On Accommodating Mobile Hosts in an Integrated Services Packet Network

Anup K. Talukdar, B. R. Badrinath*

Department of Computer Science
Rutgers University
New Brunswick, NJ 08903

Arup Acharya

C&C Research Labs
NEC USA
Princeton, NJ 08540

Abstract

This paper considers the support of real-time applications to mobile hosts in an Integrated Services Packet Network. We have proposed a service model for mobile hosts that can support adaptive applications which can withstand wide range of available bandwidth, as well as applications which require mobility independent service guarantees. We describe an admission control scheme and a reservation protocol for implementing this service model. Our admission control scheme achieves high utilization of network resources.

1 Introduction

Recent progress in computing technology and wireless digital communication has made portable computers easily available. This has led to an intensive research in the area of *Mobile Computing* to provide mobile users access to an inter network. The research, so far, has focussed on the problem of maintaining connectivity at the network and transport layer inspite of the mobility of the mobile hosts[11, 3, 2].

Also, there have been several proposals for supporting real-time applications in an Integrated Services Packet Network (ISPN). Typical applications that require real-time services include audio library, image browsing, video conferencing and video-on-demand. These multimedia applications require a bound, which may be absolute or statistical, on the delivery delay of each packet. Clark et al.[4] have described an architecture for an ISPN that supports real-time traffic. In this architecture they have described a service model for real-time traffic with two service classes, *guaranteed and predictive.* Jamin et al.[6] have described a scheme to implement the service model described in [4].

As portable computers become more powerful and the accessibility of a fixed network from a mobile host becomes easier, the number of mobile users will grow and additionally the mobile users will demand the same real-time services available to fixed hosts. In this paper, we investigate the problems of providing real-time services to mobile hosts in an ISPN. As described in [4], to handle real-time services an enhanced architecture is required for a network. This architecture has five key components. The first component is the nature of *service commitments* the network can provide. The second component is the *service interface parameters* passed between the network and the flow endpoints; this includes both the characterization of the quality of service the network can provide and the characterization of the behavior of the endpoints of the flow. The third component is the functionality of the network elements, namely *scheduling algorithms*, required to meet the service commitments. The fourth component is the *admission control* mechanism by which the service commitments are established at each network element. The last component is the *reservation protocol* to setup the required states in the network elements for providing the required service. In this paper, we discuss how mobile users can be accommodated within this architecture.

This paper is organized as follows. In Section 2 we discuss the problems of providing real-time services to mobile users in an ISPN. In Section 3 we outline

*This research work was supported in part by DARPA under contract number DAAH04-95-1-0596, NSF grant numbers CCR 95-09620, IRIS 95-09816.

the related works in this area and describe our approach to the problem. In Section 4 we describe the service interfaces and propose the service models required for the ISPN with mobile users and describe the scheduling algorithms used for the service classes; Section 5 describes the admission control mechanisms for our proposed service models and mobile handoff techniques; in Section 6 we describe reservation protocols; we conclude in Section 7 with some future directions of this work.

2 Real-time services to mobile users : Problems

In an ISPN, the main Quality of Service (QoS) parameter is the delay experienced by a packet traveling from a sender to a receiver. The packet delay has three components. The first component, *propagation delay*, is the propagation time of the packet at the speed of light; this is fixed once the data path is chosen. The data path, usually, consists of multiple hops, connected by switches or routers. The second component is the delay in transmission at each switch waiting for the entire packet to arrive before it can be transmitted onto the next link; this delay depends on the packet size. The third component, *congestion delay*, arises due to the statistical sharing of the finite link bandwidth by packets belonging to multiple flows: packets arriving at a switch are buffered in a service queue until the outgoing link is available. *Congestion delay* of a packet is the time spent by it in the service queue; this delay depends on the number of flows using a link, the link capacity and the traffic generation rate of different sources. To provide real-time services, this congestion delay must be bounded or minimized.

In [4], real-time applications have been classified into two classes, *tolerant* applications which can adapt to packet delays and thus can tolerate occasional delay bound violations, and *intolerant* applications which cannot tolerate any delay bound violations. Correspondingly, they have defined two service models: *guaranteed* service for intolerant applications and *predictive* service for tolerant applications. In an ISPN consisting of static hosts QoS guarantee is provided by reserving sufficient bandwidth at each link along the path from the sender to the receiver for the duration of the session, so that the congestion delay at each switch is bounded. In a mobile computing network, a geograhic area is divided into several cells. Each cell is served by a *base station* (also called *mobility agent*) connected to the fixed network. A mobile host maintains connectivity with the fixed network through the base station of the cell in which it is currently located. When a mobile host moves from one location to another, the delivery delay of a packet is affected in two ways. The first factor is that the propagation delay may change due to the change in the length of the path from the sender to the receiver. Secondly, the congestion delay at the switches along the new path may be different. Therefore, when a mobile host initiates a session with a certain QoS guarantee by reserving link bandwidth along the path from the sender to its current location, the QoS guarantee is valid only in that location; when the mobile host moves to a new location, the QoS guarantee is not valid in the new location.

There are two broad ways to provide service guarantees to mobile hosts. The first approach is location dependent: the QoS is guaranteed specific to a location, i.e., QoS guarantee is maintained as long as the mobile host stays at the location from where it initiated the session. To obtain such a service the mobile host makes a reservation for a certain QoS from its current location. As soon as it moves to a new location, the QoS guarantee is no longer valid. The application has to renegotiate its desired QoS at the new location. If sufficient resources are not available along the new data flow path, the mobile host may suffer service degradation, which we call *handoff failure.* The second approach is location independent: the mobile host receives the same QoS guarantee at all locations (given by a mobility profile[1]) where it may move during the lifetime of the session. For example, suppose at time t_1, the mobile host is at cell C_1 where it initiates a session with a certain QoS guarantee, and at time t_2 it is located in cell C_2 where $t_1 < t_2$. It will receive the same QoS guarantee at time t_2 as long as C_2 is within its mobility profile. To obtain mobility independent QoS, a mobile host must make reservations along the data flow path from the sender to each location it may visit. However this leads to very low utilization of resources because, although data is physically flowing to all locations, data is being used only at the current location of the mobile. The utilization of the network resources can be significantly improved if the reserved resources of unused data flows could be used by other flows. However, the users who are utilizing these unused resources may suffer service degradation when the original reservers start using these resources. For example, suppose a mobile host M_1 has made reservation at the two cells, C_1 and C_2, and it is currently located at C_1 and hence using the reserved resources at C_1. In this situation, another host M_2 which is located in C_2 can use the resource reserved by M_1 in

[1]mobility profile of a mobile host is the set of locations the mobile host is expected to visit

C_2. But when M_1 moves into C_2, M_2 must release the resources reserved for M_1 at C_2.

3 Related works and our approach

Recently there has been some work addressing the problem of providing QoS to mobile hosts. In [1] an architecture for a high-speed mobile ATM network using a new concept known as the *virtual connection tree* is proposed and an analytical method of admission control for mobile hosts is described. In [12], Singh has introduced two new QoS parameters, *loss profile* and *probability of seamless communication*, arising due to the mobility of the mobile host and described a network architecture and a suite of transport level services to satisfy these QoS parameters. Lee, in [7], has presented an "adaptive reserved service" framework for use in integrated services networks to support mobile connections carrying multimedia traffic. They have proposed an adaptive link partitioning scheme to reserve a fraction of the link bandwidth to enhance handoff success for mobile hosts. In [8], Levine et al. have proposed resource allocation and admission control schemes based on a new concept, called *Shadow Cluster*, to improve the QoS of mobile calls by reducing the number of dropped calls in an wireless-ATM network. In [9], an adaptive resource management algorithm is described for indoor mobile environments.

These research have focussed on providing QoS guarantees to adaptive applications which can withstand wide range of available bandwidth. In these works, the main objective of the admission control and resource allocation schemes is to reduce the rate of handoff failures of the mobile hosts without significant degradation in the utilization level of the network bandwidth. Mobile applications subscribing to those services may suffer variable degrees of QoS degradation as they move from one location to another. Therefore these schemes are not suitable for more rigid applications which require QoS guarantees that are not affected by mobility of hosts.

In this paper we propose an architecture that accommodates those applications which are adaptive to variable degradation in QoS and also those applications which require mobility independent QoS guarantees in the same network. A salient feature of this architecture is that, a mobile host is required to negotiate its desired QoS only at the start of a data flow session and eliminates the need of fresh reservation every time it moves to a new location during the lifetime of the session. The admission control scheme, we propose, attempts to increase utilization of network bandwidth by statistical multiplexing of link bandwidth among flows of different classes of applications.

4 Service Interfaces and Service Models

To obtain a certain QoS in an ISPN, the endpoints of a data flow is required to provide characterization of the data traffic they will generate, so that the network can reserve sufficient resources for the flow. This characterization is done with a *token bucket filter*. A token bucket filter has two parameters: its token generation rate, r, and the depth of its bucket, b. Tokens are generated at the fixed rate r and accumulate into the bucket. When the bucket becomes full, i.e. contains b tokens, newly generated tokens are discarded. When a packet of size p is generated, it can be released into the network only after removing p tokens from the bucket. In addition to securing the required number of tokens from the bucket the traffic is constrained to enter the network at a maximum rate of $C > r$, called *peak rate*. A traffic flow conforms to a token bucket filter (r, b) if there are enough tokens in the bucket whenever a packet is generated.

In a fixed network the required resources are reserved at the switches along the data flow path. In a mobile environment the data flow path changes as a mobile host moves from one cell to another. To provide real-time services to a mobile host, it is necessary to reserve resources along all possible data paths which may be used during the lifetime of a connection. Therefore, in addition to the traffic characterization, the mobile host is also required to provide its mobility characterization to the network. In this paper we assume that, the mobility of an user is predictable so that mobility can be characterized precisely by *mobility specification* which consists of the set of cells the mobile host is expected to visit during the lifetime of the flow. In the rest of this section we describe the service models for the mobile multimedia applications and the scheduling algorithms used for these service models.

4.1 Service Models

The service models, we propose, are derived from the two service models, *guaranteed* and *predictive* services, described in [4] by Clark et al. We define three service classes to which mobile users may subscribe. These are:

1. Mobility Independent Guaranteed service (MIG class): A mobile user admitted to this service class will receive *guaranteed* service with respect to packet delay bounds as long as its moves are limited to its *mobility specification* and it is conforming to its traffic characterization. This class is appropriate for the intolerant applications which require absolute bound on packet delay.

2. Mobility Independent Predictive service (MIP class): A mobile user admitted to this class will receive *predictive* service with respect to packet delay bound as long as its moves are limited to its *mobility specification* and it is conforming to its traffic characterization. This class is appropriate for those tolerant applications which require fairly reliable delay bounds in all cells it might visit and does not want to be affected by mobility of the hosts.

3. Mobility Dependent Predictive service (MDP class): A mobile user admitted to this service class will receive *predictive* service with high probability in all cells it may visit during the lifetime of its connection as long as its moves are limited to its *mobility specification* and it is conforming to its traffic characterization. However it may occasionally fail to get *predictive* service and experience severe degradation of QoS. This class is appropriate for tolerant applications, which can tolerate the effects of delay violations due to mobility of hosts.

The *predictive* flows (both in MIP and MDP classes) are categorized into a number of levels with widely separated delay bounds. The delay bounds of the corresponding levels of MIP and MDP flows are same, i.e. MIP level j and MDP level j have the same delay bounds.

To implement the above service models for mobile hosts in ISPN, it is not enough to reserve resources along the path from the sender to the current location of the mobile host; it is necessary to make reservation along the paths to other locations where the mobile host may visit. However, it is not necessary to initiate the data flow along each of those paths; data flow is initiated only along the path from the sender to the current location of the mobile host. Thus we define two types of flows:

1. *Active* : A flow is active at a switch along a data path if resources are reserved for the flow at the switches along the data path and data is being transmitted to a receiver along that path. The corresponding reservation is called an active reservation.

2. *Passive* : A flow is passive at a switch along a data path if resources are reserved for the flow at the switches along the data path but the data is not passing through the switch. The corresponding reservation is called a passive reservation.

One approach to reservation in an ISPN is to use the *a priori* worst-case traffic characterization of each reserving flows to compute the resource requirements to provide the requested services. However,if the flow traffic is bursty, the average flow rate is significantly less than their *a priori* worst-case characterization and as a result the network utilization will be very low. To achieve higher utilization, Jamin et al.[6] have proposed a measurement-based admission control scheme for predictive services. This scheme uses *a priori* characterization of new flows, but uses measurement to characterize existing flow. They have shown significant gain in utilization when there is a high degree of multiplexing.

In our proposed reservation scheme, we use the above measurement-based admission control algorithm. For a flow of MIG or MIP class, the required amount of bandwidth is reserved along the paths from the sender to all locations in the *mobility specification* of the receiving mobile host. However the flow is active only along the path to the current location. To increase the utilization of the network bandwidth, we allow the flows of the MDP class to use the reserved but unused resources of the passive reservation of MIG and MIP classes. When a passive flow becomes active at a switch (i.e. the mobile who made the passive reservation moves into the cell), some MDP flow may suffer QoS degradation if the MDP flows were using the unused resources of the passive flow.

4.2 Scheduling Algorithm

From a scheduling point of view, the flows at a switch are divided into two classes: *guaranteed and predictive*. The MIG flows belong to the guaranteed service class; the MIP and MDP flows belong to the predictive class. There are multiple levels of predictive service with widely separated delay bounds. The scheduling algorithm considers only the active flows at a switch. The unified scheduling algorithm of [4] is used, i.e., guaranteed service is provided by the Weighted Fair Queuing scheduling discipline described in [10, 5] and the predictive service is provided by the priority queuing algorithm using the bandwidth leftover from serving guaranteed service. The flows within each predictive level are scheduled by FIFO algorithm.

5 Admission Control Scheme

The goal of the admission control algorithm should be to admit as many mobile users as possible with the requested QoS and achieve a very high utilization of the resources. However, when a new mobile user is admitted, the system should ensure that it meets its prior commitments to previously admitted users.

To admit a new flow it is necessary to perform admission control at all switches and base stations along the paths from the sender to the cells where the mobile host may visit. The admission control criteria at a switch for a flow will depend on the following two factors: *service class* and whether the flow is *active* or *passive* at that switch when the flow starts. A flow is admitted to the network only if it passes the admission control criteria at each switch or base station where the flow requires reservation.

Our admission control algorithm is derived from the measurement-based admission control scheme for ISPN proposed by Jamin et al. in [6]. In Reference [6], Jamin et al. have described a technique to measure the queuing delays, usage rates of the guaranteed service class and each predictive class. These quantities have been used to compute, for each predictive class, the *equivalent token bucket filter* that characterizes the currently existing aggregate traffic pattern of the active flows of that class. In their scheme, the maximal delay bounds of predictive flows are approximated by replacing the worst-case parameters in the analytical models with *equivalent token bucket filter.* A new flow is admitted only if it can get the requested service and the delay bounds of the existing flows are not violated after admitting this new flow. The admission control criteria in our service models will use the same *equivalent token bucket filter* parameters whenever possible. In addition, we augment the measurement process of [6] to measure the parameter, *token generation rate* of the equivalent token bucket filter for each individual active flow. Before describing the admission control criteria for the different service classes in our model, we first describe how the worst-case delays of predictive services are computed.

5.1 Worst-case delay computation

We assume that there are N levels of predictive services. The delay bound of level i is smaller than the delay bound of level j for $i < j$. The aggregate *token bucket filter* parameter of predictive class j is (r_j, b_j) and μ is the link capacity such that $\sum_{j=1}^{N} r_j \leq \mu$. Parekh, in [10], has shown that in a network with arbitrary topology, under WFQ scheduling discipline the delay bounds of a guaranteed flow depends only on its reserved flow rate and its bucket depth. This implies that, WFQ isolates the guaranteed flows from each other and in addition it isolates all the guaranteed flows from all predictive traffic.

When the packet size is small, i.e. the transmission time of each packet is sufficiently small compared to other delays and hence can be ignored, the worst-case delay, D_j, of priority queue level j is given by the following theorem in [10]:

Theorem 1 *Parekh [10]: The worst-case class j delay, with FIFO discipline within the class and assuming infinite peak rates for the sources, is*

$$D_j^* = \frac{\sum_{i=1}^{j} b_i}{\mu - \sum_{i=1}^{j-1} r_i}$$

for each class j. Further, this delay is achieved for a strict priority service discipline under which class j has the least priority.

When a mobile host h requests to initiate a flow α, it specifies its traffic characteristics with a token bucket filter (r_α, b_α) and mobility specification with a set of cells C_h. In addition it specifies the class of service and if it requests MIP or MDP service, it also specifies the desired delay bound. Depending on the requested service class, the switches and the base stations along the paths from the sender to each of the cell in C_h checks appropriate admission control conditions. The new flow will be admitted only if the admission control criteria are satisfied at all of those switches and base stations. Before describing the admission control criteria for each service class, we describe the notations used in those conditions:

- μ = total link bandwidth, v = utilization target of the link bandwidth
- $MIP_i^a(MIP_i^p)$ = set of active(passive) flows in MIP class i at a switch
- $MIP_i = MIP_i^a \cup MIP_i^p$
- $MIG^a(MIG^p)$ = set of active(passive) flows in MIG class at a switch
- $MIG = MIG^a \cup MIG^p$
- $MDP_i^a(MDP_i^p)$ = set of active(passive) flows in MDP class i at a switch
- $MDP_i = MDP_i^a \cup MDP_i^p$
- $P_i^a = MIP_i^a \cup MDP_i^a$ = set of active flows of predictive class i
- $\nu_C, \hat{\nu}_C, b_C, \hat{b}_C$ = total aggregate bandwidth, measured aggregate bandwidth, aggregate bucket depth, measured aggregate bucket depth of all flows in class C respectively

- $b_\beta, \hat{b}_\beta, r_\beta, \hat{r}_\beta$ = bucket depth, measured bucket depth, requested traffic rate, measured traffic rate of flow β respectively

The parameter v, utilization target, is used to keep some portion of the link bandwidth unreserved. This unreserved bandwidth can be used by the best-effort traffic.

5.2 Admission Control Criteria for MIG service

The flow α, active or passive, is admitted to the MIG service class, if all of the following conditions are satisfied at the switch or base station:

1. The total reserved bandwidth of all flows of MIG class, both active and passive, including r_α does not exceed the targeted utilization of the link bandwidth : $v\mu > r_\alpha + \nu_{MIG}$

2. The sum of the requested flow rate r_α and the current usage should not exceed the targeted link utilization: $v\mu > r_\alpha + \hat{\nu}_{MIG} + \sum_{i=1}^{N} \hat{\nu}_{P_i^a}$

3. After the flow α is admitted, the delay bound of MIP level j, $D^*_{MIP_j}(1 \le j \le N)$, should not be violated when all flows of MIG and MIP class which have made reservation at this switch are in active state. The new delay of MIP level j, D'_{MIP_j}, after α is admitted, is given by:

$$D'_{MIP_j} = \frac{\sum_{i=1}^{j} \hat{b}_{MIP_i}}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{j-1} \hat{\nu}_{MIP_i} - r_\alpha} =$$

$$\frac{\sum_{i=1}^{j} (\sum_{\beta \in MIP_i^a} \hat{b}_\beta + \sum_{\beta \in MIP_i^p} \hat{b}_\beta)}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{j-1} (\sum_{\beta \in MIP_i^a} \hat{r}_\beta + \sum_{\beta \in MIP_i^p} \hat{r}_\beta) - r_\alpha}$$

Although we can estimate the bucket depth in the equivalent token bucket filter of the aggregate traffic of each predictive class from the observed packet delay for that class, it is difficult to estimate the bucket depth in the equivalent token bucket filter of each individual flow in that class. Therefore in the above equation we make the following substitutions :

   (a) $\sum_{\beta \in MIP_i^a} \hat{b}_\beta$ is replaced by $MIN(\sum_{\beta \in MIP_i^a} b_\beta, \sum_{\beta \in P_i^a} \hat{b}_\beta)$

   (b) $\sum_{\beta \in MIP_i^p} \hat{b}_\beta$ is replaced by $\sum_{\beta \in MIP_i^p} b_\beta$

   All other quantities in the above equation can be estimated by measurement. Therefore, the condition to be satisfied is: $D^*_{MIP_j} > D'_{MIP_j}$ for $1 \le j \le N$.

4. The delay bound $D^*_{MDP_j}$ of the MDP level $j, (1 \le j \le N)$, should not be violated when this new flow is admitted. The new delay of MDP level j,$D'_{MDP_j^a}$, after admitting α is computed by taking into account only the active flows of all classes and is given by the following equation:

$$D'_{MDP_j^a} = \frac{\sum_{i-1}^{j} \hat{b}_{P_i^a}}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{j-1} \hat{\nu}_{P_i^a} - r_\alpha}$$

   The quantity $\sum_{i-1}^{j} \hat{b}_{P_i}$ can be computed from the observed delays of the predictive class traffic. Therefore the condition to be satisfied is $D^*_{MDP_j} > D'_{MDP_j^a}$ for $1 \le j \le N$.

5.3 Admission Control Criteria for MIP service

The flow α, active or passive, will be admitted to the MIP class k at a switch or base station if all of the following conditions are satisfied:

1. The sum of the flow's requested rate r_α and the current usage should not exceed the targeted link utilization:
   $v\mu > r_\alpha + \hat{\nu}_{MIG} + \sum_{i=1}^{N} \hat{\nu}_{P_i^a}$

2. After α is admitted, the delay bound $D^*_{MIP_j}$ of the MIP level j, $j \ge k$ should not be violated when all flows of MIG and MIP class who have made reservation at this switch are in active state. The delay bound of the MIP class k, D'_{MIP_k}, is estimated by the following equation after α is admitted:

$$D^{'}_{MIP_k} = \frac{\sum_{i=1}^{k} \hat{b}_{MIP_i} + b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{k-1} \hat{\nu}_{MIP_i}} =$$

$$\frac{\sum_{i=1}^{k} (\sum_{\beta \in MIP_i^a} \hat{b}_\beta + \sum_{\beta \in MIP_i^p} \hat{b}_\beta) + b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{k-1} \sum_{\beta \in MIP_i^a} \hat{r}_\beta}$$

For $k < j \leq N$, the estimated delay, after α is admitted, is given by :

$$D^{'}_{MIP_j} = \frac{\sum_{i=1}^{j} \hat{b}_{MIP_i} + b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{j-1} \hat{\nu}_{MIP_i} - r_\alpha} =$$

$$\frac{\sum_{i=1}^{j} (\sum_{\beta \in MIP_i^a} \hat{b}_\beta + \sum_{\beta \in MIP_i^p} \hat{b}_\beta) + b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{j-1} \sum_{\beta \in MIP_i^a} \hat{r}_\beta - r_\alpha}$$

In the above equations we make the following substitutions:

(a) $\sum_{\beta \in MIP_i^a} \hat{b}_\beta$ is replaced by $MIN(\sum_{\beta \in MIP_i^a} b_\beta, \sum_{\beta \in P_i^a} \hat{b}_\beta)$

(b) $\sum_{\beta \in MIP_i^p} \hat{b}_\beta$ is replaced by $\sum_{\beta \in MIP_i^p} b_\beta$

Therefore the condition to be satisfied is:
$D^*_{MIP_j} > D^{'}_{MIP_j}$, for $k \leq j \leq N$.

3. The delay bound $D^*_{MDP_j}$ of the MDP level j, $j \geq k$ which are active at this switch should not be violated when the new flow α is admitted. The new delay of the MDP class k after admitting α is estimated by the following equation (considering only active flows):

$$D^{'}_{MDP_k^a} = \frac{\sum_{i=1}^{k} \hat{b}_{P_i^a} + b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{k-1} \hat{\nu}_{P_i^a}} =$$

$$D_{MDP_k^a} + \frac{b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{k-1} \hat{\nu}_{P_i^a}}$$

The new delay of the MDP class j, $k < j \leq N$, after admitting α is estimated by the following equation (considering only active flows):

$$D^{'}_{MDP_j^a} = \frac{\sum_{i=1}^{j} \hat{b}_{P_i^a} + b_\alpha}{\mu - \hat{\nu}_{MIG} - \sum_{i=1}^{j-1} \hat{\nu}_{P_i^a} - r_\alpha}$$

Therefore the condition to be satisfied is:
$D^*_{MDP_j} > D^{'}_{MDP_j^a}$, for $k \leq j \leq N$.

5.4 Admission Control Criteria for MDP service

The flow α will be admitted to the MDP class k at a switch in an active state, if all of the following conditions are satisfied:

1. The sum of the flows requested rate r_α and the current usage of the link bandwidth does not exceed the targeted link utilization:
$v\mu > r_\alpha + \hat{\nu}_{MIG} + \sum_{i=1}^{N} \hat{\nu}_{P_i^a}$
2. The delay bounds of the MDP levels k and above should not be violated after admitting this flow; i.e. this condition is same as the condition 3 in the previous subsection.

The flow α will be admitted to the MDP class k at a switch in a passive state without any admission control; it will only be entered into the set of MDP flows who have made passive reservation.

5.5 Handling Handoffs

When a mobile host with an existing flow in a particular service class moves into a new cell within its mobility specification, the delay bounds of the already existing flows in the new cell may be violated or the utilization target may be exceeded. Since the flows of MIG and MIP service classes cannot tolerate the consequences of mobility, we move flows of MDP classes to service classes with higher delay or convert them to best-effort traffic. Whenever a mobile host leaves a cell or a flow is terminated, the existing flows are given preferences to upgrade to their requested class rather than admitting new flows. When a mobile host with an MDP class flow enters a new cell within its mobility specification which cannot provide its existing QoS, the QoS of the mobile host degrades.

6 Reservation Protocol

In ISPN, a reservation protocol, $RSVP$ [14], has been proposed that sets up reservation state at each router along the path from a sender to a receiver. It uses two types of messages to setup the reservation states in the routers, $PATH$ message to setup the data flow path, and $RESV$ message to make the bandwidth reservation. To accommodate mobile hosts, we propose a reservation protocol, $MRSVP$, which is an extension of $RSVP$. It has the following characteristics :

1) It makes an active reservation from the current base station of the mobile host to the sender along which data flow will start immediately.

2) It also makes passive reservations along the data paths from all other base stations in the *mobility specification* of the mobile host to the sender.

We briefly outline the protocol $MRSVP$. The details of $MRSVP$ can be found in [13]. Let S be the *mobility specification* of a mobile host m. We assume that the underlying routing protocol is the IETF Mobile-IP routing protocol[11], in which data packets are routed to a mobile hosts via mobility agents, *home agent* and *foreign agent*. The protocol works as follows:

1. The sender sends a $PATH$ message to m or the multicast group address.
2. On receiving a $PATH$ message for the first time, m will send the flow specification to all mobility agents in S. If the reservation is for a unicast flow, m will send S to its home agent. If the reservation is for a multicast flow, m will ask the mobility agents of S to join the multicast group. Then it will send a $RESV$ message to the sender to make an active reservation. On receiving subsequent $PATH$ messages, it will only send the $RESV$ message to the sender.
3. On receiving S, the home agent will tunnel all $PATH$ messages for m to all mobility agents in S.
4. On receiving a $PATH$ message for a non-local mobile host, a mobility agent will send a $RESV$ message to the sender to make a passive reservation.
5. When m detects a significant change in the data flow rate, it will send the new flow specification to all mobility agents in S.

7 Conclusion and Future Work

In this paper we have investigated the problems of providing real-time service guarantees to the mobile hosts in an Integrated Services Packet Network. We presented a service model for real-time services to mobile hosts that can accommodate highly adaptive applications as well as applications which requires mobility independent QoS guarantees. We proposed an admission control scheme for this model that can achieve very high utilization of network bandwidth.

We are currently running simulation experiments to check the validity of our service model and its implementation scheme. Some of the issues we are currently looking at are, handoff techniques for reducing QoS degradation rate, the different ways to reduce handoff failure and the relationships among network utilization, handoff failure rate, QoS degradation rate and the degree of multiplexing.

References

[1] Acampora, A. S. and Naghshineh, M., An Architecture and methodology for Mobile-Executed Handoff in Cellular ATM Networks., IEEE JSAC, Vol 12, No. 8, October 1994.

[2] Bakre, A. and Badrinath, B. R., Handoff and system support for indirect TCP/IP, In Proc. of the 2nd Usenix symposium on mobile and location independent computing, pp. 11-24, April 1995.

[3] Balakrishnan, Hari, et.al., A comparison of mechanisms for improving TCP performance over wireless links, In Proc. SIGCOMM'96, August 1996.

[4] Clark. D.D., Shenker. S. and Zhang. L., Supporting Real-Time Applications In An Integrated Services Packet Network : Architecture and Mechanism., Proc. SIGCOMM '92, 1992.

[5] Demers. A., Keshav. S. and Shenker. S., Analysis and simulation of a Fair Queuing Algorithm, Proc. SIGCOMM '89, Sept. 1989

[6] Jamin. S., Danzig. P.B., Shenker. S and Zhang. L., A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks., Proc. SIGCOMM '95, 1995.

[7] Lee, K., Adaptive Network Support for Mobile Multimedia, In Proc. of the 1st Annual International Conference on Mobile Computing and Networking, pp. 62-74, November 1995.

[8] Levine, D. A., Akyldiz, I. F. and Naghshineh, M., The Shadow Cluster Concept for Resource Allocation and Call Admission in ATM-Based Wireless Networks, In Proc. of the 1st Annual International Conference on Mobile Computing and Networking, pp. 62-74, November 1995.

[9] Lu, S. and Bharghavan, V., Adaptive Resource Management Algorithms for Indoor Mobile Computing Environments, Proc. SIGCOMM'96, August 1996.

[10] Parekh. A. K., A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks., PhD Theses, MIT, LIDS, Tech. Report LIDS-TR-2089 1992.

[11] Perkins, Charlie, IP Mobility Support, IETF Draft, May 1996.

[12] Singh, S., Quality of Service guarantees in Mobile Computing, To appear in Journal of Computer Communications.

[13] Talukdar, A. K., Badrinath, B. R. and Acharya, A., Admission Control Algorithms and Reservation Protocol for Supporting Mobile Hosts in an Integrated Services Packet Network, Dept. of Computer Science Technical Report, Rutgers University.

[14] Zhang, L., Deering. S., Estrin. D., Shenker. S. and Zappala. D., RSVP: A New Resource ReSerVation Protocol. IEEE Network Sept. 1993.