program and organizing committees of a large number of international conferences. He is also the Founder of the International Conference on Personal Indoor and Mobile Radio Communications (PIMRC). He is a Fellow Member of the Royal Academy of Engineering and the IEE.