Wireless Networks 7, 425–436, 2001

Local and Global Handovers Based on In-Band Signaling in Wireless ATM Networks *

M. AJMONE MARSAN and C.F. CHIASSERINI
Dipartimento di Elettronica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

A. FUMAGALLI **
School of Engineering and Computer Science, University of Texas at Dallas, P.O. Box 830688, MS EC33, Richardson, TX 75083-0688, USA

R. LO CIGNO and M. MUNAFÒ
Dipartimento di Elettronica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Abstract. This paper presents a handover protocol for wireless ATM networks, which makes use of in-band signaling, i.e., of ATM resource management cells, to process network handovers and guarantee the in-sequence and loss-free delivery of the ATM cells containing user data. The goal of the proposed approach is to minimize the modifications of the ATM signaling standard required to overlay user mobility onto the fixed network infrastructure, and provide for a gradual upgrade of the fixed network to handle mobility. The proposed protocol handles both local handovers, in which the connection access point needs not migrate to a new ATM local exchange, and global handovers, in which the connection access point must migrate to a new local exchange. The handover scheme is devised so as to grant in-sequence delivery of cells. The performance of the network during handover is analyzed in case of connections requiring loss-free operation. The considered performance figures are the cell transmission delay introduced by the handover and the cell buffering requirements posed to the network. The behavior of the proposed protocol in presence of multiple handovers is studied via simulation, while a simple analytical method is derived for the performance evaluation of a single handover in isolation.

Keywords: wireless ATM networks, handover, in-band signaling, Quality of Service, protocol simulation

1. Introduction

The tremendous growth of cellular systems for wireless mobile telephony is a worldwide phenomenon. The number of GSM (Global System for Mobile communications) subscribers in Europe is rapidly approaching the number of fixed Public Switched Telephone Network (PSTN) terminals while in some countries the number of GSM phones is already larger than the number of fixed phones. Beside Europe, GSM networks are available in several countries worldwide, including Australia, China and most of America. Analogous wireless systems in the USA and Japan have similarly impressive growth rates. This enormous technical and commercial success of mobile telephony services is considered to be only the first wave of the personal communications era, soon leading to an equally strong demand for, and growth of, multimedia services delivered to mobile users.

Several alternative technologies are being considered for the provision of narrowband and wideband integrated multimedia services to mobile users, some being just short-term solutions. The most promising long-term solution to integrate Broadband Integrated Services Digital Network (B-ISDN) multimedia services and wireless networks is the so-termed Wireless ATM (W-ATM). Evidence of this trend can be found in the enormous work devoted by the ATM Forum to this subject. Moreover, specific projects such as R2067-MBS of the RACE program [1], focused on the design of the Mobile Broadband System (MBS), a wireless cellular network fully integrated into B-ISDN. The main target of this project is to provide an end-to-end transport service of ATM cells to and from mobile users, with the clear advantage of obtaining an overall homogeneous system, and the possibility to use standard ATM equipment within the mobile terminal. Several other research projects like those described in [2–7] are studying and prototyping large bandwidth (up to 25 Mbit/s) cellular networks, integrated within B-ISDN, shaking the traditional idea that mobile networks must be narrowband and specialized.

In this paper we take a similar viewpoint, and focus on a solution that supports the delivery of multimedia services to mobile ATM user terminals, a goal that is difficult to achieve due to a number of factors. First of all, integrated multimedia services require quite high user data rates (from 2 Mbit/s up), that are common in high-speed wired networks, but are themselves a problem over a radio interface. In addition, the integration of mobility within B-ISDN implies dynamic reestablishment of the ATM Virtual Circuit

* This work was supported in part by a research contract between CSELT and Politecnico di Torino and in part by the Italian Government through the Ministry for University and Research (MURST) and through the National Research Council (CNR). A preliminary version of this paper has been presented at the *Workshop on Nomadic Computing Satellite of IPPS'97* in Geneva, Switzerland, April 5, 1997.

** This work was carried out when A. Fumagalli was at the Electronics Department of Politecnico di Torino.

(VC) during the time span of the mobile terminal handover. Notice that this VC reestablishment procedure must ensure in-sequence and loss-free delivery of the ATM cells containing user data.

From the architectural point of view, one of the most demanding challenges of W-ATM with respect to existing wireless networks is the need for providing effective handover procedures within the ATM network (in this paper termed *network handovers*), besides the traditional radio handovers. A promising solution to this problem is based on the enhancement of the ATM switches belonging to the fixed network to handle wireless communications [8–11]. In particular, some (or all) switches in the ATM network must be *mobility-aware*, by virtue of implementing services and protocols that allow mobility. This approach implies the extension of the “standard” ATM services all the way to the mobile terminal with a full integration of wireless and mobility functions into the user, control and management planes of the protocol stack. This seamless integration of both fixed and wireless (mobile) connections requires some enhancement in the current ATM signaling protocols, by either upgrading the ATM Forum UNI 3.1/4.0 and ITU Q.2931, or finding some valid alternative to standard signaling.

Some initial design work on mobility management in ATM networks is reported in [10,12,13], and recently the ATM Forum released a Draft Standard [14]. For instance, in [8] mobility is achieved with an enhanced ATM signaling protocol termed “Q.2931+”, which incorporates some advanced functions that are necessary to carry out handover control and quality of service renegotiation.

This paper focuses on a more recent handover protocol, originally proposed by the authors in [15], whose objective is twofold: minimize the required changes in the standard ATM switches and allow incremental reestablishment procedures [16] to more efficiently serve the mobile user terminal that undergoes a sequence of network handovers. With the considered protocol, the wireless ATM segment becomes an integral part of the wired ATM network while, at the same time, required enhancements in the signaling protocol standards and recommendations are kept minimal.

The above objective is achieved through the use of in-band signaling [17]. In [17] in-band signaling is used to synchronize the transmission of data cells during the network handover in order to guarantee in-sequence cell delivery over the connection, whereas the information exchange among the nodes along the connection path is realized through both control signaling and in-band signaling. In our proposal, the handover protocol is entirely based on dedicated cells that are transmitted with the data flow of the connection. The dedicated cells, termed Mobility Enhancement Signaling (MES) cells, are Resource Management (RM) cells similar to those used in the Available Bit Rate (ABR) and ATM Block Transfer (ABT) ATM transfer capabilities.

The handover protocol, outlined in [15] and analyzed in more detail in this paper, is designed to perform both *local* and *global* handovers. In the former case, the mobile terminal is moving to a destination Base Station (BS) that is connected to the same ATM Local Exchange (LE) to which the current BS is connected. The complexity of the local handover procedure is limited as only one LE is involved and must switch the mobile connection from the port attached to the source BS to the port attached to the destination BS. In the latter case, the mobile terminal moves to a destination BS that is connected to a LE different from the original LE. Consequently, the connection path over the wired network must be modified to reach the destination LE. In either case the connection interruption during the network handover may be significant and may produce a considerable traffic disruption.

While [15] was mainly concerned with the description of the in-band signaling solution to support mobility within ATM networks, this paper carries out a detailed evaluation of the impact of the handover duration on traffic disruption and on the buffering requirements posed to the network. We follow two distinct approaches complementing each other. In section 4 we present a simple analytical model, that enables the derivation of the maximum handover duration, as well as the maximum buffering requirements posed to the network by a single cell loss-free handover. In section 5 the detailed analysis of the global buffer requirements posed to the network by concurrent handovers is carried out through the simulation of two different scenarios: one modeling multiple local handovers and the other modeling multiple global handovers. The handovers are triggered by a number of mobile terminals that are free to move from one BS to another. Concurrent handover of both terminals exchanging cells over the same connection is possible in the simulated environment.

As we shall demonstrate, the proposed in-band signaling technique can be used in conjunction with reasonably short buffers at the destination BS to guarantee in-sequence and virtually loss-free cell delivery during both local and global handovers, even when the user’s offered load approaches the radio channel capacity. The proposed handover protocol does not require handover buffers at the ATM LEs.

2. Architecture overview

As the focus of the paper is on the connection rerouting at the ATM layer, the details of the radio handover, e.g., the criteria for initiating the handover, are not considered. This choice was determined by the explicit objective of designing a handover procedure that is independent of the underlying radio technology.

We consider an air interface underneath ATM that allows mobile terminals to be part of a B-ISDN system by transmitting information in the form of ATM cells. Examples of air interfaces satisfying these requirements are being developed as part of the RACE R2067-MBS project [1] or the ACTS AC085 Magic WAND project [18]. This approach offers the significant advantage of leading to a simple and homogeneous network architecture with end-to-end ATM delivery between nodes making use of standard ATM service access points. This means that at the mobile terminal each cell is

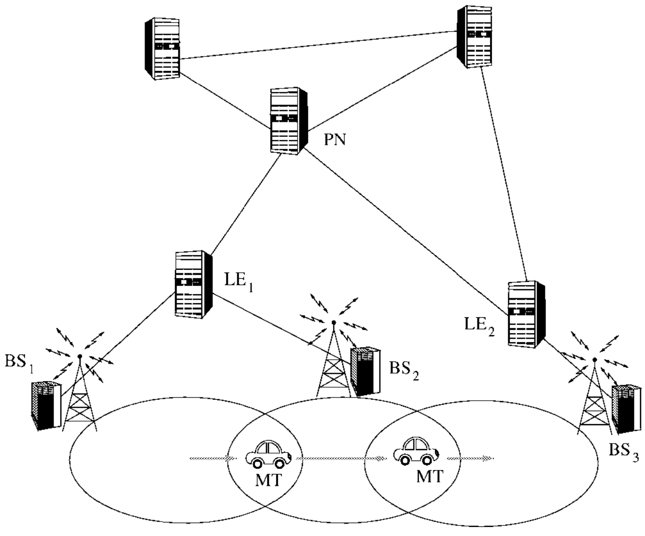

Figure 1. W-ATM scenario.

transmitted over the radio channel (*wireless segment*) connecting the mobile terminal to the wired portion of the network, and the same is subsequently routed through the wired ATM network (*fixed network segment*).

As shown in figure 1, a W-ATM network consists of Mobile Terminals (MTs), Base Stations (BSs), that provide the interface between the wired and the wireless portions of the network, and ATM switches. ATM switches and BSs are wired together and form the fixed network segment, whereas MTs can directly communicate only with BSs. Each BS communicates with all the MTs in its corresponding macro-cell (possibly comprising a number of micro-cells and/or pico-cells). A number of BSs can be connected to the same ATM switch – in this paper referred to as the Local Exchange (LE) with respect to the considered MT.

In the protocol stack shown in figure 2, the radio interface between the BS and the MT comprises the Radio Physical Layer (RPHY), the Radio Multiple Access Control Layer (RMAC) and the Radio Logical Link Control Layer (RLLC), globally referred to as the Radio Access Layer (RAL). The end-to-end virtual channel connection between mobile/fixed terminals requires AAL Layer functions only at the end terminals. BSs and ATM switches (including LEs) support only the ATM layer, with the additional capability of extracting from and inserting into the connection data flow the MES cells of the handover protocol. The protocol architecture shown in figure 2 is very general and reflects the common features of the protocol stacks proposed by different W-ATM projects and studies carried out in Europe [1,9,18,19] and in the USA [8,20].

3. Handover protocol based on in-band signaling

We distinguish between two types of handover depending on the user location. If the MT moves to a neighboring micro- or pico-cell belonging to the same BS area (i.e., macro-cell), the handover is managed by the radio interface without any involvement of the ATM layer. In this case, the procedure is called *radio handover* as only a change of radio channel within the coverage of the same BS takes place. Conversely, if the MT enters a border micro- or pico-cell, which belongs to two or more overlapping BS areas, the so-called *network handover* must be initiated. In the latter case we further consider two possible scenarios. If both the source and destination BSs are under the same LE jurisdiction, the network handover is defined as *local*, otherwise it is defined as *global*. From the viewpoint of the MT there is no difference in handling either network handover.

We recall that the proposed network handover relies upon incremental reestablishment of the connection path. This technique is appealing because it requires only the establishment of a new partial path (without involvement of the remote host) which connects the MT to a portion of the original connection path, thus allowing VCs to be partly reused. Note that, because of spatial locality in movement, it is very likely that the reestablished path to the new location of the mobile user shares most of the VPs in the old path. As a consequence, this technique is expected to be fast and efficient.

The next two subsections describe the proposed handover protocol based on in-band signaling, that guarantees in-sequence and loss-free delivery of data cells at both end terminals. As already mentioned, the in-band signaling relies upon special Resource Management cells, named MES

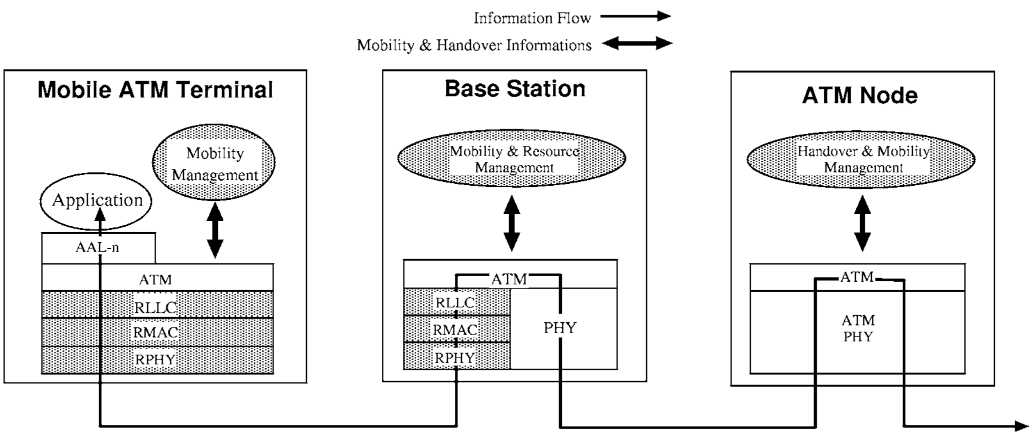

Figure 2. Protocol stack of the fixed and wireless network segments.

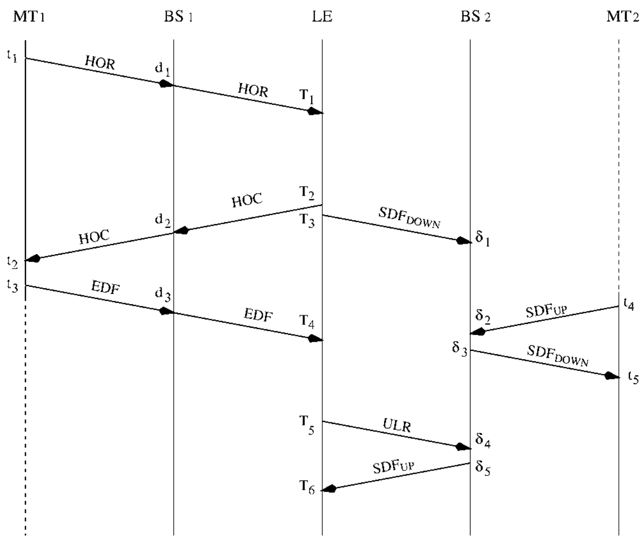

Figure 3. Sequence of messages for the LNH protocol.

cells, that are exchanged by the parties involved in the network handover. The messages exchanged during the handover procedure are listed below. A detailed description of the proposed format of the MES cells is given in appendix.

Handover Request (HOR): sent by the MT to initiate the handover procedure. This message includes both the identifier of the source BS, to which the mobile is connected, and the identifier of the destination BS, towards which the mobile is moving.

Handover Confirm (HOC): sent by the network to inform the MT that the handover request has been accepted and the handover can be executed. After this message transmission, the downstream connection, i.e., the data flow towards the MT, is switched over the new connection path towards the destination BS.

End Data Flow (EDF): sent by the MT to terminate the upstream transmission (i.e., the data flow from the MT) over the old connection path.

Start Data Flow Uplink (SDF_{UP}): sent by the MT to the destination BS indicating the beginning of the upstream transmission over the new radio link. This message serves also as a trigger to begin the radio downstream transmission at the destination BS.

Start Data Flow Downlink (SDF_{DOWN}): sent by the network to the MT to initiate the downstream transmission over the new wired connection path. The message is buffered by the destination BS while the radio link is not available.

UpLink Ready (ULR): sent by the network to the MT when the incremental (wired) connection path for the upstream transmission is ready.

3.1. Local Network Handover (LNH) protocol

Local Network Handover (LNH) takes place when the MT moves from the macro-cell controlled by the source BS, denoted by BS_1, to the macro-cell controlled by the destination BS, denoted by BS_2, and both BSs are connected to the same LE. MES cells are exchanged between the MT, BS_1, BS_2 and LE, and their circulation is restricted to the radio segment and the portion of the fixed network comprising the two BSs and the LE. The MT user–network interface consists of two transmission queues, one reserved for data cells and the other reserved for MES cells. Data cells are transmitted only when the queue reserved for MES cells is empty. At the BSs and LE, data and MES cells are buffered in the same queue in order to maintain cell flow order. The sequence of messages exchanged in the LNH protocol is shown in figure 3. We assume that radio channel c1 is used by the terminal to communicate with BS_1 and radio channel c2 is used to communicate with BS_2. To specify on which radio channel the various messages are transmitted, in figure 3 the MT is denoted as MT_1 and MT_2, respectively.

The handover protocol starts with a HOR message sent by the MT over the connection. While waiting for the HOC message that confirms the network acceptance to process the handover, the MT continues to send (receive) data cells over the original upstream (downstream) connection. Upon reception of the HOR message, the LE attempts to establish the incremental connection through BS_2 (time lapse $[T_1, T_2]$). In this phase network resources are reserved at BS_2 to meet the QoS requirements of the ongoing connection, e.g., the radio bandwidth necessary to continue the wireless connection through BS_2. If sufficient network resources are not available at BS_2, the handover request is refused with a negative HOC. This case is not considered in the paper, as it requires proper counter-actions that are beyond the scope of our study. Conversely, in case of successful establishment of the incremental connection, the LE sends a (positive) HOC to the MT through BS_1, and immediately updates the downstream routing table. From this moment, downstream cells arriving at the LE are directly forwarded to BS_2. Upon reception of the HOC message, the MT terminates the upstream transmission through BS_1 by sending the EDF message. The EDF message informs the LE that no more upstream data cells will arrive through BS_1. At this point, the LE upstream routing table is updated (time lapse $[T_4, T_5]$). Soon afterwards, the ULR message is sent to BS_2 to indicate that the upstream transmission on the wired segment has been reestablished.

Between instants t_3 and t_4 the radio connection is not available: the upstream transmission is suspended at the MT which stores the generated data cells in the transmission buffer. Similarly, the downstream radio transmission from BS_2 towards the MT is not yet operational, thus, cells arriving from the network must be stored in the *downstream handover buffer* at BS_2, until the new radio link is established (time lapse $[\delta_1, \delta_3]$). Once the radio link is established, the MT sends the SDF_{UP} message and immediately resumes the upstream transmission. At BS_2 the upstream cells sent by the MT are temporarily (time lapse $[\delta_2, \delta_5]$) buffered in the *upstream handover buffer* until the incremental wired connection path through the LE is established – this condition is notified by the reception of the ULR message. This mechanism

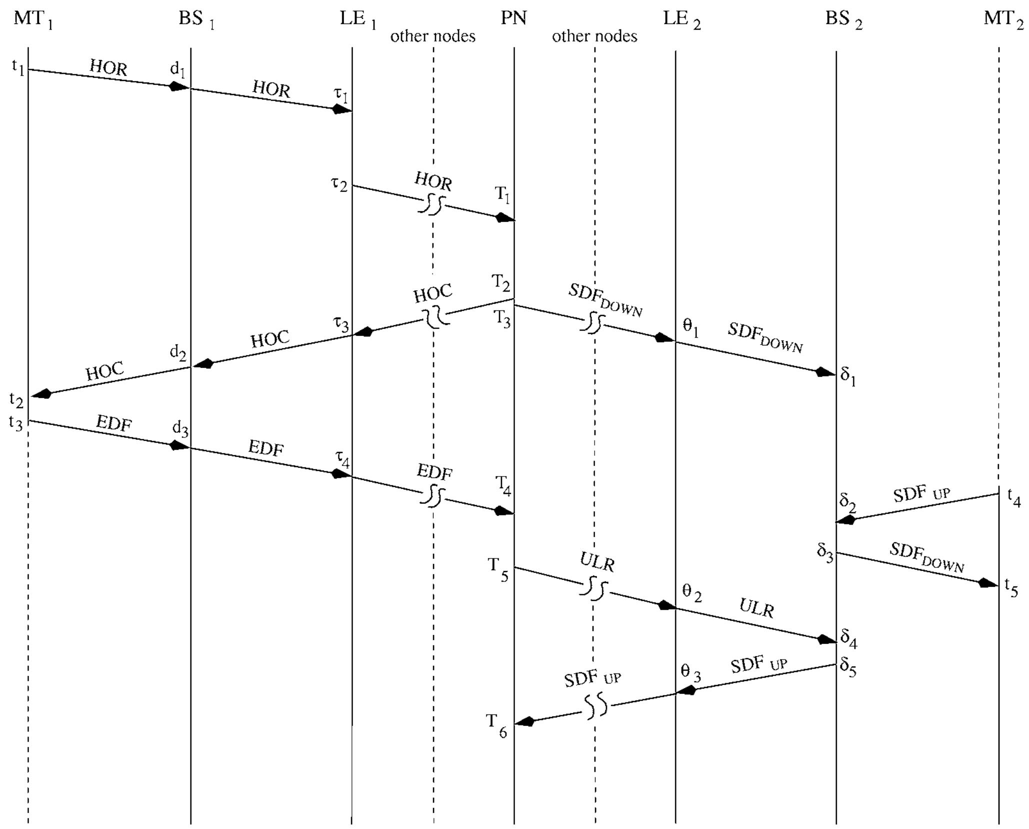

Figure 4. Sequence of messages for the GNH protocol.

and the upstream handover buffer are necessary to guarantee that upstream cells sent over the incremental path are received by the LE and transmitted to the remote terminal only after all the upstream cells transmitted over the original path have been forwarded by the LE to the remote terminal, i.e., to guarantee in-sequence cell delivery. The SDF_{UP} message is the first upstream cell transmitted by the BS_2, indicating that the handover protocol has been successfully terminated.

Notice that in the proposed protocol temporary handover buffers are required only at the destination BS.

3.2. Global Network Handover (GNH) protocol

In Global Network Handover (GNH) the MT is moving to a destination BS_2 that is connected to a LE, denoted by LE_2, that is different from the original LE, denoted by LE_1.

The LNH protocol presented above can be adapted to handle the GNH with minor changes as shown in figure 4. Upon reception of a handover request (HOR message), LE_1 looks at its table containing all the BSs under its jurisdiction. Since the destination BS does not belong to this set, the search is further extended to the BSs controlled by the neighboring LEs, i.e., the set of LEs which are connected to the BSs adjacent to the BSs under LE_1 control. Once the destination LE_2 is identified, a *Pivot Node* (PN) must be found along the original connection path, which becomes the junction switch between the existing path and the incremental path established between the PN and LE_2. We assume that the wired network runs the appropriate algorithm which, given the old path and the destination LE_2, finds the optimal PN and the incremental path connecting the PN to the destination LE_2. The details of this step in the handover procedure are not considered here, as the PN search algorithm is out of the scope of this paper. The algorithms described in [21] or any other suitable algorithm can be adopted to serve this purpose. If necessary the algorithm can be run "off-line" storing the results in a dedicated lookup table so that the time needed to find the PN is kept minimal. In GNH, the PN plays the same role as the LE in the LNH protocol, while the MT and the two BSs behave exactly as in the LNH. In figure 4 the PN is identified during time lapse $[\tau_1, \tau_2]$. Once the PN has been identified, upon reception of the HOR message the PN establishes the incremental connection toward BS_2 (time lapse $[T_1, T_2]$), sends the HOC message to MT, updates its downstream routing table and sends the SDF_{DOWN} message to BS_2.

Similarly to what happens in the LNH protocol, upon reception of the SDF_{DOWN} message, BS_2 begins buffering the downstream data cells arriving from the PN. These cells remain stored in the downstream handover buffer until the new radio link is established between BS_2 and MT. On the upstream transmission path, the MT stores data cells in the transmission buffer during the radio handover (time lapse $[t_3, t_4]$). As soon as the new radio link is available between BS_2 and the MT, upstream cells transmitted by the MT are stored in the upstream handover buffer of BS_2 until the ULR message sent by the PN is received.

It must be pointed out that MES cells exchanged during the GNH are restricted to the portion of the path connecting the nodes participating in the handover procedure, i.e., the MT, the PN, BS_1, BS_2, LE_1 and LE_2. ATM switches possibly located along the connection path between the PN and the two LEs may ignore the existence of the in-band signaling cells, and handle them as normal RM cells. This unique feature of in-band signaling allows us to gradually upgrade the ATM network to handle terminals' mobility because only a subset of ATM switches may be upgraded to operate as PN.

4. Analysis of buffer requirements at the destination BS

In this section we evaluate the maximum occupancy of the upstream and downstream handover buffers as a function of relevant system parameters. We recall that the upstream handover buffer is the memory space that must be reserved at the destination BS during the handover in order to store the cells that are transmitted by the MT before the ATM upstream path is reestablished. Conversely, the downstream handover buffer is the memory space needed at the destination BS to store the cells sent towards the MT in the downstream direction before the radio link is re-established.

The analysis is based on the assumption that the radio link makes use of a TDMA protocol, but it can be extended to other types of access, such as FDMA, CDMA and channels with contention access, provided that the contention resolution protocol ensures a maximum medium access delay. Moreover, it is assumed that both handover buffers are emptied by a shaper device implemented as a Continuous Time Leaky Bucket (CTLB) or equivalently as a Generic Cell Rate Algorithm (GCRA) [22] device, whose parameters are defined according to the connection traffic contract.

The parameters used in our analysis are defined in table 1. Two additional variables are defined, $\psi_U = \max(\delta_5 - \delta_2, 0)$ and $\psi_D = \max(\delta_3 - \delta_1, 0)$. Variable ψ_U (ψ_D) represents the time lapse during which upstream (downstream) data cells may arrive to the destination BS from the MT (PN), but cannot be forwarded to the PN (MT).

Since MES cells are generated asynchronously, their transmission over the TDMA radio link is subjected to a waiting time represented by a random variables ξ_{MES} (the subscript indicates the actual type of MES cell), whose maximum value is one radio time slot, a quantity that is less than or equal to τ_{SCR}. If we consider a stand alone handover occurring in a network that is not overloaded, i.e., where queueing delays are negligible, for both local and global handovers we obtain:

$$\psi_U = \max\bigl(T_p + D_1 + \gamma_{PN} + T_{UPD} + D_2 + \gamma_{BS} - (T_{LAT} + T_p), 0\bigr), \quad (1)$$

$$\psi_D = \max\bigl(T_p + D_1 + \xi_{HOC} + \gamma_{MT} + \xi_{EDF} + T_{LAT} + T_p + \xi_{SDF_{DOWN}} - D_2, 0\bigr). \quad (2)$$

Table 1
Parameters used in the delay analysis.

γ_{PN}	MES cell processing delay at the PN
γ_{BS}	MES cell processing delay at the BS
γ_{MT}	MES cell processing delay at the MT
T_{LAT}	radio link disruption time ($t_4 - t_3$ in figures 3 and 4)
ξ_{MES}	time between a MES cell generation (arrival) at the MT (BS) and its transmission over the radio link (see section 3 for a list of the types of MES cells)
T_{UPD}	time required to update the routing tables of the ATM switch to build the incremental path of the ongoing connection
PCR	Peak Cell Rate of the shaper devices, set equal to the radio link capacity
SCR	Sustainable Cell Rate of the shaper devices
τ_{SCR}	time interval between two consecutive cell transmissions at the Sustainable Cell Rate ($\tau_{SCR} = 424/\mathrm{SCR}$)
L_o	offered load per connection
T_p	cell propagation delay over the radio link
D_i	cell propagation delay between BS_i ($i = 1, 2$) and PN (or LE in local handovers)

If we reasonably assume that $\gamma_{PN} = \gamma_{BS} = \gamma_{MT} \simeq 0$ and we calculate the upper value of (1) and (2) we obtain, respectively:

$$\psi_{U\max} = D_1 + D_2 + T_{UPD}^{\max} - T_{LAT}^{\min}, \quad (3)$$

$$\psi_{D\max} = D_1 - D_2 + T_{LAT}^{\max} + 2T_p + 3\tau_{SCR}. \quad (4)$$

Assuming that during the entire handover transient the MT and the remote terminal transmit bursts of cells at their maximum sustainable speed, i.e., at the SCR, the maximum number of cells that may accumulate in the handover buffers can be derived from (3) and (4) as described next.

Let us first consider the transient behavior of the upstream handover buffer. Figure 5 plots the number of cells in the upstream handover buffer against time for three values of SCR, namely, 1.8 (solid line), 1.9 (dashed line), and 2.0 (dotted line) Mbit/s, in the case of local network handover with PCR = 2.0 Mbit/s and offered load $L_o = 1.6$ Mbit/s.

The three curves are obtained through simulation. On the shown time scale, time = 0 corresponds to the instant when the SDF_{UP} message reaches the destination BS (δ_2). The vertical dashed line marks the time instant when the ULR message reaches the destination BS (δ_4), and triggers the upstream cell transmission. Before this event occurs, the number of cells in the buffer grows linearly in all cases while the effect of the SCR value becomes relevant only after the reception of ULR – when SCR = 2.0, i.e., equal to PCR, the buffer occupancy remains constant until MT ends the transmission of the cells belonging to the burst (time $\simeq$ 140 ms); when SCR = 1.9 or 1.8, the buffer occupancy decreases steadily, thus ensuring that the handover buffer empties in a finite time under all conditions.

The transient behavior of the downstream buffer occupancy in the case of local handover, as well as the behavior of both the upstream and the downstream buffers in the case of global handover, are similar to the one reported above: the buffer occupancy grows linearly from the SDF_{DOWN} arrival instant until the SDF_{UP} message reaches the BS.

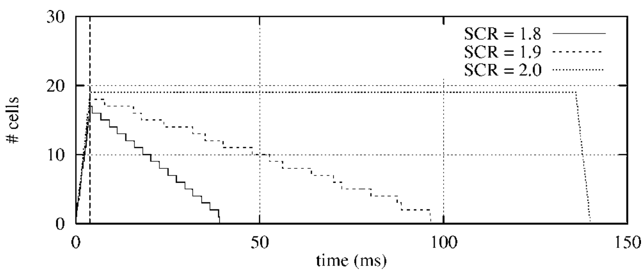

Figure 5. Upstream handover buffer occupancy versus time for three values of SCR and average offered load (L_o) 1.6 Mbit/s and PCR = 2.0 Mbit/s.

Table 2
Maximum occupancy of the upstream and downstream handover buffers for three values of SCR with PCR = 2.0 Mbit/s and L_o = 1.6 Mbit/s.

SCR	1.8	1.9	2.0
$\eta_{\mathrm{U}\max}$	20	21	22
$\eta_{\mathrm{D}\max}$	6	6	6

Therefore,

(i) dividing (3) and (4) by τ_{SCR}, and

(ii) increasing the result by 1 to take into account the $\mathrm{SDF_{DOWN}}$ cell buffered at δ_1 in the downstream buffer and the $\mathrm{SDF_{UP}}$ cell buffered at δ_3 in the upstream buffer,

the maximum occupancy of the two handover buffers is, respectively:

$$\eta_{\mathrm{U}\max} = \frac{D_1 + D_2 + T_{\mathrm{UPD}}^{\max} - T_{\mathrm{LAT}}^{\min}}{\tau_{\mathrm{SCR}}} + 1, \qquad (5)$$

$$\eta_{\mathrm{D}\max} = \frac{D_1 - D_2 + T_{\mathrm{LAT}}^{\max} + 2T_{\mathrm{p}}}{\tau_{\mathrm{SCR}}} + 4. \qquad (6)$$

The above equations are used to validate the simulations results of a stand alone local network handover. Table 2 shows the maximum occupancies derived through simulation by assuming $D_1 = D_2 = 0.25$ ms, the offered load $L_o = 1.6$ Mbit/s, and PCR = 2.0 Mbit/s, $T_{\mathrm{UPD}}^{\max} = 4$ ms, $T_{\mathrm{p}} = 8.5$ µs, $T_{\mathrm{LAT}}^{\min} = 17$ µs, and $T_{\mathrm{LAT}}^{\max} = 425$ µs. The simulation results perfectly match with the numerical results obtained by using (5) and (6).

5. Buffer requirements for simultaneous and repetitive network handovers

The analysis carried out in the previous section provides some insight into buffer requirements at the BS during a stand alone network handover. A detailed performance study that takes into account the occurrence of multiple handovers towards the same BS and the concatenation of frequent handovers originated by the same terminal is however mandatory and carried out via simulation in this section. Simulation results are obtained by emulating the proposed handover procedure in CLASS[1] [23,24] and observing several thousand handover procedures.

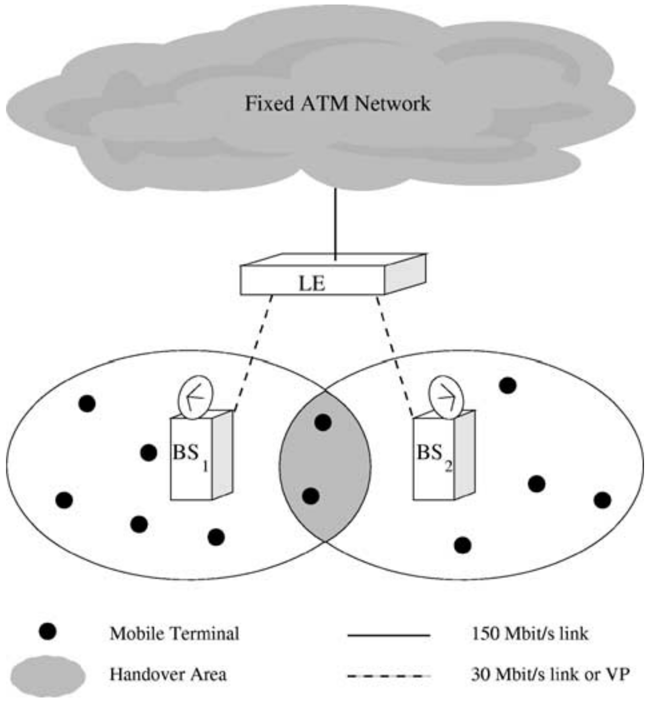

Figure 6. Simulation analysis setup for local network handover.

The setup used for the local handover analysis is shown in figure 6. The analysis of the system performance is carried out under the following general assumptions. Links connecting BSs to the LE have capacity equal to 30 Mbit/s, and links connecting neighboring ATM switches have transmission rate equal to 150 Mbit/s. The propagation delay between BSs and the LE is 0.25 ms, accounting for a distance of about 50 km. In each simulation run, 10 MTs per BS are modeled, and they are free to move at random from one BS to another. ATM switches are non-blocking with fixed crossover delay, therefore, such delay does not influence the handover performance and will not be considered in the model. The time needed by the PN (or the LE in case of LNH) to update the routing tables, for both downstream and upstream connections, is a truncated exponential random variable with minimum value equal to 1 ms, mean value equal to 4 ms, and maximum value equal to 10 ms. Values measured on small size switches [25] indicate that the routing table update requires a time that varies slightly around 4 ms. However, we speculate that in real networks this value will depend on the switch type and on its workload, thus a random variable is a better representative than a constant value. The radio bandwidth allocated to each MT

[1] CLASS is a cell level simulator of ATM networks developed at the Electronics Department of Politecnico di Torino in cooperation with CSELT (the research center of Telecom Italia) and the Technical University of Budapest; CLASS stands for Cell Level ATM network Services Simulator; it is entirely written in C language and portable on most computing platforms. CLASS is freely available for research and education: information and instructions on how to obtain it are available at the URL `http://www1.tlc.polito.it/class/`.

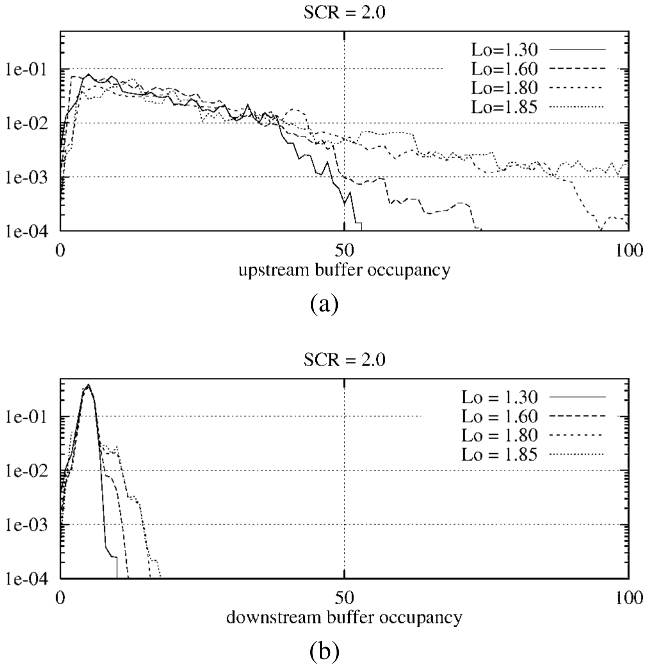

Figure 7. Histogram of the upstream and downstream handover buffer occupancy for varying values of the average offered load (L_o) and SCR = 2.0 Mbit/s during local handovers.

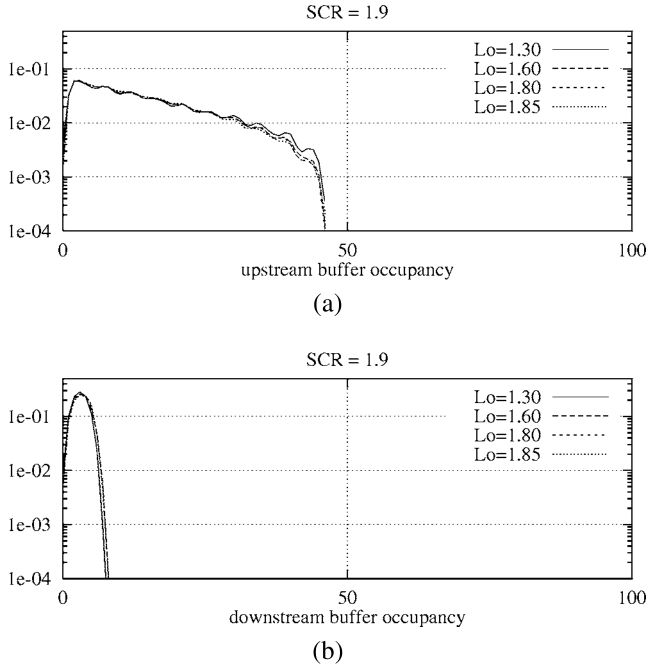

Figure 8. Histogram of the upstream and downstream handover buffer occupancy for varying values of the average offered load (L_o) and SCR = 1.9 Mbit/s during local handovers.

is 2 Mbit/s for either upstream and downstream transmissions. The propagation delay on the radio channel is roughly 8.5 μs, which corresponds to a distance between the MT and the BS of about 2.5 km. The radio handover latency (i.e., absence of transmission and reception at the MT) is a random variable uniformly distributed between twice the radio propagation delay (17 μs) and twice the cell transmission time at the radio interface (425 μs). Every MT generates *messages* (i.e., PDUs of a higher layer) which are segmented at the AAL level – AAL5 is considered for this purpose. Messages are generated according to a Poisson point process. The message length is modeled by a truncated geometric random variable whose mean is 20 ATM cells and whose truncation value is 100 cells. The average offered load L_o of each MT is assumed variable in order to study the influence of this parameter on system performance. The PCR of the VC between the BS and the LE during the network handover is equal to the capacity of the radio link, while the SCR is a varying parameter in the model.

It is assumed that a MT which just finished a handover might begin a new handover procedure even if the handover buffers at the current BS have not been emptied yet. The time interval between two consecutive handovers of the same terminal is assumed to be uniformly distributed between 0.3 and 3 s. Although nowadays unrealistic, these small values were chosen to explore the behavior of the system under the occurrence of highly frequent handovers, a scenario that may become a reality in next generation B-ISDN networks.

Figures 7 and 8 show the estimated probability density function or histogram of the upstream and downstream handover buffer occupancy for two values of SCR (1.9 and 2.0 Mbit/s) and varying values of the average load L_o offered to the connection. Figure 7 shows that for SCR = 2.0 Mbit/s the buffer occupancy distribution has heavy tails, especially when the offered load approaches the sustainable cell rate, SCR. Indeed, in this case the MES cells exchanged during the handover protocol must queue in the handover buffers at BS_1, so that the time intervals $[\delta_1, \delta_3]$ and $[\delta_2, \delta_5]$ may arbitrarily increase (see figure 3). The resulting concatenation of handovers induces heavy tails of the probability distribution. The tail weight grows with the offered load; however, as shown in figure 8, a negotiated SCR 5% smaller than the radio link capacity is sufficient to reduce the probability of two overlapping handover procedures for the same terminal, thus ensuring that the handover buffers are quickly emptied even when the value of L_o approaches SCR, and greatly reducing the weight of the distribution tails. The analysis of the system behavior in the conditions depicted in figures 7 and 8 suggests that the SCR parameter negotiated for ATM connections in mobile networks must always be smaller than the radio link capacity, even when this may seem wasteful.

Cells queueing in the network does not necessarily imply that the channel disruption due to the handover procedure is long: it just indicates that the transfer delay of cells along the connection is temporarily increased. Cells buffering may in any case be an annoying side effect for the network, since it implies that some network entity (base stations in our proposal) need additional buffering capabilities to manage handovers.

The occupancy of the upstream buffer is sensibly higher than that of the downstream buffer. This behavior, as indicated in (1) and (2), can be explained by observing that,

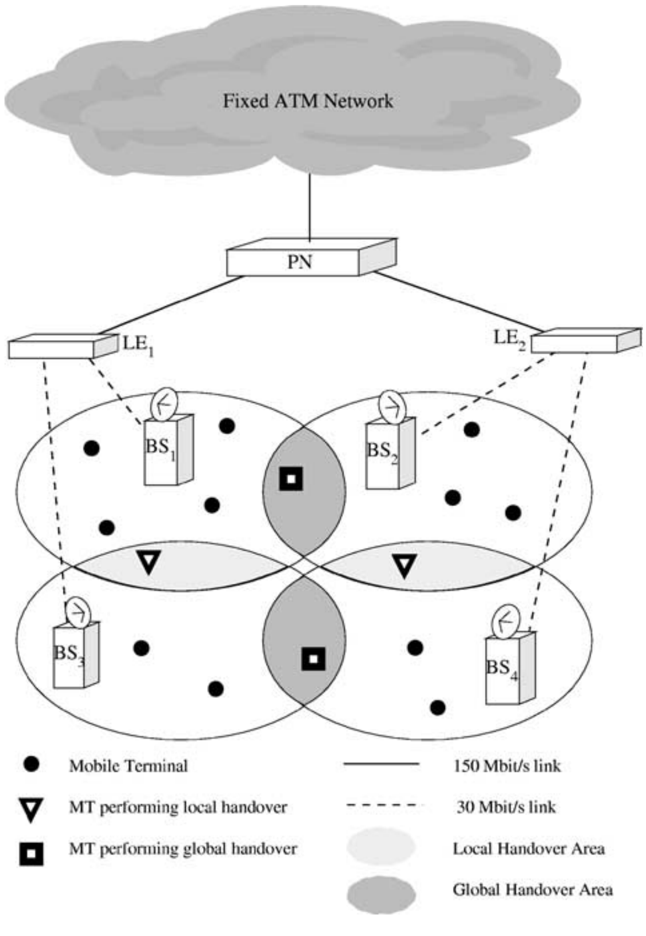

Figure 9. Simulation analysis setup for global network handover.

in the case of local handover, time ψ_D depends mainly on the radio handover latency, while time ψ_U is also a function of the time necessary for updating the routing tables and the distance between the BSs and the LE. From this fact stems the consideration that, in order to ameliorate the QoS provided during network handovers, it is not necessary to provide ultra fast radio handover procedures, but it is of the paramount importance to provide fast procedures for the update of the routing tables within ATM switches and to place the BSs not too far from the LEs.

We now focus on global handovers, showing how buffer requirements depend on the distance between LEs and the PN. Figure 9 shows the system setup used to simulate global handovers. The portion of the wired network affected by the handover is now wider and comprises the MT, two BSs, two LEs, and the PN. Since upstream and downstream buffer occupancies are functions of $D_1 + D_2$ and $D_1 - D_2$, respectively, we report only the results related to the upstream buffer, i.e., the most critical. In fact, the downstream buffer behavior for the adopted global handover scenario is very similar to the one shown for local handovers, and generally speaking, it never represents a critical issue in connections with balanced traffic.

Figure 10 shows the histogram of the upstream buffer occupancy when SCR = 1.9 Mbit/s and PCR = 2.0 Mbit/s for values of propagation delay ($D_1 = D_2$) ranging from 0.5 to 50 ms (corresponding to a link length ranging between 100 and 10,000 km). This last value is surely very large if LEs and the PN are directly connected, but it may become a reasonable value if we consider that there may be several intermediate switches between the two. The delay caused by the pivot search does not influence the presented results since during this phase the MT is still connected to the source BS.

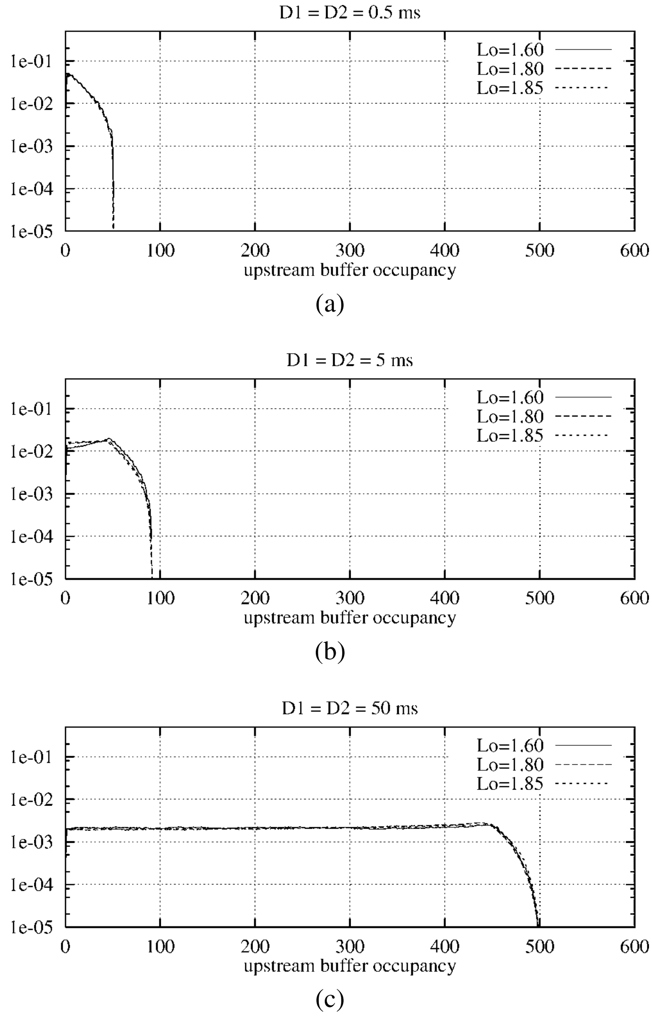

Figure 10. Histogram of the upstream handover buffer occupancy in global handovers for SCR = 1.9 Mbit/s, PCR = 2.0 Mbit/s and varying values of the average offered load (L_o), D_1 and D_2.

Table 3
Average handover procedure duration $\overline{D}_{ho}$ versus the propagation delay between LEs and the PN in case of $D_1 = D_2$.

D_i [ms]	0.5	1	2	3	4	5
$\overline{D}_{ho}$ [ms]	12.2	14.7	19.7	24.7	29.7	34.7

In general, longer distances between ATM switches would result in additional buffer occupancy, without substantial modifications of its qualitative distribution.

It is interesting to note that even in global handovers the buffer occupancy does not depend significantly on the average connection load.

Table 3 reports the average handover duration $\overline{D}_{ho}$ as a function of the average distance between LEs and PN. We only report the case when $D_1 = D_2$, but other simulated cases yielded similar results. The handover duration increases linearly with D_i and is dominated by the term $5D_i$. $\overline{D}_{\rm ho}$ is the *total* handover duration, from the emission of the first signaling message (t_1 in figure 4) to the reception of the last one (T_6 in figure 4). Channel disruptions are significantly shorter, in particular the uplink channel disruption is proportional to $2D_i$ as can be evinced from figure 4. Taking into account that the channel disruption time in GSM networks can be up to 100 ms, we can conclude that the proposed handover procedure can be safely used for real time services even when the PN is several hundred kilometers from the mobile terminal.

Though in this paper we focus on loss-free handovers, that we deem to be the target for W-ATM networks, it may be interesting to evaluate the number of packets that ought to be dropped if the handover duration is larger than a given target value $\Delta_{\rm t}$. $\Delta_{\rm t}$ represents a possible delay jitter experienced by the cells during the handover. The cell delay jitter disappears once the handover buffer is emptied. The delay jitter allowed on a real time connection is related to the size of the service playout buffer; hence, a scenario where the playout buffer is very small poses tight limits on the maximum handover duration. The number of dropped cells in this case is directly related to the handover duration. If $D_{\rm ho}$ is the effective handover duration and $\Delta_{\rm t}$ is the maximum allowed delay jitter, then the maximum number of dropped cells is $\mathrm{PCR} \times (D_{\rm ho} - \Delta_{\rm t})$. For instance, if we take $\mathrm{PCR} = 2$ Mbit/s, $\Delta_{\rm t} = 50$ ms (a value acceptable for conversational services) and let $D_{\rm ho}$ increase from 50 to 100 ms, the number of dropped cells grows from 0 to 95.

One last observation on the radically different buffering requirements posed on the downstream and upstream buffers is in order. The rationale of the different requirements is evident if one considers (3) and (4). The upstream buffer occupancy is proportional to the *sum* of the propagation delays of the links connecting the BSs to the PN. The downstream buffer occupancy is proportional to the *difference* between the propagation delay of the link connecting the source BS to the PN, and the propagation delay of the link connecting the destination BS to the PN. Although in the presence of balanced bidirectional traffic the resulting average buffer occupancy of the upstream buffer is higher than that of the downstream buffer and, consequently, the cell delay jitter observed during handover may vary considerably between the two flows, real time services and data applications are not severely affected as long as the maximum jitter is maintained below predetermined acceptable values.

6. Conclusions

This paper presented and analyzed a protocol for handling network handover procedures in a mobile wireless ATM environment. Two different types of network handover were considered, namely *local* and *global*. The former handover type refers to a user movement that does not entail migration to a new local exchange, but only to a new base station. The latter handover type refers to a user movement that entails migration to a new base station and a new local exchange; in this event, the path followed by ATM cells within the fixed network must be modified to reach the new local exchange. The protocol makes use of in-band signaling which imposes minimal modifications on the ATM signaling standards and recommendations.

Results obtained through simulation show that multiple terminals' handovers can be simultaneously performed while guaranteeing in-sequence and loss-free cell delivery. This objective is accomplished already with small handover buffers at the base station. In conclusion, on the one hand the proposed solution requires minimal enhancement of the standard ATM signaling, on the other it offers a scalable approach to handle multiple handovers simultaneously occurring within the same base station and local exchange.

The protocol analyzed in this paper is conceived for handling unicast ATM connections. The extension to multicast connections or multimedia connection supported by several ATM VCs is straightforward, though it may pose additional burdens to the network. In presence of native ATM multicast support, cells duplication is performed directly by ATM nodes, hence, we can assume that there is only a single VC on the radio link that is handled as if it was a unicast connection. When several VCs are transmitted over the radio link, two cases are possible. In the first case all the VCs are trunked together: the handover procedure is basically identical to the one described in the paper, apart from the necessity to keep track of the VC correlation. In the second case, the network handover is performed on a subset of VCs, while the others remain active through the source base station waiting for a postponed handover. This latter case is surely more complex, as it requires the network to keep track of the correlation between VCs that may enter the network at different access points. This is a matter for future studies.

Acknowledgements

The authors would like to thank J. Kurose and R. Ramjee for providing an early copy of [25].

Appendix. MES cells: a detailed description

In figure 11 the proposed format of the Mobile Enhancement Signaling (MES) cells is described [24].

The ID field of MES cells must be set to a value consistent with the ITU-T and ATM Forum specifications [22,26] and that identifies the use of these ATM cells for Mobility Enhanced Services. Field names and possible values are as follows.

Header: ATM cell header.

ID: Standard Resource Management cells Protocol Identifier.

Octet	Field		Description
	7 4	3 0	
1–5	Header		ATM header
6	ID		Protocol Identifier
7	MES_TYPE		Message type
8	FLOW_ID		Data flow identifier
9	PRI	QoS_MOD	Priority | Quality of Service modifier
10	PFL	N.D.	Pivot Flag | Not defined
11–18	S_BSA		Source Base Station address
19–26	T1_BSA		First Destination Base Station address
27–34	T2_BSA		Second Destination Base Station address
35–42	T3_BSA		Third Destination Base Station address
44–51	N.D.		Not defined
52	N.D.	CRC_1	Not defined | First part of the CRC code
53	CRC_2		Second part of the CRC code

Figure 11. Mobility Enhanced Service (MES) cells format.

MES_TYPE: This field identifies the handover protocol message carried by the MES cell. Allowed values identify all the different messages required for the handover protocol:

- HOR: HandOver Request;
- HOC: HandOver Confirm;
- EDF: End of the Data Flow on the upstream connection through the source base station;
- SDF_{DOWN}: Start Data Flow on the new DOWNlink connection;
- SDF_{UP}: Start Data Flow on the new UPlink connection;
- ULR: New UpLink connection Ready.

FLOW_ID: This field can be used to manage the coordinated handover of several connections. If the value is 0 (zero) the handover of the connection to which MES cells belong is managed independently from all others.

PRI: 4-bit field indicating how badly the handover is needed by the mobile terminal; it is used to manage handover priorities in case of network congestion.

QOS_MOD: 4-bit field indicating the possibility of a renegotiation of the Quality of Service parameters for the connection; 0 indicates that no renegotiation is needed or permitted, other values are still to be defined.

PFL: 4-bit field used to manage the MES cell propagation depth: depending on its value the MES cell will be removed from the connection by a different entity (e.g., the Pivot Node or the Destination Base Station).

S_BSA: Source Base Station Address (8 bytes).

T1_BSA: First Destination Base Station Address (8 bytes).

T2_BSA: Second Destination Base Station Address (8 bytes).

T3_BSA: Third Destination Base Station Address (8 bytes).

CRC: Standard Resource Management cells Cyclic Redundancy Code.

The presence of multiple destination BS fields is optional (unused fields may be set to zero) and reflects the possibility for a mobile terminal to indicate more than one possible target to complete the handover[2]. The use of multiple destination BSs increases the probability to successfully complete the handover in the presence of network congestion.

References

[1] M. Progler, MBS air interface principles, in: *RACE Mobile Telecom Summit*, Cascais, Portugal (November 1995).

[2] D. Raychaudhuri and N.D. Wilson, ATM-based transport architecture for multiservices wireless personal communication networks, IEEE Journal on Selected Areas in Communications 12(8) (October 1994) 1401–1414.

[3] D. Raychaudhuri, Wireless ATM networks: Architecture, system design and prototyping, IEEE Personal Communications 3(4), Special Issue on Wireless ATM (August 1996) 42–49.

[4] D.D. Falconer, A system architecture for broadband millimeter-wave access to an ATM LAN, IEEE Personal Communications 3(4), Special Issue on Wireless ATM (August 1996) 36–41.

[5] M. Umehira, A. Hashimoto, H. Matsue and M. Nakura, A proposal of an ATM wireless access system for tetherless multimedia services, Mobile Networks and Applications 1(3), Special Issue on Wireless ATM (December 1996) 335–347.

[6] J. Naylon, D. Gilmurray, J. Porter and A. Hopper, Low-latency handover in a wireless ATM LAN, IEEE Journal on Selected Areas in Communications 16(6) (August 1998) 909–921.

[7] J. Mikkonen, J. Aldis, G. Awater, A. Lunn and D. Hutchinson, The magic WAND – funcional overview, IEEE Journal on Selected Areas in Communications 16(6) (August 1998) 953–972.

[8] R. Yuan, S.K. Biswas, L.J. French, J. Li and D. Raychaudhuri, A signaling and control architecture for mobility support in wireless ATM networks, Mobile Networks and Applications 1(3), Special Issue on Wireless ATM (December 1996) 287–298.

[9] H. Mitts, H. Hansen, J. Immonen and S. Veikkolainen, Lossless handover for wireless ATM, Mobile Networks and Applications 1(3), Special Issue on Wireless ATM (December 1996) 299–312.

[10] A.S. Acampora and M. Naghshineh, An architecture and methodology for mobile-executed handoff in cellular ATM networks, IEEE Journal on Selected Areas in Communications 12(8) (October 1994) 1365–1375.

[11] E. Buitenwerf, G. Colombo, H. Mitts and P. Wright, UMTS: fixed network issues and design options, IEEE Personal Communications 2(1) (February 1995) 30–37.

[2] This is very similar to what happens in GSM networks, in which radio frequency measures are not limited to one possible destination.

[12] L.V. Hauwemeiren, L. Vercauteren, A. Saidi and T.V. Landegem, Requirements for mobility support in ATM, in: *Proc. GLOBECOM'94*, San Francisco, CA (November 1994).

[13] B.A. Akyol and D.C. Cox, Signaling alternatives in a wireless ATM network, IEEE Journal on Selected Areas in Communications 15(1) (January 1997) 35–49.

[14] B. Rajagopalan (ed.), Draft Wireless ATM Capability Set 1 Specification, ATM Forum BTD-WATM-01.13 (January 13, 2000).

[15] M. Ajmone Marsan, C.F. Chiasserini, A. Fumagalli, R. Lo Cigno and M. Munafò, Local and global handovers for mobile management in wireless ATM networks, IEEE Personal Communications 4(5), Special Issue on Mobility and Location: An European Perspective (October 1997) 16–24.

[16] C.-K. Toh, A hybrid handover protocol for local area wireless ATM networks, Mobile Networks and Applications 1(3), Special Issue on Wireless ATM (December 1996) 313–334.

[17] O.T.W. Yu and V.C.M. Leung, Connection architecture and protocols to support efficient handoffs over an ATM/B-ISDN personal communications network, Mobile Networks and Applications 1(2) (October 1996) 123–139.

[18] Detailed information about the Magic WAND project is available at `http://www.tik.ee.ethz.ch/~wand`

[19] B. Walke, D. Petras and D. Plassmann, Wireless ATM: Air interface and network protocol of the mobile broadband system, IEEE Personal Communications 3(4), Special Issue on Wireless ATM (August 1996) 36–41.

[20] D. Raychaudhuri, Wireless ATM networks: Architecture, system design and prototyping, IEEE Personal Communications 3(4), Special Issue on Wireless ATM (August 1996) 42–49.

[21] C.-K. Toh, Crossover switch discovery for wireless ATM LANs, Mobile Networks and Applications 1 (1996) 141–165.

[22] ITU-TSS Study Group 13, Recommendation I.371, Traffic control and congestion control in B-ISDN, Geneve, Switzerland (July 1995).

[23] M. Ajmone Marsan, A. Bianco, T.V. Do, L. Jereb, R. Lo Cigno and M. Munafò, ATM simulation with CLASS, Performance Modeling Tools, Performance Evaluation 24 (1995) 137–159.

[24] C.F. Chiasserini, R. Lo Cigno and M. Munafò, On the implementation of mobility services in CLASS: Architecture, validation and results, in: *Fifth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, UK (21–23 July 1997).

[25] R. Ramjee, T. La Porta, J. Kurose and D. Towsley, Performance evaluation of connection rerouting schemes for ATM-based wireless networks, IEEE/ACM Transactions on Networking 6(3) (June 1998) 249–261.

[26] ATM Forum/af-tm-0056.000, Traffic management specification, Version 4.0 (April 1996).

Marco Ajmone Marsan is a Full Professor at the Electronics Department of Politecnico di Torino, Italy. He has coauthored over 250 journal and conference papers in the areas of communications and computer science, as well as the two books *Performance Models of Multiprocessor Systems* published by the MIT Press, and *Modelling with Generalized Stochastic Petri Nets* published by Wiley. He received the best paper award at the Third International Conference on Distributed Computing Systems in Miami, FL, in 1982. His current interests are in the fields of performance evaluation of communication networks and their protocols. M. Ajmone Marsan is a fellow of IEEE.
E-mail: ajmone@polito.it

Carla-Fabiana Chiasserini did her graduate work at the Politecnico di Torino, Italy, receiving the Ph.D. degree in 1999. Since then she has been with the Department of Electronics Engineering at Politecnico di Torino. She was at the Center for Wireless Communications, University of California, San Diego, as a visiting researcher in 1999 and 2000. Her research interests include architectures, protocols and performance analysis of wireless networks for integrated multi-media services.
E-mail: chiasserini@polito.it

Andrea Fumagalli received his Ph.D. from the Politecnico di Torino, Italy, in 1992. From 1987 to 1989, he worked as a Consultant for CSELT. Between 1990 and 1992, he was a Visiting Researcher at the University of Massachusetts. From 1992 to 1998, he was on the Faculty of the Electronics Engineering Department at the Politecnico di Torino. Since 1997, he has been an Associate Professor of Electrical Engineering at the University of Texas at Dallas. Primarily dealing with network architectures and protocols, Dr. Fumagalli's research has been supported by national and international funding agencies, including ESPRIT, EIT, NSF, DARPA, DoD and the Texas Higher Education Coordinating Board. Dr. Fumagalli is an IEEE Distinguished Lecturer.
E-mail: andreaf@utdallas.edu

Renato Lo Cigno is an Assistant Professor at the Electronics Department of Politecnico di Torino. He received a Dr.Ing. degree in Electronic Engineering from Politecnico di Torino in 1988. Since then he has been with the telecommunication research group of the Electronics Department of Politecnico di Torino, first under various research grants and contracts, then as a Research Engineer and finally as an Assistant Professor. From June 1998 to February 1999, he was at the CS Department at UCLA as Visiting Scholar under grant CNR 203.15.8. He is a coauthor of about 50 journal and conference papers in the area of communication networks and systems. His current research interests are in performance evaluation of wired and wireless networks, modeling and simulation techniques and flow and congestion control. R. Lo Cigno is a member of IEEE.
E-mail: locigno@polito.it

Maurizio Munafò is a Research Engineer in the Electronics Department of Politecnico di Torino. He holds a Dr.Ing. degree in electronic engineering since 1991 and a Ph.D. in telecommunications engineering since 1994, both from Politecnico di Torino. Since November 1991 he has been with the Electronics Department of Politecnico di Torino, where he has been involved in the development of an ATM networks simulator. He is a coauthor of about 30 journal and conference papers in the area of communication networks and systems. His current research interests are in simulation and performance analysis of communication systems and QoS routing algorithms. M. Munafò is a member of IEEE.
E-mail: munafo@polito.it