Wireless Networks 9, 341–352, 2003

Quality of Service and Mobility for the Wireless Internet

J. ANTONIO GARCIA-MACIAS, FRANCK ROUSSEAU, GILLES BERGER-SABBATEL, LEYLA TOUMI and ANDRZEJ DUDA *
LSR-IMAG Laboratory, CNRS and Institut National Polytechnique de Grenoble, Grenoble, France

Abstract. Our paper explores the issue of how to provide appropriate quality of service mechanisms closely integrated with flexible mobility management in wireless local area networks. We consider them as access networks of choice for the high performance Wireless Mobile Internet. We present a hierarchical QoS architecture that extends *Differentiated Services* (DiffServ) to mobile hosts in a wireless environment. Our approach is based on controlling several parameters of a wireless LAN cell: the limited geographical span to ensure the same high bit rate for all hosts, the constrained rate of traffic sources to limit the use of the channel in function of the required QoS and the limited number of active hosts to keep the load sufficiently low. The QoS management is coupled with mobility management at the IP level. We use a micro-mobility scheme implemented in the IPv6 layer with fast hand-offs between adjacent cells. Micro-mobility avoids address translation, traffic tunneling, and enables fast hand-offs. We give some details of experiments to show the quality of service differentiation over the 802.11b network.

Keywords: QoS support in wireless access networks, micro-mobility management mechanisms, differentiated services *DiffServ*, 802.11 WLAN, IPv6

1. Introduction

Providing ubiquitous Internet access to mobile hosts becomes increasingly important because of new emerging applications: mobile information access, real-time multimedia communications, networked games, immersion worlds, cooperative work, and some others not yet invented. Many of such applications require better quality of service than the current *Best Effort*, however providing such quality of service to mobile hosts is a difficult problem because of the radio channel characteristics and complexity of mobility management. We focus on wireless local area networks such as IEEE 802.11 that have many advantages as access networks to the Wireless Mobile Internet: they provide higher nominal bandwidth (11 Mb/s) than the future UMTS and can easily be deployed as hot spots for high density areas. The purpose of this paper is to explore the issue of how we can provide appropriate quality of service mechanisms closely integrated with flexible mobility management in wireless local area networks.

The current approach to providing quality of service in the global Internet is based on *Differentiated Services* (*DiffServ*) [5]. Its principle is to classify and mark up the traffic at the entrance of the backbone network so that it can be processed differently in backbone routers and obtain different performance for each assigned class. Performance of *DiffServ* relies on sufficient provisioning of network resources in the backbone. This model also assumes that resources in access networks (the networks between a host and the backbone) are over-provisioned as usually it is the case for current local area networks (LAN). However, if a mobile host is connected to a wireless LAN such as IEEE 802.11 or Bluetooth, the radio channel becomes a critical part of the whole architecture and may severely affect the end-to-end performance. Although IEEE 802.11 provides a means for allocating a part of the radio channel bandwidth to some hosts (PCF – Point Coordination Function), we are interested in using the commonly available access method (DCF – Distributed Coordination Function) that is oriented towards fair sharing of the common communication channel. In this way, we just use 802.11 as any other available link for transferring IP packets. Our approach to providing QoS in such an environment is to extend the *DiffServ* model to wireless access networks so that we can provide consistent IP level quality of service to mobile hosts.

In the rest of the paper, we describe how the differentiated services model can be extended to a wireless LAN so that mobile hosts can benefit from differentiated performance classes in a similar way to wired networks. Providing QoS support in a wireless environment is not easy mainly because of the varying performance of the radio channel and the channel access method that shares the channel equally between all hosts. Our approach is based on controlling several parameters of the wireless LAN cell: the limited geographical span to ensure the same high bit rate for all hosts, the constrained rate of traffic sources to limit the use of the channel in function of the required QoS, and the limited number of active hosts to keep the load sufficiently low.

Each cell of a wireless LAN is managed by an *Access Router* that forwards packets between mobile hosts in a cell. Mobile hosts and access routers are provided with *DiffServ* mechanisms so that traffic sources can be constrained in a configurable manner. Mobile hosts use a lightweight in-band signaling protocol to request bandwidth allocations from the Access Router. Based on this information it configures the traffic shapers in function of the current allocation. The number of active sources is also subject to admission control. We

* Corresponding author.
E-mail: duda@imag.fr

use a hierarchical QoS architecture in which Access Routers manage fast changing local situations and cooperate with an *Edge Router* that fixes long term policies for Access Routers: admission control rules, mobility contexts, pre-reservation of resources.

Another issue concerns mobility management that should be coupled with QoS management. As we propose to manage QoS at IP level, we have also chosen to manage mobility at the same level. For such mechanisms to be efficient, we need an efficient mobility management scheme optimized for QoS. So we need to rethink our approach to mobility management. The traditional approach of Mobile IP provides a solution to global mobility [19,20], however, it does not take into account QoS requirements. In fact, *Home Agents* and *Foreign Agents* allows to deliver traffic to a mobile host by using indirection and tunneling in both cases: limited local movements between adjacent wireless cells and global world-wide mobility. Triangular routing, address translation, and complex interaction between agents make Mobile IP unsuitable for integration with quality of service support in a wireless LAN environment [10,12].

We propose to limit the scope of mobility management to the local case and make it efficient enough so that we can couple it with QoS management. This approach follows recent work on micro-mobility whose rationale comes from the observation that most of the mobility is limited to local areas, exceptional global movements can be dealt with as nomadicity by acquiring new addresses. Integration of mobility management with QoS makes it possible to take into account a richer set of parameters to initiate hand-offs. For example, the decision to switch to another cell can be made not only based on the signal to noise ratio, but also on the current load in a cell, the level of available resources, the state of pre-reservations, and on some administrative policies. We use a micro-mobility scheme implemented in the IPv6 layer with fast hand-offs between adjacent cells. Micro-mobility avoids address translation, traffic tunneling, and enables fast hand-offs. Coupled with the QoS management, it contributes to the overall end-to-end performance.

We start with the analysis of the 802.11 wireless LAN (section 2), then we discuss the related work (section 3) and we present the hierarchical QoS architecture that provides differentiated services to mobile hosts in a wireless environment (section 4). We also present the micro-mobility scheme integrated with QoS management (section 5). Furthermore, we give details of implementation and experiments that show how we achieve differentiation of services over the 802.11 wireless LAN (section 6). Finally, we present some conclusions (section 7).

2. Quality of service in IEEE 802.11b networks

Our goal is to provide appropriate quality of service mechanisms closely integrated with flexible mobility management in wireless local area networks. A wireless LAN environment has specific characteristics that make it difficult to provide an adequate quality of service. The IEEE 802.11 standard defines two access methods: the Distributed Coordination Function (DCF) that uses CSMA/CA to allow for contended access to the wireless media and the Point Coordination Function (PCF) providing for uncontended access via arbitration by a Point Coordinator, which resides in the Access Point. The DCF method provides best effort type of service whereas the PCF guarantees a time-bounded service. Both methods may coexist: a contention period follows a contention-free period. The PCF is the method especially well suited for real-time traffic, unfortunately it is not implemented in current 802.11 products. Moreover, simulation studies [15,16] show that PCF has rather poor performance compared to other control methods such as EDCF (Enhanced Distributed Coordinator Function) defined in the scope of IEEE 802.11e standard. Other research efforts target at providing some QoS support at the MAC level considers modification of some parameters of the DCF method [1,3].

Our approach is quite different – we provide QoS at IP level and use the best effort type DCF method at the MAC level. We can build such support by controlling several parameters of wireless LAN cells. To explain this, we start with analyzing the characteristics of the DCF access method from the performance point of view.

First of all, the DCF access method raises the problem of the access overhead that increases with the number of active hosts. The access method is based on the CSMA/CA principle in which a host wishing to transmit senses the channel, waits a period of time (DIFS – Distributed Inter Frame Space) then transmits if the medium is still free. If the packet is received correctly, the receiving host sends an ACK frame after another period of time (SIFS – Short Inter Frame Space). If the ACK frame is not received by the sending host, a collision is assumed to have occurred and the data packet is transmitted again after waiting another random amount of time.

If a single host transmits a data frame, the transmission time will be the following (we suppose 802.11b with the bit rate of 11 Mb/s [2] and we neglect propagation times; this analysis follows [4,8,23]):

$$T_{\text{single}} = t_{\text{pr}} + t_{\text{tr}} + SIFS + ACK + DIFS, \tag{1}$$

where t_{pr} is the preamble time (144 µs), t_{tr} is the frame transmission time (size/bit rate), $SIFS = 10$ µs, ACK is the ACK transmission time (210 µs), and $DIFS = 30$ µs. If we assume the frame size of 1500 bytes of data (data frame of total 1534 bytes), proportion r of the useful bandwidth in this case will be:

$$r = \frac{t_{\text{tr}}}{T_{\text{single}}} = \frac{1.11 \text{ ms}}{1.51 \text{ ms}} = 0.735. \tag{2}$$

So, a single host sending over a 11 Mb/s radio channel will have the useful bandwidth of 8.08 Mb/s.

If there are multiple hosts attempting to access the channel, one host may sense busy channel or collide with the transmission of another host. In such cases, the host executes the exponential backoff algorithm to wait a random interval distributed uniformly between $[0, CW - 1] \times slot$, $CW_{\min} = 32$,

$CW_{\max} = 1024$, and $slot = 20$ µs (these parameters are for the Direct Sequence Spread Spectrum physical layer). Each time the host chooses a slot and happens to collide, it will double CW up to $CW_{\max}$. So, if there are m hosts in a cell, the efficiency will degrade with the number of hosts because of collisions. The transmission time experienced by a single host when competing with $m-1$ other hosts will be increased by some interval $w(m)$ that accounts for the time spent in collisions and backoff procedure (the analytical formulae for this duration is difficult to derive [8]):

$$T_{\text{multiple}}(m) = t_{\text{pr}} + t_{\text{tr}} + SIFS + ACK + DIFS + w(m). \quad (3)$$

This means that the proportion of the useful bandwidth as seen by a single host will also depend on the number of hosts:

$$r(m) = \frac{t_{\text{tr}}}{T_{\text{multiple}}(m)}. \quad (4)$$

For example, if we assume 1500 bytes of data and one collision on the average, i.e., $w(m) = 0.31$ ms, the efficiency experienced by a single host decreases to 0.61, and the useful bandwidth to 6.71 Mb/s.

To evaluate the overhead of the contention, we have measured the performance of a 802.11b cell. Figure 1 presents the useful bandwidth measured at the transport layer experienced by a single host in a cell with two or three competing hosts (all hosts tries to send as much data as possible – the traffic is generated by greedy TCP sources). For one host and the packet length of 1500 bytes, the bandwidth is roughly 6.5 Mb/s. When there are two hosts, it decreases to 5 Mb/s and it further degrades to 2 Mb/s in the case of three hosts. These figures are much smaller than the limits analyzed above, because they include the overhead of all protocol layers.

We can conclude from this analysis, that if we want to manage bandwidth allocations in a 802.11 cell, we have to take into account the fact that its useful bandwidth strongly depends on the number of active hosts.

Fact 1. To provide quality of service over the 802.11 link, the number of hosts allowed to use the channel should be limited.

Another problem of 802.11 is related to the performance of the radio channel that is time and location dependent due to factors such as the distance between the source and the destination, signal interference and fading. Some wireless LANs make use of different modulation and error control techniques so that these factors manifest themselves as variation in bandwidth perceived at the network layer. However, the most popular 802.11 products do not provide such a support. Instead, they are able to degrade the bit rate when repeated frame drops are detected (e.g., WaveLAN can degrade from 11 Mb/s to 5.5, 2, or 1 Mb/s). However, as the channel access probability is equal for all hosts, hosts that send at low rates penalize hosts that use the high rate. Table 1 shows the measured performance of a 802.11b cell with two hosts that use different rates (the throughput is measured at the TCP layer). We can see that the low rate host penalizes the high rate host and both hosts obtain a small proportion of the nominal bandwidth.

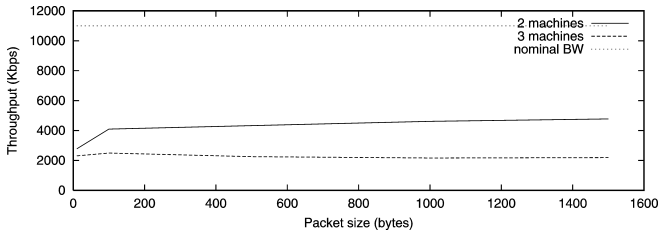


Figure 1. The useful bandwidth of a host in a shared 802.11b cell.

Table 1
802.11b performance, hosts of different rate.

Host rates	Measured throughput
11 Mb/s, 11 Mb/s	5 Mb/s
11 Mb/s, 1 Mb/s	0.84 Mb/s

This means that if we want to provide a satisfactory QoS behavior, we have to restrict the usage of the 802.11 link to an area in which all hosts can send at the same high rate, e.g., 11 Mb/s.

Fact 2. To provide quality of service over the 802.11 link, the geographical area in which mobile hosts communicate should be limited so that all hosts use the same high bit rate.

The DCF access method of 802.11 is designed to provide mobile hosts with a fair share of the radio channel capacity. If we want to provide different performance behavior to traffic sources at mobile hosts, we need to constrain them in a configurable way so that sources of low priority benefit from different resource allocations than high priority ones. For example, we can use traffic shapers to constrain sources at mobile hosts and keep in this way the aggregated traffic lower than the available link capacity.

Fact 3. To provide quality of service over the 802.11 link, traffic sources should be constrained by configuring traffic shapers in hosts to obtain desired QoS effects.

In addition to that, QoS management should be reactive enough to adapt to varying conditions in a cell such as starting or terminating a traffic source, arrival or departure of a host in/from a cell. Based on performance conditions in a cell and in its neighbors we can also make proper decisions on whether a mobile host should hand-off to adjacent cells or not.

3. Related work

The problem of providing quality of service in IP networks has received considerable attention. However, supporting QoS over wireless links and integrating QoS mechanisms with mobility is still an open problem addressed by the IETF community [17]. Recent surveys analyze different issues and identify research directions [10,11]. Our analysis follows their conclusions and applies them to the problem of providing QoS for the Wireless Internet based on the *DiffServ*

architecture. Several authors have investigated a completely different approach to QoS differentiation in 802.11 networks by extending or modifying the MAC layer [1,3]. However, these solutions cannot apply to the networks that use current 802.11 products.

The traditional approach to mobility based on *Mobile IP* provides a solution to global mobility [19,20], however, it does not take into account QoS requirements. In fact, *Home Agents* and *Foreign Agents* allows the delivery of traffic to a mobile host by using indirection and tunneling in both cases: limited local movements between adjacent wireless cells and global world-wide mobility. Triangular routing, address translation, and complex interaction between agents make Mobile IP unsuitable for integration with quality of service support in a wireless LAN environment [11,12,18].

Our mobility management scheme is similar to those studied in the HAWAII project [21]. HAWAII proposes four schemes: MSF, SSF, UNF, and MNF. In MSF, hand-off is initiated via the old base station and results in transient loops, whereas SSF requires more descriptive routing tables. UNF and MNF rely on the capacity of the mobile host to communicate with both base stations: the old and the new one. When a mobile host hand-offs into a new cell, routing tables in routers involved in the movement are modified starting from the new base station. The HAWAII mobility schemes have been only validated by simulation and they do not provide any specific QoS support. If integrated with QoS management, the schemes allow the mobile host to start using resources in the target cell without any admission control.

Cellular IP is another approach for handling micro-mobility [9,22]. However, it requires specialized routers in a local domain and its functioning relies on a gateway acting as a Mobile IP Foreign Agent. The gateway is a critical element on which depends the reliability of the whole domain. Moreover, Cellular IP only supports best effort traffic.

Our signaling protocol described later is inspired by Insignia that defines a IP-based QoS framework for mobile ad-hoc networks [14]. Insignia is based on in-band signaling and soft-state resource management to support highly dynamic environments with time varying network topology, node connectivity, and end-to-end QoS. Its simple QoS model is based on providing mobile hosts with adaptive services: the allocation of a minimum bandwidth and the possibility to enhance to some maximum bandwidth.

4. Hierarchical QoS architecture

Based on the analysis of the 802.11 wireless link, we can address the issue of the QoS architecture for such access networks. What kind of a QoS model can we provide over a wireless LAN link? Current approaches to IP quality of service include the *IntServ* [7] and *DiffServ* architectures. The *IntServ* architecture defines mechanisms for per-flow QoS management and provides tight performance guarantees for high priority flows. It uses RSVP as signaling protocol. Unlike the *IntServ* model, the *DiffServ* architecture defines aggregated behavior for a limited number of performance classes for which only statistical differentiation is provided. *DiffServ* does not require any signaling protocol, resource allocation being defined statically by means of SLA (*Service Level Agreements*) between administrative domains.

Although it would be possible to build QoS support for wireless access networks based on the *IntServ* architecture, we think that *DiffServ* is a better candidate for several reasons. First, the characteristics of the wireless LAN environment preclude any tight bounds on performance measures, for example it would be useless to reserve sufficient resources via RSVP to guarantee a worst case delay for a high priority flow, if we cannot guarantee the delay on the wireless link. Instead, we think that a QoS model that does not define any absolute guarantee and only proposes a statistical differentiation fits better (this point of view is also shared by several participants of the IAB Wireless Internetworking Workshop [17]). Furthermore, using RSVP as a signaling protocol rises several issues including signaling overhead and setup delays on roaming events [17]. Finally, as *DiffServ* will be deployed in the global wired Internet, extending *DiffServ* to wireless access networks will provide consistent end-to-end QoS behavior without the need for mapping between QoS classes of different models.

So, we propose to build the QoS architecture for wireless access networks on the *DiffServ* model and use basic mechanisms of *DiffServ* such as traffic shapers to constrain sources at mobile hosts. We briefly describe the architecture of *DiffServ* and the implementation that we use for providing QoS to mobile hosts.

4.1. Differentiated Services

The architecture of *DiffServ* distinguishes two parts: the core network composed of one or several ISPs, packet forwarding done by core routers and the access network connecting end hosts to an edge router (cf. figure 2). Performance agreements between different administrative domains (SLA – *Service Level Agreements*) allow to statically reserve sufficient resources to support statistical performance guarantees of different traffic classes. Core routers forward packets according to different BA (*Behavior Aggregates*) – QoS classes that group flows of similar properties. Performance perceived by each class depends on the type of processing at core routers specified in a PHB (*Per Hop Behavior*). Edge routers perform classification of the incoming traffic and marking according to application types, source and destination addresses or ports or other criteria. Incoming traffic is checked against a TCA (*Traffic Conditioning Agreement*), a profile of the traffic defined in the SLA. Traffic exceeding a given TCA can be dropped, marked as out of profile, or marked with a lower priority class.

We use an implementation of the *DiffServ* edge and core router functions developed in a Next Generation Internet project [13]. It is based on an IPv6 stack and has slightly different properties than those defined by IETF. The main difference is the number of AF classes: we define only one

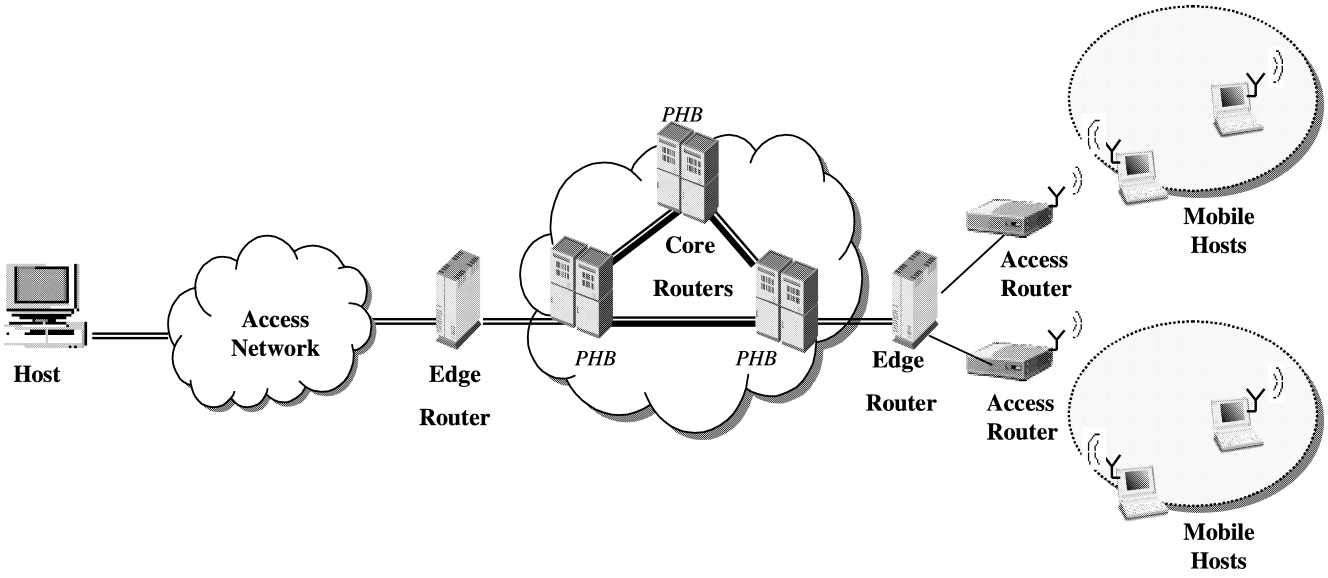


Figure 2. Differentiated services for wireless access networks.

AF class instead of four in *DiffServ* and two spatial priorities (drop probability thresholds) instead of three. This simplification appears to us as a right tradeoff between a sufficient choice of different services and user readability – it is difficult to make clear distinction between all 14 classes proposed in *DiffServ*. The second difference concerns buffer management techniques. The AF queue is managed using the PBS (*Partial Buffer Sharing*) policy: only conformant packets are accepted when the queue size is greater than a given threshold. We follow this simple approach because RED techniques proposed in *DiffServ* are still subject to controversy [6] and their parameters are difficult to tune. If needed, we can easily increase the number of AF classes and replace PBS with RED.

We define three classes:

- EF (*Expedited Forwarding*). It provides flows with small delay and jitter as well as with low packet drop rate that is suitable for interactive real-time applications. To achieve such performance, EF packets have higher priority than other classes. EF flows are rate envelope multiplexed: waiting probability of EF packets is kept low by controlling the number of admitted flows based on their peak rate and by providing enough resources (link capacity).
- AF (*Assured Forwarding*). It defines a QoS class for elastic flows that do not have the strict requirements of EF flows, but need a minimum bandwidth. If the network is not congested, AF flows may obtain more bandwidth.
- BE (*Best Effort*). This class, which exists in the current Internet, does not provide any QoS guarantee.

The edge router functions are presented in figure 3. Incoming packets are classified and marked with a DSCP (*Differentiated Services Codepoint*). TCA specifies rules for classification and metering. Shaping of the EF class is done by a FIFO queue with a small size. Some bursts can be tolerated, however, packets arriving when the queue is full are dropped. Packets leave the queue according to a given peak rate. The TCA for the AF class contains a token bucket that defines the mean rate and burst tolerance. Traffic exceeding the rate is marked as out of profile and can be eliminated by core routers in case of congestion. The BE class is not controlled at all.

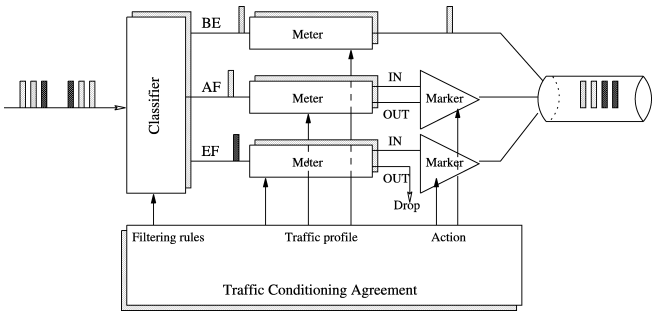


Figure 3. Edge router functions.

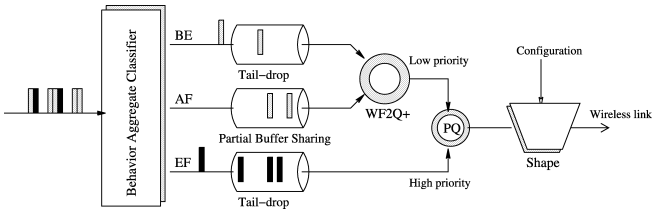


Figure 4. Core router functions.

The architecture of the core router is presented in figure 4. It is composed of three queues for each class of the traffic. The EF class has a static priority higher than AF and BE. The AF and BE classes are scheduled according to a variant of WFQ (*Weighted Fair Queueing*): WF2Q+ (*Worst-case Fair Weighted Fair Queueing*) [24]. The proportion of the bandwidth allocated to the AF and BE classes is configurable, for example, 60% and 40%. We use a tail-drop policy for the EF and BE queues: a packet is dropped when the queue is full. Conformant and non-conformant packets of the AF class are

subject to the PBS (*Partial Buffer Sharing*) policy. In this way, all AF packets may benefit from available resources, however in case of congestion only conformant packets will be allowed in the network. The output traffic is limited by a token bucket to fit the rate of the output link. Note that the EF class benefits from a fixed part of the available bandwidth and the AF and BE classes share the bandwidth not used by EF flows.

Our goal is to extend the *DiffServ* model to a wireless environment so that we can provide consistent IP level quality of service to mobile hosts (cf. figure 2). We can rely on the implementation of *DiffServ* that gives us a set of mechanisms for managing quality of service in a wireless LAN environment: classification and marking, packet scheduling and traffic shaping. Using this mechanisms we can constrain the traffic sent over the wireless LAN and process each performance class with respect to its PHB.

However, there is also a problem of resource management: how to provision sufficient resources to guarantee some QoS parameters. The *DiffServ* performance guarantees rely on sufficient provisioning of network resources with respect to accepted SLAs. In the case of wireless LANS such as 802.11, the bandwidth of the wireless link becomes a critical resource and to provide some statistical QoS guarantees, we should add some form of admission control and signaling. These functions should be specialized in order to take into account the characteristics of wireless LANs, e.g., the fact that the available bandwidth decreases with the number of active hosts. In our architecture, an Access Routers acts as a QoS manager for a cell by performing admission control and configuring the *DiffServ* mechanisms of mobile hosts.

A wireless LAN raises also the issue of fast varying conditions: they change fast when mobile hosts arrive in a cell or the users activate applications. This means that the manager of QoS has to keep track of the current load in a cell and dynamically configure scheduling mechanisms and traffic shapers in all hosts of a cell. Moreover, admission control should be done based on the current state of resources in a cell, for example, a hand-off to a given cell can be denied or granted in a degraded mode if there are no sufficient resources to satisfy the moving mobile host. Furthermore, the QoS management should be tightly coupled with mobility management so that the overall end-to-end performance perceived by mobile hosts be acceptable for each QoS class. Our hierarchical architecture addresses all these issues.

We assume that the wireless access network is composed of several wireless LAN cells. Interconnection of cells is done at the IP level so that mobility between cells is managed at the network layer. The reason for this is that the QoS management should be tightly coupled with mobility management. As the former is done at the IP level, we have to manage mobility also at the IP level. Since mobility is basically a routing problem, IP seems to be the right level to deal with local mobility.

Figure 5 shows the elements of our architecture. Each wireless cell is managed by an *Access Router* that forwards packets between mobile hosts in a cell and connects it to an *Edge Router* via a wired over-provisioned LAN. All mobile hosts and Access Routers are provided with the *DiffServ* mechanisms (the edge and core router functions) so that traffic sources are controlled in function of varying conditions of a cell: parameters of traffic shapers and scheduling mechanisms resulting in bandwidth allocations for QoS classes can be adjusted to provide requested performance behavior.

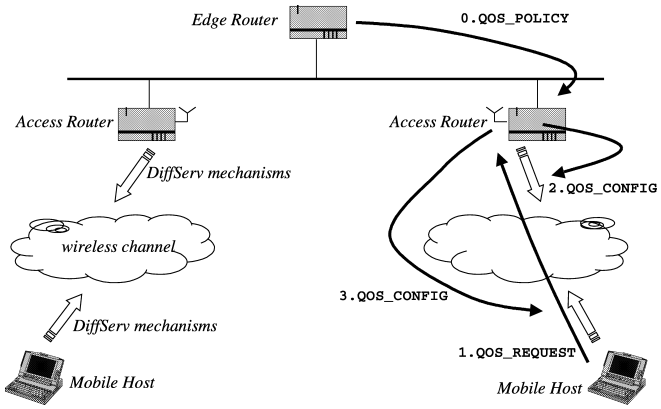


Figure 5. Hierarchical QoS architecture.

The proposed QoS architecture is hierarchical because we can identify two time scales and two levels of management: *intra-cell management* and *inter-cell management*. We can observe that the state of a wireless cell can change rapidly. For example, available resources may decrease due to a movement of a mobile host or after launching a new application. The first level of QoS management (intra-cell management) is thus local to one cell and performed by the Access Router that manages fast changing local situations. Mobile hosts inform the Access Router on the required bandwidth and the Access Router, in turn, configure their QoS mechanisms.

The second level (inter-cell management) concerns a set of wireless cells connected to an Edge Router. At this level, the conditions change slowly, for example, when some resources should be reserved on a given path over several cells or we want to change admission control rules. This global management is done by the Edge Router that fixes long term policies for Access Routers.

4.2. Local QoS management

We assume that hosts within a cell communicate using a MAC layer such as IEEE 802.11. As we have seen, to provide quality of service we should constrain the IP traffic of different classes sent over the link. However, the available bandwidth of the link depends on the number of active hosts in a cell and on the aggregated traffic of each class. So, the Access Router in charge of QoS management in a cell should be informed about the bandwidth required by each mobile host, keep track of the number of host, and configure the parameters of the *DiffServ* mechanisms to obtain desired behavior.

4.2.1. Bandwidth allocation

The QoS allocation problem can be stated as follows: given available bit rate capacity C and $x_{i,\text{class}}$, traffic rate of class EF, AF, BE requested by source i, find proportions r_{EF}, r_{AF} and r_{BE} of the bandwidth to be allocated to each respective class:

$$\begin{aligned} x_{\text{EF}} &= \sum x_{i,\text{EF}} \leqslant r_{\text{EF}} r(m) C, \\ x_{\text{AF}} &= \sum x_{i,\text{AF}} \leqslant r_{\text{AF}} r(m) C, \\ x_{\text{BE}} &= \sum x_{i,\text{BE}} \leqslant r_{\text{BE}} r(m) C, \\ r_{\text{EF}} &+ r_{\text{AF}} + r_{\text{BE}} = 1 - \delta, \end{aligned} \tag{5}$$

where δ accounts for overprovisioning of the allocation and $r(m)$ is the proportion of the effective bandwidth if m hosts are active.

To perform bandwidth allocation, we have measured the proportion of the useful bandwidth in function of the number of hosts in the 802.11b wireless LAN (see figure 1 introduced in section 2). Based on these statistics we can configure the *DiffServ* mechanisms of EF, AF and BE classes to limit their aggregated output rate to $r_{\text{EF}} r(m) C$, $r_{\text{AF}} r(m) C$, and $r_{\text{EF}} r(m) C$, respectively.

Bandwidth allocation follows the *soft-state* principle. The Access Router interprets requests for QoS allocation (`QOS_REQUEST` included in each data packet as explain later – in-band signaling) and satisfies them if possible by appropriate configuration of *DiffServ* mechanisms (`QOS_CONFIG` sent in control packets – out-of-band signaling). The QoS management module in the mobile host configures the output rate of the EF and AF/BE classes and fixes the proportion between the AF and BE classes. The configuration may concern only a given mobile host and the Access Router or even all mobile hosts in a cell, for example, if a new request requires the modification of the QoS parameters in all hosts. An allocation is given for a time interval and when a mobile host stops sending packets, its allocation is canceled after the interval.

4.3. Global QoS management

The Edge Router acts as a global QoS manager for Access Routers managing cells. It sets policies to be followed by Access Routers such as admission control and reservation of resources (`QOS_POLICY`). For example, we can imagine that a priority mobile host reserves sufficient resources in cells on a given path. In this case, we can configure the *DiffServ* mechanisms in the mobile hosts in the cell to limit the current AF/BE traffic. When the mobile handoffs to the next cell, it benefits from the part of the already allocated bandwidth. We propose the following rules for reservations:

1. Reserve a given bandwidth in all cells.
2. Reserve a given bandwidth in the cells on a given path.
3. Reserve a given bandwidth in the cells on a frequent mobility path (found from mobility observation).
4. Reserve a given bandwidth in the neighbor cells.
5. Reserve a given bandwidth in one cell.

Rule 4 is the default rule for the EF traffic. The AF class has lower performance requirements, so rule 5 will be its default rule. Based on the current state of reservations, the Edge Router may adapt policies that fix the number of admitted hosts in a cell, however, Access Routers do not reject flows of mobile hosts already accepted in a cell.

4.4. QoS signaling

Managing QoS as well as mobility (described in the next section) requires information exchange at the IPv6 level between all elements of our architecture: mobile hosts, Access Routers, and the Edge Router. Figure 6 shows the protocol structure of a mobile host. An Access Router has a similar structure, but without the Traffic Profile Configuration. The IPv6 stack includes two modules: QoS and mobility management, and *DiffServ* mechanisms. Mapping between a `FlowId` used by an application and a required traffic profile is defined in a configuration table, e.g., the administrator can specify that a real-time multimedia application that uses a given `FlowId` requires 64 Kb/s bandwidth allocated to the EF class. The table is used to configure the classification mechanism of *DiffServ*. Cooperation between the management modules is done by means of a signaling protocol that either uses data packets for communication (in-band signaling) or generates ICMP control packets (out-of-band signaling). The in-band approach allows to take into account fast changing situations in a wireless cell. The signaling information has the format presented in table 2. Bandwidth allocation units can be configured for a given cell, they can be smaller or greater depending on the kind of applications that use the cell. We experimented with the unit of 64 Kb/s.

In-band signaling consists of inserting commands into data packets transmitted between a mobile host and an Access Router. Two solutions are possible: for short commands such as `QOS_REQUEST`, we can encode this information into a part of the IPv6 Flow Label field (note that the use of the Flow Label field is restricted only to a local access network, this information is taken away from the packets sent to the core network). The rest of the field is still used for a flow identifier to distinguish between flows. Another solution uses

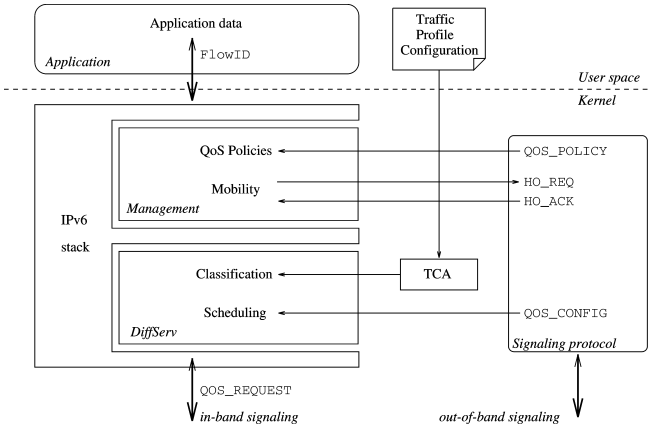


Figure 6. Protocol structure of a mobile host.

Table 2
Signaling protocol format.

Command	Parameters
QOS_REQUEST	bandwidth in allocation units
QOS_CONFIG	EF rate, AF/BE rate, AF weight
QOS_POLICY	bandwidth in allocation units, traffic class, source address, policy type
HO_REQUEST	target AR, current QoS allocations
HO_ACK	source AR, host route
HO_DENY	source AR

header extensions to hold the signaling information. Recall that this information is local to a wireless access network and it is removed by the Access Router before sending packets the Edge Router.

If there is no data traffic, we need another way of signaling. We propose to define a new type of ICMPv6 to contain signaling commands. This solution is also required if a signaling command should be sent to a remote entity, which is the case, for example, of the hand-off request.

5. Mobility management for fast hand-offs

As we have stated before, one of the design requirements for our mobility management scheme was its integration with QoS support. Fast hand-offs can only be achieved when a mobile host keeps its IP address when moving to another cell. To do this, the routes in the wired backbone should be updated to reflect the new location of the host. Careful preparation of the new route in advance makes it possible to avoid lost packets and reduces the hand-off delay.

We describe below the operation of our mobility management protocol during a hand-off (cf. figure 7 in which the topology is simple – there is no intermediate routers between access routers and the cross-over router, the Edge Router). We assume that Access Routers send periodical beacons that provide mobile hosts with the identity of possible target Access Routers for a hand-off and enables measuring of the signal to noise ratio.

- *Hand-off initiation.* At some instant the mobile host decides to move to another cell. This decision can be based on some standard parameters such as the signal to noise ratio or it can take into account QoS parameters: the load or the number of hosts in the current and in the adjacent cell. We assume that mobile hosts can set the roaming mode on 802.11 cards to receive beacons from neighbor cells which allows to measure the signal to noise ratio. When the decision to move is taken, the mobile host sends a hand-off request (HO_REQ) to the target Access Router (AR2) via its current Access Router (AR1) to setup a new route (step 1). The request contains the address of the Access Router of the target cell and the request for bandwidth allocation.
- *Hand-off request propagation.* The current Access Router (AR1) propagates the hand-off request to the target Access Router (AR2) that checks whether the request can be satisfied or not. For example, if there are not enough resources, the hand-off may be denied or granted in degraded mode, e.g., instead of an EF allocation, a host obtains an AF allocation. To avoid such a situation, which may severely affect QoS performance, Access Routers can pre-reserve resources in adjacent cells according to set up policies.

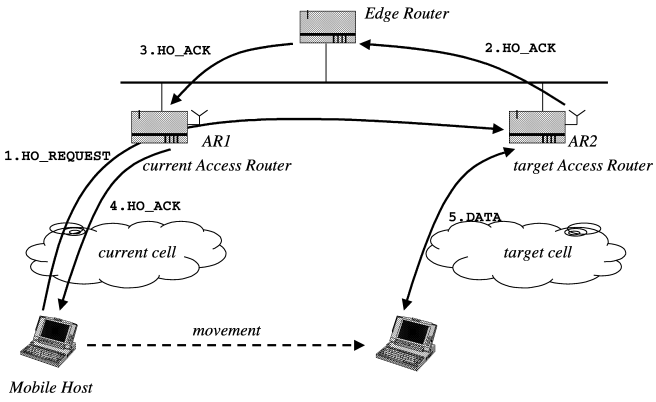


Figure 7. Hand-off protocol.

- *Hand-off granted.* If the hand-off request is accepted, the target router modifies its routing table by inserting a host route for the mobile host. The request is acknowledged to the mobile host (HO_ACK) via the cross-over router and the current Access Router (AR1) (step 2).
- *New route setup.* After accepting the mobile host, the target Access Router (AR2) relays the acknowledgement containing the new host route that should be set up in all routers in the wired backbone up to the cross-over router (step 3). All routers update their routing tables by inserting a host route that goes via the target Access Router (AR2) to reflect the new location of the mobile host. At this instant, the traffic from hosts behind the Edge Router can be forwarded to the target cell using the new route.
- *Old route deletion.* The cross-over router forwards the acknowledgement to all routers on the old route to the previous Access Router (step 4). The routers changes the old route in the routing tables. At this instant, the traffic from the previous cell can be forwarded to the target cell using the new route.
- *End of hand-off.* When receiving the acknowledgement the mobile host changes its routing table by specifying the target Access Router (AR2) as its default router and changes the channel to be used in the target cell. At this instant, the mobile host is able to communicate with mobiles in the target cell.

5.1. Discussion

Our mobility management scheme is similar to those studied in the HAWAII project [21]. At the beginning, we considered the UNF scheme, however it does not take into account the QoS management – before using a cell, the new Access Router has to check whether the QoS requirements of the mobile host can be satisfied or not. So, in our mobility scheme, we initiate a hand-off by contacting the current Access Router

before using any resource of the target cell. The mobile host changes its routes and starts using the target cell after the target Access Router has granted permission. This means that there are enough resources to satisfy the QoS requirements of the mobile host.

The order of route updates prevents transient routing loops or the creation of multiple traffic streams during hand-off similarly to the HAWAII UNF and MNF schemes [21]. As the route updates are done before the mobile host changes the transmission channel, it receives all packets along the old route.

Moreover, the scheme is optimized so that the traffic can be delivered as soon as possible to the new location: after the first route setup at the target Access Router (AR2), some part of the traffic to the mobile host can be already delivered; after step 3 and 4, the rest of the traffic is rerouted to the new location. There is, however, a caveat in this scheme: packets going along the new route can be sent by the target Access Router before the mobile host changes channels and is able to receive them. This may only happen during a short period between the instant of the route update and the beginning of the communication in the target cell. We are currently investigating how the Access Router can prevent losses.

6. Implementation and experience

We have implemented the *DiffServ* mechanisms and the mobility management scheme on FreeBSD 3.2 notebooks that use a shared 11 Mb/s 802.11b wireless link. However, in the current prototype they are not yet integrated. The *DiffServ* mechanisms are implemented in the IPv6 stack so we were able to measure performance of service differentiation presented below. We are currently working on the implementation of signaling protocols and the integration with mobility management.

6.1. Measured performance of service differentiation

We have measured performance of service differentiation in the following experiment (see figure 8). A mobile host has two traffic sources: an UDP source generating priority EF traffic of rate 300 Kb/s with short 50 Bytes packets (a simple request-response test application) and a TCP source generating elastic BE traffic (`netperf` tool for measuring useful bandwidth with 1KB packets). In the first experiment, the QoS control mechanisms are disabled. Figure 9 presents the bandwidth obtained by the BE source measured at the application layer. The BE class is in competition with the EF class and gains most of the available bandwidth – we can see that its bandwidth stays around 5 Mb/s. We also show the round trip delay (RTT) of the EF class. Until $\texttt{SeqNum} = 100$ the EF class is in competition with the greedy BE class. We can observe that the EF class is severely disturbed by the BE class, because both classes are scheduled according to the FIFO policy. At $\texttt{SeqNum} = 100$, the BE source stops sending, so that the RTT of the EF class becomes shorter, around 2.5 ms, and much more predictable.

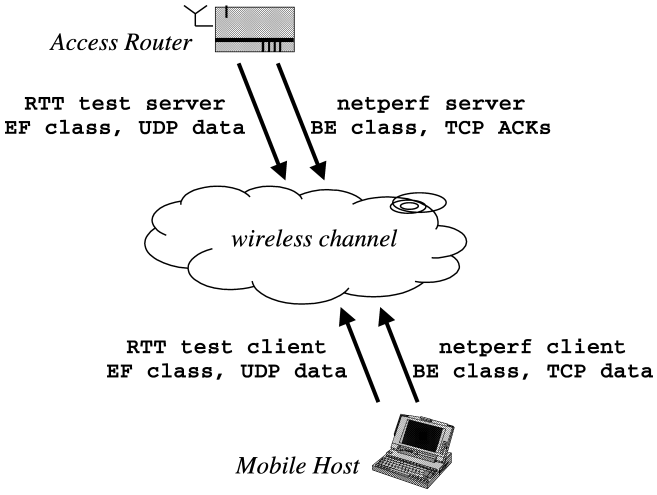


Figure 8. Experimentation set up.

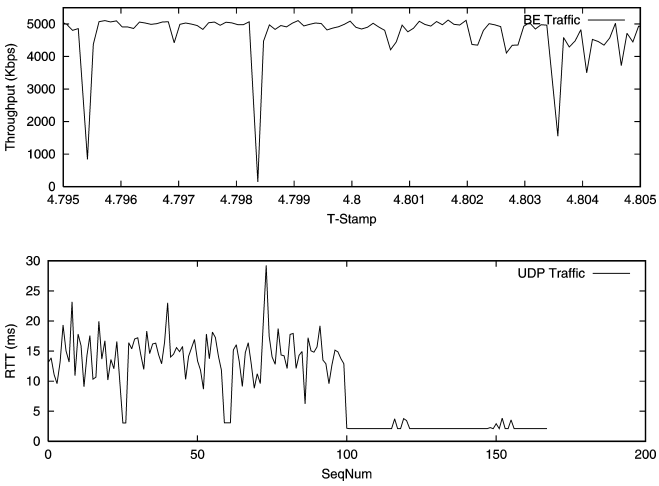


Figure 9. No QoS control; bandwidth of the BE class and RTT of the EF class.

The second experiment tests the isolation of both classes by means of the *DiffServ* control mechanisms. The output traffic shaper limits the bandwidth of the BE class to 2.4 Mb/s. Figure 10 shows the bandwidth obtained by the BE source, which is effectively maintained around 2.4 Mb/s. It can also be seen that the RTT of the EF class is much less disturbed by the BE class. It is still greater than 2.5 ms, because of the competition with the BE class (the priority policy is not preemptive and an EF packet may wait an interval corresponding to the residual waiting time). As previously, the BE source stops sending in the middle of the observation ($\texttt{SeqNum} = 130$). These measures show that it is possible to isolate different QoS classes and obtain satisfactory performance.

Figure 11 shows similar results for the BE and AF class: a UDP source sends AF traffic of rate 100 Kb/s, a TCP source generates elastic BE traffic. The output traffic shaper limits the output bandwidth to 2.4 Mb/s.

In another experiment we compared service differentiation between all three classes. An UDP source generates priority EF traffic of rate 100 Kb/s, another UDP source

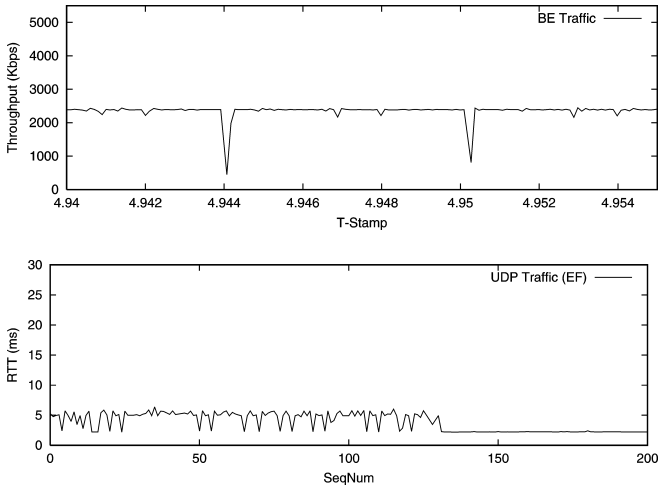


Figure 10. QoS control, bandwidth of the BE class and RTT of the EF class.

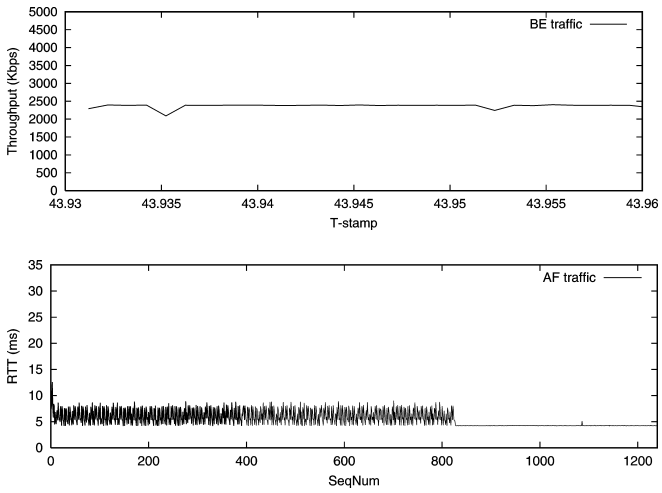


Figure 11. QoS control, bandwidth of the BE class and RTT of the AF class.

sends AF traffic of rate 100 Kb/s, and a TCP source generates elastic BE traffic. The output traffic shaper limits the output bandwidth to 2.4 Mb/s. Figure 12 shows the bandwidth obtained by the BE source, which is effectively maintained around 2.4 Mb/s. Figures 13 and 14 present the round trip delay (RTT) of the AF and EF class, respectively. Until `SeqNum` = `900` the AF class is in competition with the greedy BE class, afterwards all three classes are active, and finally at `SeqNum` = `1400` the BE source stops sending. We can observe that the EF class is only slightly disturbed by the other classes.

Finally we have measured performance of service differentiation when traffic is generated on two mobile hosts. The conditions are similar to those of figure 10: one mobile host with a UDP source generating priority EF traffic of rate 100 Kb/s and another mobile host with TCP source generating elastic BE traffic.

Figure 15 presents the bandwidth obtained by the BE source and the round trip delay (RTT) of the EF class. Until `SeqNum` = `780` the EF class is in competition with the greedy BE class and then the BE source stops sending. We

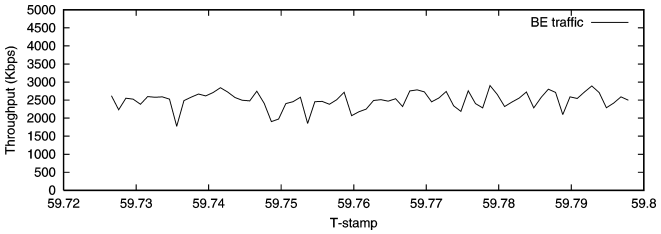


Figure 12. QoS control, bandwidth of the BE class.

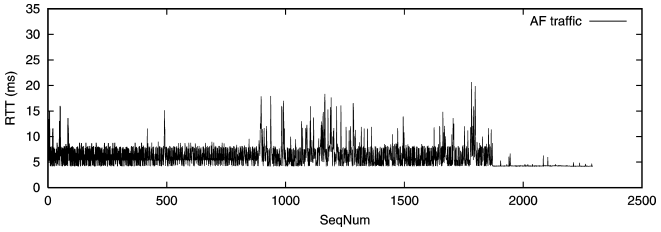


Figure 13. QoS control, RTT of the AF class.

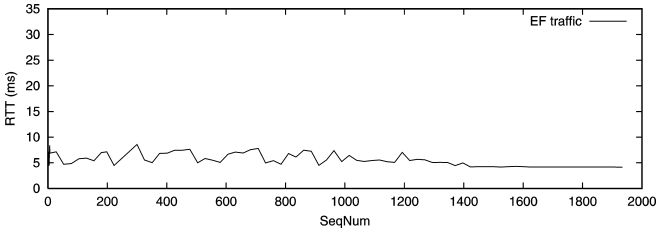


Figure 14. QoS control, RTT of the EF class.

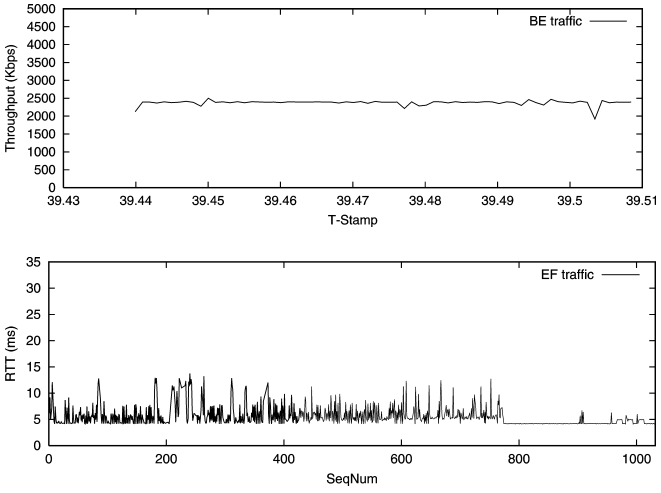


Figure 15. QoS control, distributed sources, bandwidth of the BE class and RTT of the EF class.

can seen that even in the case of distributed sources, the EF class is only slightly disturbed by BE traffic.

6.2. Hand-off protocol

The hand-off protocol has been prototyped in IPv4 over WaveLAN cards in the ad-hoc mode, which is the only mode that allowed to obtain signal to noise measurements in neighbor cells. However, this mode does not allow changing chan-

nels so that our neighbor cells have to use the same communication channel. Due to this limitation, the prototyped hand-off protocol was simpler, because the mobile host could listen to neighbor Access Routers simultaneously: when the target Access Router accepts a hand-off, it propagates the route update to the cross-over router and sends the acknowledgment to the mobile host directly. The hand-off protocol was implemented in user space using UDP. Deamons executing on routers wait for hand-off messages and perform route updates as requested. Note that contrary to the scheme described in section 5, there are no lost packets during the hand-off, since the mobile host uses the same channel in both cells. Obviously, this mode of operation is not desirable in general, because we want to provide sufficient bandwidth to QoS enabled applications.

We have measured the performance of the hand-off between two overlapping cells that use the same communication channel as described above. The measurements only include the cost of mobility management and they do not account for QoS resource allocation nor for configuration of *DiffServ* mechanisms. The mean hand-off latency is of the order of 5 ms which is fairly low compared to the performance of Mobile IP [12,18].

7. Conclusions and future work

Wireless local area networks provide many advantages compared to the proposed wide area global mobility solutions such as UMTS. We believe that their increasing deployment will create a basis for the high performance Wireless Mobile Internet. The only missing functionality is the support for quality of service and mobility.

Currently there are several different proposals for handling mobility in IP networks as well as for providing better than Best Effort QoS. However, they present separated efforts in both domains. In this paper, we have proposed a contribution towards the integrated management of QoS and mobility based on a hierarchical architecture. Our first results show that we are able to provide substantially better performance to the priority traffic, isolate different QoS classes, and manage mobility efficiently.

Acknowledgements

This work has been supported by the French Ministry of Industry, National Network of Telecommunication Research (RNRT) via the @IRS project: "*Integrated Architecture for Networks and Services*". Useful discussions with Jean-Luc Richer and constructive remarks from the referees are gratefully acknowledged.

References

[1] I. Aad and C. Castellucia, Differentiation mechanisms for IEEE 802.11, in: *INFOCOM* (2001).

[2] ANSI/IEEE, 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications (2000).

[3] M. Barry et al., Distributed control algorithms for service differentiation in wireless packet networks, in: *INFOCOM* (2001).

[4] G. Bianchi, Performance analysis of the IEEE 802.11 Distributed Coordination Function, JSAC Wireless Series 18(3) (2000).

[5] S. Blake et al., An architecture for Differentiated Services, Internet RFC 2475 (1998).

[6] T. Bonald et al., Analytic evaluation of RED performance, in: *INFOCOM'2000* (March 2000) pp. 1415–1424.

[7] R. Braden et al., Integrated Services in the Internet architecture: an overview, Internet RFC 1633 (1994).

[8] F. Cali, M. Conti and E. Gregori, IEEE 802.11 wireless LAN: Capacity analysis and Protocol enhancement, in: *INFOCOM* (1998).

[9] A.T. Campbell et al., An overview of cellular IP, in: *IEEE Wireless Communications and Networks Conference, WCNC* (1999) pp. 606–611.

[10] D. Chalmers et al., A survey of Quality of Service in mobile computing environments, IEEE Online Communication Surveys 2(2) (1999).

[11] J. Chan et al., The challenges of provisioning real-time services in wireless Internet, Telecommunications Journal of Australia 50(3) (2000).

[12] A. Helal et al., Towards integrating wireless LANs with wireless WANs using mobile IP, in: *IEEE Wireless Communications and Networks Conference, WCNC* (2000).

[13] @IRS, Integrated architecture for networks and services, RNRT Project (2001), `http://www-rp.lip6.fr/airs/`

[14] S. Lee and A.T. Campbell, INSIGNIA: in-band signaling support for QOS in mobile ad hoc networks, in: *Mobile Multimedia Communications, MoMuC* (1998).

[15] A. Lindgren et al., Evaluation of Quality of Service schemes for IEEE 802.11 wireless LANs, in: *Proceedings of the 26th Annual IEEE Conference on Local Computer Networks (LCN 2001)* (November 2001).

[16] A. Lindgren et al., Quality of Service schemes for IEEE 802.11 – a simulation study, in: *Proceedings of the Ninth International Workshop on Quality of Service (IWQoS 2001)* (June 2001).

[17] D. Mitzel, Overview of 2000 IAB Wireless Internetworking Workshop, Internet RFC 3002 (2000).

[18] S. Mukkamalla and B. Raman, Latency and scaling issues in mobile IP, ICEBERG Technical Report (2001).

[19] C.E. Perkins, Mobile IP specification, Internet RFC 2002 (1996).

[20] C.E. Perkins and D.B. Johnson, Mobility support in IPv6, in: *Mobile Computing and Networking* (1996) pp. 27–37.

[21] R. Ramjee et al., HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks, in: *IEEE Internat. Conf. Network Protocols* (1999).

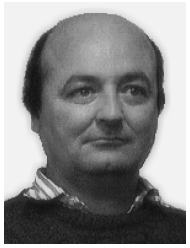
[22] A. Valko, Cellular IP – a new approach to Internet host mobility, ACM Computer Communication Review (1999).

[23] J. Weinmiller et al., Performance study of access control in wireless LANs – IEEE 802.11 DFWMAC and ETSI RES 10 HIPERLAN, Mobile Networks and Applications 2(1) (1997) 55–67.

[24] B. Zhang et al., WF2Q: Worst-case fair weighted fair queueing, in: *INFOCOM 96* (1996).

J. Antonio Garcıa-Macıas is a General Director of Fundación Teleddes, Ensenada, México. He received his Ph.D. from INPG in 2002. Previously, he was an industry consultant in telematics. His research interests include mobile and wireless networks.

Franck Rousseau is an Assistant Professor at INPG-Ensimag (Ecole Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble). He is a member of the LSR laboratory in Grenoble. He received his Ph.D. from INPG in 1998. Previously, he was a member of technical staff at the Open Group Research Institute in Grenoble. His research interests include wireless networks and communication applications.



Gilles Berger-Sabbatel is a Chargé de Recherche at CNRS and a member of the LSR laboratory in Grenoble. He received his Ph.D. in 1978 and his Habilitation diploma in 1988 both from the INPG. His research interests include parallel computing, distributed systems and networks.

Leyla Toumi is a Ph.D. candidate at INPG. She is a member of the LSR laboratory in Grenoble. Her research interests include quality of service and scheduling mechanisms in heterogeneous networks.

Andrzej Duda is a Professor at INPG-Ensimag (Ecole Nationale Supérieure d'Informatique et de Mathématiques Appliquées de Grenoble). He is a member of the LSR laboratory in Grenoble. He received his Ph.D. from the Université de Paris-Sud in 1984 and his Habilitation diploma from the Grenoble University in 1994. Previously, he was an Assistant Professor at the Université de Paris-Sud, a Chargé de Recherche at CNRS, and a Visiting Scientist at the MIT Laboratory for Computer Science. His research interests include performance evaluation, distributed systems, multimedia, and networks.