Abstract

Mobility is one of the new services that have been introduced due to the development and poularity of the Internet. However, the current Internet lacks quality of service support. This article presents an end-to-end QoS architecture for roaming terminals. This architecture is based on Mobile IP Reservation Protocol (MIR) [10], Hierarchical Mobile IP (HMIP) and Diffserv. MIR operates within the wireless environment inside a HMIP domain while Diffserv mechanisms provide end-to-end QoS. MIR addresses the problem of bandwidth and reservation in order to provide users of a shared medium with a guaranteed bandwidth. Resources are reserved in the cells where mobile nodes are likely to go.

Keywords— Mobility, Quality of Service, Diffserv, Mobile IP, HMIP.

1. Introduction

The major drawback of the current Internet is the lack of Quality of Service support. However, as the Internet evolves from a research tool to a business tool, new requirements and pressures are being placed on Internet protocols to support business needs. At the forefront of these requirements is what is commonly referred to as QoS, whose goal is to support services that have specific qualitative needs. Mobile environments are more complicated to deal with than fixed environments since they introduce new constraints such as mobility, reliability, throughput, etc.

We present a new protocol, MIR (Mobile IP Reservation Protocol) which addresses the problem of bandwidth allocation and reservation so as to provide users of a shared medium with a guaranteed bandwidth. In order to provide end-to-end QoS, MIR is integrated in a HMIP/Diffserv environment.

This article is organized as follows: related work is presented in section 2; in section 3, we describe MIR protocol, the results obtained through simulation and modeling, and possible enhancements; section 4 depicts how HMIP and Diffserv are combined with MIR to provide end-to-end QoS; finally, we conclude and present future work.

2. Related work

2.1. Mobile QoS support

Talukdar [16] proposes frameworks and modifications to RSVP [2] that ensure good call dropping probabilities for mobile hosts. This scheme requires additional service classes, major changes to RSVP, a knowledge of the mobility profile of users and high bandwidth overheads for control signalling. This work assumes that the mobility of a user is predictable so that a mobility specification can be defined. This specification consists of the set of locations the mobile host is expected to visit during the lifetime of a flow. Three service classes are defined: MIG (Mobility Independent Guaranteed service), MIP (Mobility Independent Predictive service) and MDP (Mobility Dependent Predictive service). As long as terminals conform to this mobility specification, users who subscribe to the mobility independent service classes are not affected by host mobility; the others may experience changes in their QoS parameters when handoffs occur. However, to provide good service guarantees to mobility independent classes, resources are reserved all along the possible paths which may be used during the connection lifetime. To improve network utilization, mobility dependent service classes may use the reserved but unused mobility independent resources.

Awduche [1] enhances RSVP to mobile environments. The approach is similar to the one described in [16]. Reservation classes are extended and three new classes are defined: committed, quiescent and transient reservations. Quiescent reservations are unused resources which can be allocated by others (transient reservations) as long as the reservation remains quiescent. The reservation is activated when a mobile node enters a cell in which a quiescent reservation was made for it. The quiescent reservation then pre-empts the other terminals which had allocated it’s quiescent resources. However this scheme uses mobility management agents to maintain mobility profiles and to send signals to virtual receivers which set up the quiescent reservations. Other proposals to support QoS in mobile environments based on RSVP include [8] and [9]. However, these approaches are not scalable.
When congestion occurs, the network actually discards packets randomly to reduce congestion. Because the network chooses packets at random, those within higher bandwidth flows have a greater probability of being dropped than those within lower bandwidth flows. Discarding packets quickly reduces congestion by reducing the traffic of the biggest consumers of resources. Unfortunately, the packets dropped are likely to belong to users who pay to have greater bandwidth. Thus, [14] proposes that users specify at the beginning of a flow their preferences concerning possible data losses. These specifications are called loss profiles. A profile is used to allocate bandwidth to the users in a cell. For example, if a mobile node uses an audio application, it can afford to lose some packets regularly. When compressed video is used, random losses would forbid good decoding; thus, it is better to lose a burst. The bandwidth must therefore be allocated according to those needs and the packets are not lost randomly. However this approach is not scalable and does not fit if the users do not wish to have a degraded service. Moreover, latency and jitter, which are relevant especially in the case of duplex audio traffic are not addressed.

### 2.2. Micro mobility approaches

Mobile IP protocol has been developed for computers that are moving. In the standard Mobile IP, mobile nodes have to report their every movement in the foreign network to their home networks. This causes huge amount of signalling traffic and disturbing latency during handoffs. Because of these problems, several protocol proposals have been defined to solve this so-called micro mobility problem. In all of these solutions the home network does not have to know the exact location of the mobile node. Instead the home network only has to know in which visited network the mobile node is located and the local micro mobility is managed inside the visited network. These approaches include Hierarchical Mobile IP ([4] and [5]), Cellular IP [3] and Hawaii [15]. However, although handoff latency is reduced, these schemes do not cope with some quality of service issues like seamless bandwidth reservations with a roaming terminal.

### 3. MIR (Mobile IP Reservation Protocol)

#### 3.1. CLEP (Control Load Ethernet Protocol)

Ethernet or IEEE 802.11 [7] do not guarantee any type of quality of service. It is impossible for a specific flow to have a fixed throughput since this type of medium guarantees some kind of fairness. The Control Load Ethernet Protocol (CLEP) described in [6] is an implementation of the Controlled Load service over Ethernet defined by Wroclawski in [17]. It provides the client data flow with a quality of service approximating the quality of service this flow would receive on an unloaded network. The load control is done using token bucket filters on outgoing interfaces of network elements.

![Figure 1. Architecture of a network element](image)

CLEP’s load control is done using token bucket filters on outgoing interfaces of network elements (cf. Figure 1). Network elements use a distributed protocol to manage the parameters of the token bucket filters and efficiently use the link. All flows (best effort and privileged flows) use the controlled-load service to control packets admission in the Ethernet network. Packets are admitted only if there is enough bandwidth for them. This admission control can be achieved with token buckets filters for each flow.

There are two queues for best effort traffic – instead of one for privileged flows – because standard best effort traffic and flows needing very low rates but a high priority are distinguished. The latter have strict time or reliability constraints. Each network element assigns a rate to best effort and privileged flows. These parameters change when a new element appears in the network or when a new reservation is needed. Every time a parameter varies, the new available bandwidth is computed again.

This protocol has been implemented and simulated. It shows a clear improvement of the quality of service of the applications [6]. Performance on a bus technology are very satisfying. The rates of the applications remain stable and there are very few packet losses even when other stations start emitting datagrams.
3.2. MIR (Mobile IP Reservation Protocol)

MIR [10] is an extension of CLEP to mobile environments; it is intended to provide statistical QoS guarantees to mobile users. Although CLEP provides hard bandwidth guarantees to the sources, it is impossible to give absolute guarantees in a mobile environment since the mobility of the terminals is not bound. If all the users which were admitted in separate cells move in the same area, they will compete to use the same resources. However, if all these users benefit from a high level QoS (if they have negociated a certain bandwidth where they were admitted in the network), they will probably not be able to benefit from this high level of QoS if they move into the same cell. Thus, QoS in such an environment can only be provided statistically: for example, a privileged flow can benefit from a certain bandwidth with hard guarantees if the terminal is not mobile and with a low probability of being blocked if the terminal executes one or more handoffs.

However, it must be noted that the handoff blocking probability is a function (but is not) the flow blocking probability. Actually, a flow will experience many handoffs during its lifetime. Each handoff is subject to a specific blocking probability so that the flow’s blocking probability is higher than that of a single handoff. Figure 2 highlights this observation. It represents the flow blocking probability as a function of the number of handoffs when the handoff blocking probability is fixed.

MIR is distributed and allows each cell to be managed separately depending on the mobility of the terminals in that cell. A cell can thus adapt to the mobility requirements of the terminals and divide its bandwidth accordingly. MIR is based on a model where Mobile IP and IEEE 802.11 are used. Each cell is an IP domain (which means that each base station is also a foreign agent) but we could also design a protocol in which many cells could belong to the same IP domain. IEEE 802.11 is also used as a general model for transmitting data on a wireless medium.

Adapting CLEP to mobile environments means taking into account resource reservation during handoffs. To offer end-to-end QoS guarantees, two assumptions can be made: if the wired part of the network is not considered as a bottleneck since it may offer a greater bandwidth than the wireless cells, only the wireless cells need to run MIR and can be provided with quality of service guarantees.

3.3. Resource Allocation Policy

In order to provide a higher priority to handoffs rather than new flows, the bandwidth of each cell should be divided into two parts. When the load is smaller than a threshold S1, new flows and privileged handoffs are accepted. When the load is greater than S1, new flows are no longer accepted. Passive reservations should be made for flows from other cells which might execute a handoff. A passive reservation is a reservation made by a foreign agent for a mobile that may register with this foreign agent (cf. Figure 4). As soon as a mobile hears an advertisement from a new foreign agent, it should ask this agent to perform a passive reservation for its privileged flows. This reservation is not actually used. It becomes active only when the terminal executes a handoff in the cell handled by the new foreign agent. Thus, an active reservation is a reservation which is currently used to send data. A passive reservation becomes active when the terminal executes a Mobile IP handoff. This scheme is used to establish a QoS connection as soon as a Mobile IP handoff occurs since the tokens are reserved in advance by the foreign agent of the new cell. Since passive reservations may never become active, it is possible to reserve more bandwidth than what is actually available (threshold S2 in Figure 3).

MIR has been modelled using Markov chains, as described in [10]. This model is a three dimensional chain. In which the behavior of a single cell is studied. All the events that may occur are modelled: a new connection (privileged or best effort), a passive and an active handoff. Each state of the model has three dimensions: the first (i) is the number
of best effort tokens, the second (j) is the number of privileged active tokens and the last (k) is the number of passive reservation tokens. Once the Markov chain was built, the dropping probabilities were computed using the GTH algorithm, which calculates the stationary distribution of the chain numerically.

Bandwidth Available Bandwidth

Threshold S1

Privileged and best effort flows (new flows)

Passive Reservations

Active Reservations

Figure 3. Bandwidth allocation in a cell

The protocol has proven its ability to handle mobile QoS in a local environment. Once a privileged connection has been accepted, it can roam in the network with a low probability of being dropped. Figure 5 was obtained with a highly loaded network. When S1 increases, new call blocking probabilities decrease (since more new connections are accepted in a cell) but the handoff dropping probabilities increase (Figure 5). Moreover, handoffs have a lower blocking probability than new connections, which was one of the protocol's goals. When S2 increases, the blocking probability of a passive handoff decreases whereas the blocking probability of an active handoff increases.

Since the load of the network is not constant, these thresholds must be managed dynamically according to the network's needs.

MIR privileges handoffs rather than new connections. However, any privileged flow may be dropped during a handoff (with a low probability). It is impossible to guarantee a flow will not be dropped unless the terminals respect a mobility specification. Predicting terminal mobility allows to make reservations all along the path the mobile will follow as proposed in [16]. However, we believe it is impossible to know in advance where a mobile will go. Moreover, this policy under-utilizes the global bandwidth of the network, unless the unused reserved resources can be used by other connections with a lower priority. But this scheme introduces extra management overhead which is not justified since in case of high mobility and priority, it becomes impossible for many connections to benefit from a premium service. On the other hand, MIR deals properly with high mobility networks since its thresholds vary dynamically.

Figure 3. Bandwidth allocation in a cell

Figure 4. Negotiation of QoS parameters while the mobile moves

Blocking probability as a function of S1 (Bw=15, S2=22)

Figure 5. Dropping probabilities when Bw=15 and S2=22

3.4. Managing the thresholds

To achieve the best possible performance, each cell should manage the values of the thresholds dynamically. A Foreign Agent counts the number of handoffs which occurred during a certain time to measure mobility in a cell. This helps adjust the thresholds. S1 is determined by the handoff dropping probability the network should offer and S2 by the number of passive handoffs which do not become active. This can easily be evaluated if a mobile advertises its new foreign agent of its previous location (previous foreign agent). Each foreign agent counts the number of passive handoffs it granted – ps\textsubscript{ok} – and refused – ps\textsubscript{ko}. The number of active handoffs is also measured. If the passive dropping probability is too high, S2 should be increased. If
the active dropping probability (the probability that an active reservation is refused) is too high, S1 should be decreased.

In order to enhance threshold management and cope with the specific requirements of each cell, the value of S2 should depend on the cell from which the mobile comes. We use the following scheme: a foreign agent should count the number of reservations it grants – \( act_{ok} \) – or denies – \( act_{ko} \) – together with the current foreign agent of the mobile requesting this reservation. Every foreign agent therefore knows the percentage of passive flows coming from a specific foreign agent and which try to become active. This percentage (A) is used to compute the value of S2 associated to this particular foreign agent.

\[
A = \frac{act_{ok} (FAi) + act_{ko} (FAi)}{psv_{ok} (FAi)} \quad \text{and} \quad B = \frac{act_{ok \_total} + act_{ko \_total}}{psv_{ok \_total}}
\]

If \( n\% \) of the flows become active, then the value of S2 should be greater than if \( m\% \) of the flows become active (when \( m>n \)). A nominal value of S2 is determined by the total number of passive and active handoffs (i.e., the handoffs form all the neighbor FAs). It corresponds to a percentage (B) of flows becoming active. If the mobile comes from a foreign agent unknown or for which we have too few statistics, the nominal value is used. Otherwise, the admission threshold is \( S2 + (B-A) \cdot (S2-R_{tot}) \) where \( R_{tot} \) is the maximum number of tokens which can be distributed in a cell.

### 3.5. Increasing the reliability of the protocol – protecting broadcast packets

The transmission of data packets on a wireless LAN is unreliable. This creates some performance problems in the transport layer for unicast packets. Usually, MAC level retransmissions are used to deal with this problem. For example, in IEEE 802.11, the Data/ACK handshake ensures the packets were received properly. But for broadcast and multicast packets, the transport layer cannot include any acknowledgment and retransmission mechanism due to the non defined number of recipients. The MAC layer cannot either provide such a mechanism for the same reason.

To overcome packet losses on the medium, acknowledgments and retransmissions are used. Acknowledgments are embedded in layer two so that they are guaranteed not to collide. Moreover, retransmissions can be launched very quickly since it is not necessary to wait for an acknowledgment from the recipient. The only restriction is that this assumes a unique receiver is concerned.

To overcome losses due to packet collisions on the medium, RTS/CTS is used in IEEE 802.11. This ensures the medium is free while data is being sent and solves the hidden node problem.

When using MIR, some data, although it is best effort traffic, is critical for a good operation of the protocol. These packets are the ones used to compute depths and rates of the network elements’ token buckets (MIR packets) and also Mobile IP registration messages. If they are considered as normal best effort traffic, they might experience high delay and even high dropping rates. The latter can even be caused in the network elements themselves because packets are dropped from the best effort queues when these queues are full. This traffic can thus be handled in two ways:

- **best effort traffic** with a high priority: it is in a separate best effort queue which is served before the rest of the best effort traffic,
- **privileged traffic**: a new privileged flow is defined to handle this specific traffic.

We have simulated the architecture in Figure 6 with NS [11]. FA1 is MH0’s Home Agent and FA2 is MH1’s and MH2’s Home Agent. At \( t=100 \), flow udp0 starts and reserves 300000 tokens at 150. The other flows are best effort flows. Then, MH0 starts to move towards cell 2. When the handoff occurs, udp0’s reservation is maintained in cell2; udp1 and udp2 reduce their traffic accordingly.

When broadcast Mobile IP packets have no protection (Figure 7), many Mobile IP timeouts occur and the best effort flows are irregular. It is not the case in Figure 8. However, in both cases, we have to cope with broadcast packets which are not acknowledged nor protected against
collisions. When cells are highly loaded, it becomes necessary to protect broadcast packets against collisions to avoid frequent Mobile IP deregistrations. This can be done using a classical RTS/CTS handshake with a well chosen terminal:
- if the sender is a mobile node, it should perform this handshake with the base station,
- if the sender is a base station, it should perform the handshake with any mobile node which is currently registered with it.

In this case, the node might already be gone but not yet deregistered, so if the base station does not complete the handshake properly, it should chose another station to do the handshake after a timeout has expired. Finally, if no terminal is registered with the base station, the broadcast packets should not be protected by RTS/CTS and must not be acknowledged.

Therefore, all the mobile nodes in the cell are very likely to hear either an RTS or a CTS packet. If a collision or a packet loss is detected, the broadcast packet will be immediately retransmitted by layer 2 mechanisms. Mobile IP and MIR packets have thus a very low probability of being lost.

3.6. Using bandwidth efficiently – passive reservations are re-used to send best effort traffic at the base station

A foreign agent performs passive reservations for a mobile’s privileged flow. If the mobile is the source, then the tokens are transferred from the foreign agent to the mobile when it performs its handoff. If the mobile is a recipient, the foreign agent keeps the passive tokens and uses them to send active traffic only once the mobile has executed a handoff. However, passive tokens are not used at any time to send traffic. They only act on best effort traffic in the cell in order to reduce it, so that the tokens are ready once the handoff is executed.

One way to optimize this scheme is to actually use passive tokens to send downstream best effort traffic for the base station. This is completely transparent for other terminals since a passive reservation acts like an active reservation, the only difference being that in the first case, no traffic is sent (whereas in the second case, the tokens are used to send privileged traffic). Moreover, this can drastically improve the performance since the base station is likely to be sending more traffic than any other station in the cell. This scheme allows base station best effort traffic to have a higher priority than other best effort stations, with no impact on the reservation mechanism. As soon as a flow becomes active, tokens are immediately used for it and are no longer used for best effort traffic. This is done without having to modify the token buckets of the other terminals. This means that when a flow becomes active, best effort parameters need not be computed again with the control load algorithm.

If this was not used, the base station would be considered as any other terminal and its aggregate best effort throughput would be equal to that of any other terminal in the cell. Increasing the base station’s best effort priority is obviously more efficient, since wireless traffic is notoriously asymmetric.

4. Providing statistical end-to-end QoS guarantees – a network architecture using DiffServ

End-to-end QoS cannot be provided with a simple MIR/CLEP mechanism for scalability reasons. A scheme based on DiffServ [12] or MPLS must be used. In this
section, we will describe a MIR/CLEP/HMIP/Diffserv architecture which provides end-to-end QoS.

The Diffserv architecture comprises a number of different areas. Before traffic is grouped together into an aggregation, the packets belonging to the aggregation must be identified. Within the Diffserv architecture, a router must be able to measure, shape and drop packets in a flow. Another part of the architecture describes the distinction between the edges and the core of the Diffserv network. Diffserv also defines the relationship among multiple administrative domains, which are specified in service contracts.

When integrating Mobile IP and Diffserv, many problems arise, for example: lack of dynamic configuration, network provisioning (will enough resources be available in the new location), routing (what Service Level Agreements (SLA) should be set up) and flow identification (because of the mobile IP tunnels).

We propose to use Hierarchical Mobile IP to cope with this problem. There are two main advantages in using such an approach: first, micro-mobility solutions provide improved handoff (lower latency) because as long as a mobile stays within a domain, it appears to be static for its own Home Agent. HMIP introduces a GFA (Gateway Foreign Agent) at the top of the hierarchy. The mobile can roam within the hierarchy transparently for its home agent. It registers with its home agent only to update its binding lifetime or when it roams between domains.

To provide end-to-end QoS, the architecture described in figure 8 is used:
- MIR operates only on the wireless part of the network - (between the FA and the mobile node),
- CLEP and HMIP are used within a domain,
- Diffserv operates between domains.

When a mobile is roaming within a domain, the GFA with which it is registered is responsible for negotiating dynamically a SLA with the neighbor domains through a Bandwidth Broker (BB). When a mobile node is roaming within a visited network, it wishes to benefit from the same service it has in its home network. New SLAs have to be negotiated depending on the mobile node delivery mode (bi-directional, triangular or optimal routing).

It is rather difficult to establish new SLAs dynamically, this is why HMIP is used. Most of the time, a mobile node roams within a domain without updating the SLAs between the correspondent, the home agent and the visited domain. The Diffserv architecture requires the router to perform a number of functions to support the different service categories. A router must be able to look at each packet and identify the flow to which it belongs. Examining the source and destination IP addresses and possibly the source and destination port numbers provides a means for uniquely identifying the flow. This process of flow identification is commonly referred to as classification. A special case of classification is the ability to identify the aggregation to which a packet belongs, based on the DS field.

![Figure 9. Visited domain with a GFA, and a home network with HA](image)

MIR/CLEP nodes classify flows in order to determine what QoS they should get within a domain. At the same time, they can perform the Diffserv classification and simplify the task of edge routers.

5. Conclusion and future work

We designed a new protocol called MIR which was evaluated with a Markov chains model (in [10]) and simulation using NS. MIR is a distributed protocol which guarantees a certain quality of service to mobile users. In this mobility context, it is impossible to provide absolute guarantees in terms of bandwidth since the mobility of the terminals cannot be controlled. The protocol presented in this article ensures that the QoS level requested will be met with a high probability if the thresholds are adjusted properly. Flows can be differentiated according to their QoS requests and low blocking probabilities are achieved for the privileged flows. We proposed some enhancements to the protocol which will help use bandwidth more efficiently and adapt dynamically to the network’s needs. Finally, we proposed to integrate MIR within a HMIP and Diffserv architecture in order to provide end-to-end QoS to mobile terminals.

In the future, we will evaluate the signaling overhead and the performances of this architecture.
6. References


