

Performance Models of Handover Protocols and Buffering Policies in Mobile Wireless ATM Networks

Marco Ajmone Marsan, *Fellow, IEEE*, Carla-Fabiana Chiasserini, *Member, IEEE*, and Andrea Fumagalli, *Member, IEEE*

Abstract—Due to the connection-oriented nature of the asynchronous transfer mode (ATM), one of the challenges in mobile wireless ATM (WATM) systems is the management of terminal handovers. When ATM connections are reestablished to follow terminals moving between areas covered by distinct base stations, seamless handover protocols are necessary to guarantee that ATM cells are delivered to terminals in the correct order, with cell loss rate and delay that satisfy the contracted quality of service (QoS). A promising approach to meet QoS requirements is based on the use of handover buffers at the (destination) base station, where transmitted cells are stored while the connection is being reestablished. Up to date, only simulation and experimental results are available to determine the performance of such protocols and buffering schemes. This paper presents the first attempt to develop an analytical modeling approach to estimate the performance of handover protocols making use of handover buffers at the base station. By incorporating several approximations, the proposed models allow designers to simultaneously take into account numerous system parameters, including handover buffer size, sustainable and peak cell rates of the ATM connection, terminal offered load, and time needed to reestablish the ATM connection. Analytical performance predictions are shown to closely match results of detailed simulation experiments, thus demonstrating the suitability of the proposed modeling framework for the selection of the most adequate solution to handle handover and provide the QoS required by end users.

Index Terms—Handover procedures, Petri nets, wireless asynchronous transfer mode (WATM) networks.

I. INTRODUCTION

OVER the last few years, one of the major commercial successes in the telecommunications world has been the widespread diffusion of cellular mobile telephone services, whose provision relies on sophisticated algorithms implemented by state-of-the-art dedicated computer equipment.

The cellular nature of mobile telephony stems from the subdivision of the serviced area into *cells* that are covered by the electromagnetic signal emitted by the antennas of fixed *base stations* (BSs). The mobility of users implies that it is possible for a mobile terminal (MT) to *roam* from one cell to another, while information transfer over the network is in progress. In order for the communication to continue, it is necessary that the network

be capable of transferring the connection from the old cell (the source BS) to the new one (the destination BS). This operation is normally referred to as call *handover* or *handoff*.

Upgrading the service offer to mobile users to include high-speed data communication services poses several technical challenges. A natural approach for the introduction of high-speed data communication services for mobile users is to adopt the asynchronous transfer mode (ATM) in the wireless environment, resulting in the so-called wireless ATM (WATM) networks [1]–[3].

One of the critical design issues that arise in WATM networks is mobility management [4]–[6]. In particular, when a user's MT moves from one cell to another, all ATM connections originating or terminating at the MT must be rerouted from the source BS to the destination BS. During this transition the end user may experience: 1) cell losses and delay variations due to the temporary interruption of the wireless link and 2) out-of-order cell delivery due to the rerouting of the connection through the network. The former problem arises only in hard handovers, whereas the latter arises in both hard and soft handovers¹ [7]. Handover protocols must be designed so as to guarantee virtually loss-free and in-sequence delivery of ATM cells to end users with minimal delay increases, while the ATM connections are being rerouted.

A straightforward solution to this problem consists of a handshake protocol between the connection end users, that suspends the traffic flow while the connection is being rerouted. However, if the roundtrip time of the connection is relatively long, the traffic interruption due to the handshake may not be acceptable for applications such as voice and video that have stringent quality of service (QoS) requirements. Moreover, the user at the other end of the connection must be notified of the occurring handover, thus requiring a global upgrade of the protocols at all nodes of the wired network.

A solution that circumvents these drawbacks is represented by the so-called *seamless* handover protocols [8]–[12]. A seamless handover protocol is loss free, guarantees minimal impact on cell delay, and grants in-sequence cells delivery to the destination by performing cell buffering at some network node while the connection is being rerouted, thus avoiding the need for the handshake between terminals. Some of the proposed seamless handover protocols require cell buffering at both the destination BS and the neighboring ATM switch [9], [11]. Others require

Manuscript received September 28, 1999; revised September 20, 2000. This work was supported by a contract between CSELT and Politecnico di Torino.

M. Ajmone Marsan and C. F. Chiasserini are with the Dipartimento di Elettrotecnica, Politecnico di Torino, Torino 10129, Italy.

A. Fumagalli is with the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA.

Publisher Item Identifier S 0018-9545(01)04894-0.

¹In hard handover, the wireless link is interrupted from the instant the MT disconnects from the source BS to the instant the MT connects to the destination BS; in soft handover, the MT connection over the wireless link simultaneously reaches both the source and destination BSs so that absence of wireless connection is avoided.

cell buffering only at the destination BS with the aim to minimize the necessary upgrades of the ATM switches in existing networks [8], [10], [12].

Performance evaluation of seamless handover protocols is an essential step to compare multiple approaches to designing and controlling the WATM system. In particular, the minimum buffer size required by the protocol to meet the expected QoS is a key factor in determining the solution with optimal cost. This evaluation is not trivial, as the metrics of interest, namely the cell-loss probability due to overflow of the handover buffer depends on rare events. In spite of this situation, performance results reported in the literature are mostly obtained via simulation [6], [12] or experimental prototyping [2], [3], [11], because analytical models that estimate the performance of handover protocols and buffering schemes in WATM networks are not available or they are based on quite simplistic approximations that lead to partial results [8], [9]. For example, a deterministic approach is used in [9] to estimate the performance of a number of handover protocols, assuming that the temporal characteristics of the system dynamics can be captured by constant values, whereas in real systems many of the system timings are random in nature (for example the reestablishment delay of the ATM connection). A system design based on estimates obtained under deterministic worst-case assumptions often leads to unnecessary additional system costs.

The lack of more comprehensive analytical models can be explained by considering the inherent complexity of the system under consideration, in which

- 1) traffic over the connection may be bursty;
- 2) during the handover, the wireless link and the path of the ATM connection through the fixed network may be reestablished at different time instants;
- 3) MT may have to store its generated data while moving from one wireless access point to another (if hard handover is used);
- 4) transient originated in the system by the handover may last for a relatively long time past the reestablishment of the wireless link and the ATM connection.

This paper presents the first comprehensive analytical model to obtain approximate but accurate estimates of the performance of hard handover protocols and buffering policies at the destination BS. The novelty of the proposed model relies on both the representation of the distinct phases that characterize the call handover and the consideration of a large number of system parameters. The latter includes the bursty nature of data and the traffic load generated by the application running at the terminal, the average random time necessary to reestablish the ATM connection path through the network, the average random time necessary to establish the new wireless link, the sustainable cell rate (SCR) and the peak cell rate (PCR) of the ATM connection. The formalism used to develop the models is based on two classes of timed Petri nets [13]: generalized stochastic Petri nets (GSPNs) [14] and colored GSPNs (CGSPNs) [15]. These formalisms provide a high-level graphical formalism for the development of complex Markovian models, that would be otherwise difficult to generate directly. GSPNs and CGSPNs have established their effectiveness as a modeling paradigm for the investigation of complex communication systems in a number of contexts, from

ATM switches to GSM networks (see for example [16]–[32]). They thus seem to be suitable candidates for the development of the stochastic models required by the complex scenarios arising in WATM.

With the proposed modeling approach, it is possible to estimate the minimum handover buffer size required by the handover protocol to satisfy the cell loss probability and delay bounds required by the application. In addition, the modeling framework *quantitatively* determines the best system configuration for guaranteeing the minimum delay variance on the ATM connection performing handover. Intuitively, the minimum delay variance is achieved under two conditions: 1) the handover buffer is emptied within the shortest possible time and 2) a new handover occurs only when the handover buffer of the previous handover has been emptied. The latter condition is necessary to prevent a cumulative effect of consecutive handovers on the ATM cell delay. These two conditions are met only if the maximum burst size (MBS) of the connection is sufficiently large to allow the handover buffer to be rapidly emptied and the handover rate of the mobile terminal is below the inverse of the time required to empty the handover buffer. Numerical values for these parameters can be derived with the proposed model and may help the designer choose the handover protocol and the buffering policy that minimize the cost of the base station, as well as the worst-case MBS and maximum speed of the mobile terminal to be negotiated with the user.

For demonstration purposes, the proposed modeling framework is used to quantitatively compare the performance of two well-known policies to control the handover buffer. According to the first policy—dedicated buffer (DB)—a connection requiring handover is assigned a dedicated handover buffer at the destination BS for the entire duration of the handover. According to the second policy—buffer sharing at the base station (B^2S^2)—connections concurrently requesting handover toward the same destination BS share a common handover buffer always available at the BS. As known from a number of different application contexts, solutions based on buffer sharing lead to cost-effective and efficient implementations thanks to the multiplexing of different requests upon the same pool of resources [33], [34]. In this way, the total buffer capacity necessary to satisfy the required QoS is minimized.

It must be noted, however, that the main contribution of this paper is in the model development and validation. The presented performance results are obtained only for some values of the system parameters, and should not be viewed as an attempt to completely characterize the effectiveness of the handover protocols and buffering schemes. The presented models can quite easily accommodate the parameter values used in any design project and provide an accurate quantitative characterization of the system performance, thus helping the designer choose the most performing and cost-effective solutions under a large variety of cases.

II. SYSTEM DESCRIPTION

The WATM architecture under consideration is described in Fig. 1. Three types of nodes characterize the network as follows:

- BSs that represent the access points to the fixed network;

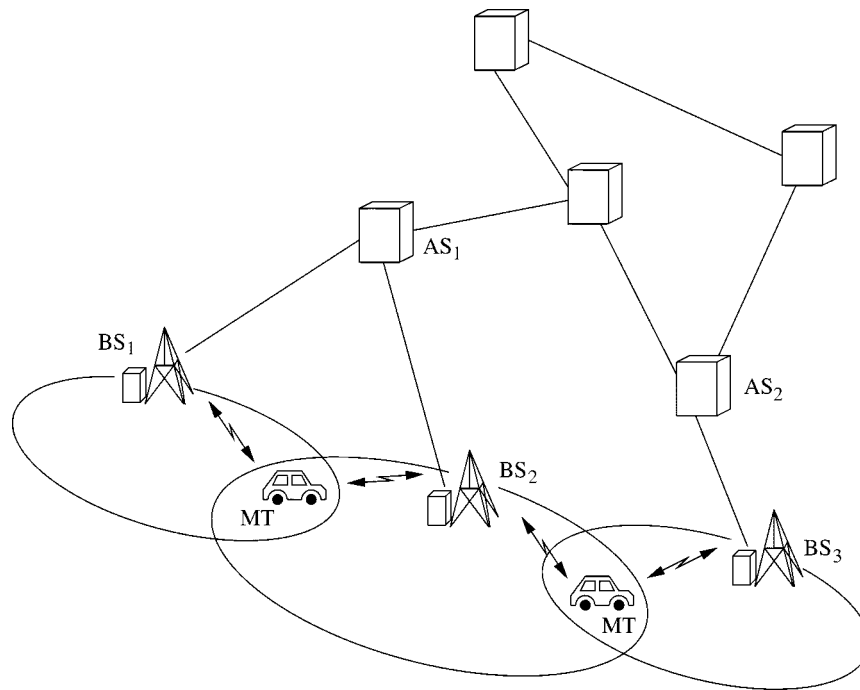


Fig. 1. The WATM architecture.

- MTs that freely roam in the geographical area covered by the cellular network (a terminal that moves from one BS to another must request a call handover for each of its established ATM connections);
- ATM switches, that are part of the wired network [these switches have mobility functionalities that allow an existing connection to be rerouted over a different (input or output) port if necessary].

Each BS is wired to one ATM switch and transmission between the two occurs according to standard ATM transfer capabilities [35], [36]. Since this paper focuses on high-speed data services, several transfer capabilities could be considered; we essentially refer to the case of ATM connections exploiting non-real-time variable bit rate (nrt-VBR), but the proposed approach can be modified to cope with other service classes. Transmission between the MT and the BS is achieved by means of a wireless interface capable of transmitting ATM cells [37]. The advantage of this approach is the resulting transparent connection between the wireless link and the wired ATM network that does not require any ATM adaptation layer (AAL) at the BS. According to this transmission scheme, cells generated by the MT are first transmitted over the wireless link, then forwarded over the *upstream* ATM connection. Similarly, cells transmitted by the terminal at the other end of the connection arrive at the BS via the *downstream* ATM connection and are then forwarded to the MT via the wireless link.

Contrary to conventional fixed ATM networks in which users require connections whose path is established at setup time and remains unchanged during the lifetime of the connection, the WATM network requires that the path of a connection is modified as the MT roams through distinct BSs. Fig. 2 shows the case of a MT that moves from BS₁ to BS₂ and, subsequently, from BS₂ to BS₃. The figure depicts the *incremental reestablishment* technique [11], which is used to update the path of the

connection(s) associated with the MT. According to this technique, when the MT moves from one BS, say BS₁, to another BS, say BS₂, an ATM switch is chosen to be the *crossover switch (CS)* between the old path (solid line) and the new connection path (solid line and dashed line). Only the portion of the new connection path between the crossover switch and BS₂ (dashed line) is established anew, while the portion of the old path overlapping with the new path of the connection is left unchanged. The same procedure is used when the MT moves from BS₂ to BS₃; the final connection path is shown as a solid and dotted line. Typically, the crossover switch is selected such that either the total number of hops of the new connection or the number of hops of the portion of the path that must be established anew is minimized.

A hard handover procedure is assumed in the system. Accordingly, during handover, connections are handled using a *break and make* approach.

III. SEAMLESS HANDOVER PROTOCOL WITH BUFFERING AT THE BASE STATION

The scope of the seamless handover protocol is to guarantee loss-free and in-sequence cell delivery to the end users during handover.

The sequence of events in the seamless handover protocol considered in this work is shown in Fig. 3. These events include actions taken by the MT, the designated CS, and the destination BS (indicated as BS₂). The MT sends a *handover request (HOR)* to the current BS (indicated as BS₁) notifying the intention to change BS. While waiting for the network reply, the MT continues to transmit and receive cells using the current connection path. During time interval $[\tau_1, \tau_2]$, an ATM switch is chosen to be the CS and informed of the handover request. After verifying that the necessary network resources are available at BS₂, the

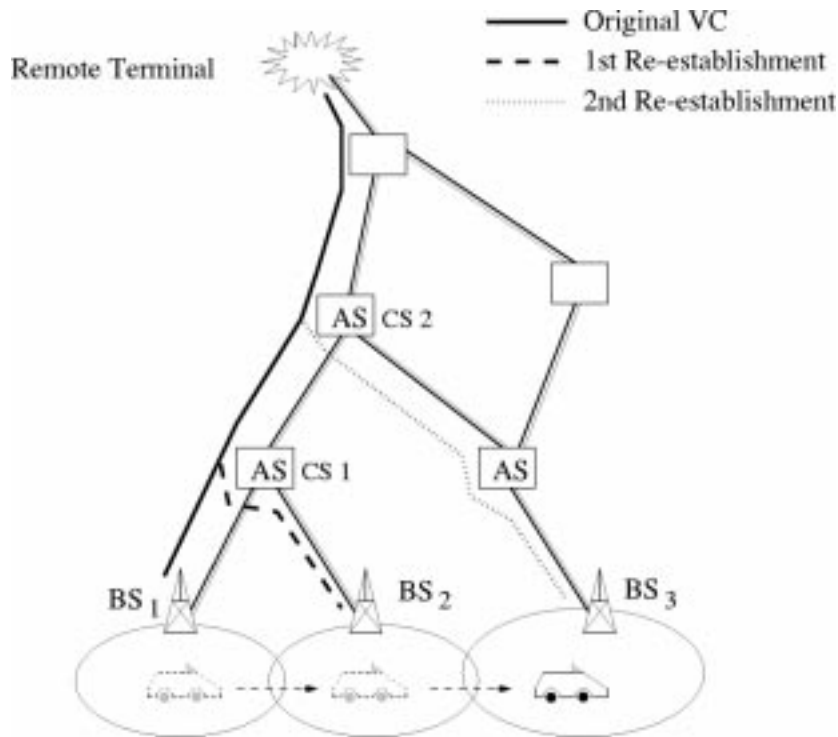


Fig. 2. The incremental path reestablishment technique.

CS notifies the MT that the handover is in progress by sending a *handover confirm (HOC)*. Soon after, the crossover switch is ready to initiate the traffic rerouting over the incremental path. Notice that the rerouting is performed on both the downstream traffic and the upstream traffic, not necessarily at the same time. Specifically, downstream cells are immediately transmitted over the incremental path, whereas the path for the upstream traffic will be rerouted only after the last upstream cell transmitted via BS₁ has reached the CS.

Upon reception of the handover confirm, the MT has received all the downstream cells transmitted via BS₁ [10]. Consequently, it disconnects from BS₁ by sending the *end upstream data flow (EDF)* message and attempts to establish a new wireless link with BS₂. Clearly, until the new wireless link is created, cells cannot be transmitted between the MT and the BSs. During this time interval, indicated as $[t_3, t_4]$, it may be necessary to store data cells at both MT (upstream traffic) and BS₂ (downstream traffic). More precisely, referring to Fig. 3, downstream cells are buffered if $\delta_1 < \delta_2$, i.e., downstream cells arrive at BS₂ before the new wireless link is established. Similarly, if $\delta_2 < \delta_4$, i.e., the *start data flow (SDF)* message reaches BS₂ earlier than the *upstream ready (USR)* message, upstream cells are buffered at BS₂ until the new upstream connection is established via the CS.

To contain the BS buffering requirement originated by each connection handover, the cell arrival rate at the BS buffer is kept below the cell departure rate as described next. At the terminal, a traffic shaping device based on the GCRA algorithm [35] maintains the terminal transmission rate below the sustainable cell rate (SCR). Cells stored in the BS buffer are transmitted at the connection peak cell rate (PCR). By controlling the gap between the PCR and the SCR, with $SCR < PCR$ it is thus possible to

determine how quickly the handover buffer will be emptied and consequently the minimum value to be contracted for the maximum burst size (MBS) of the ATM connection.

Two alternative policies are considered to control the handover buffer.

- 1) *Dedicated buffer (DB)*: for each (unidirectional) ATM connection of a MT performing handover, data cells are stored in a dedicated handover buffer.
- 2) *Buffer sharing at the base station (B²S²)*: connections of MTs concurrently requesting handover toward the same destination BS share a common handover buffer that is always available at the BS.

IV. THE GSPN APPROACH TO HANDOVER BUFFER MODELING

A significant cost of the BS is represented by the handover buffers. Optimization of the buffer size is a necessary step for designing a cost-effective solution based on the proposed seamless handover protocol. This section presents four GSPN analytical models that can be used to determine both the packet loss probability due to buffer overflow and the time required to empty the handover buffer as a function of the buffer size. The four models are derived for the upstream and downstream handover buffers in the DB and B²S² cases, respectively.

The steps followed to derive the models are:

- 2) definition of the handover phases that have significant impact on the handover buffer occupancy;
- 3) choice of the simplifying assumptions that allow the GSPN models to be built;
- 4) construction of the GSPN models and message level analysis;
- 5) ATM cell level analysis.

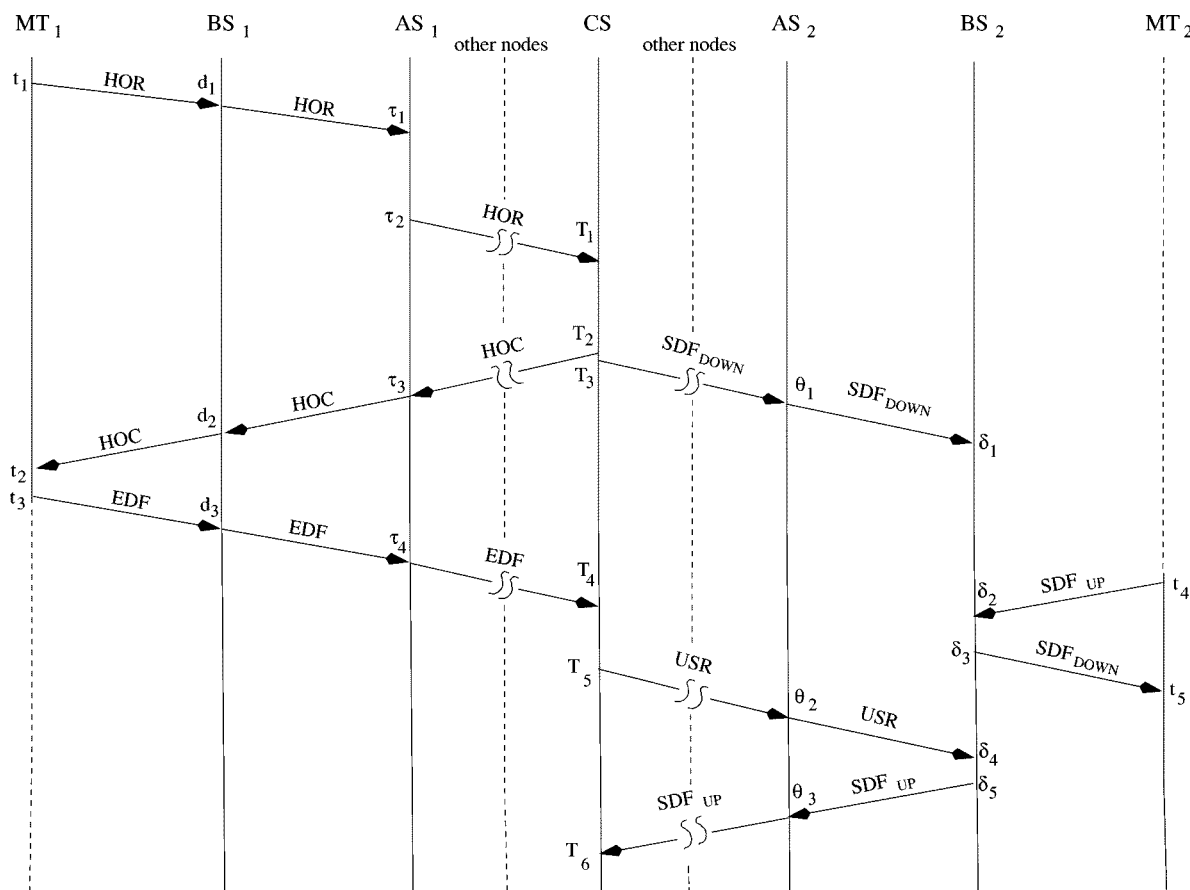


Fig. 3. Temporal diagram of the seamless handover protocol. MT₁ (MT₂) indicates that MT is connected to BS₁ (BS₂).

A key simplifying assumption of the proposed approach is the message level analysis, as opposed to the more conventional cell level analysis. The message level analysis considerably reduces the modeling complexity when compared to a cell level analysis. It thus allows the investigation of much more complex systems. Once the message level analysis is completed it is possible to obtain performance metrics at the ATM cell level with appropriate postprocessing of the message level results.

The following sections describe these steps with reference to the considered seamless handover protocol.

A. The Handover Phases

In order to study the occupancy of the handover buffers, it is first necessary to identify the different phases (or states) of a handover procedure. In the following we illustrate the *handover cycle*, defined as the sequence of phases in a single handover procedure starting with the handover request from the MT, and ending when the handover buffer is completely emptied.

As shown in Fig. 4, five phases are identified in one handover cycle.

- 1) Beginning of the handover, when MT and BS₁ still exchange cells—both the upstream and downstream handover buffers at BS₁ are bypassed by the cell flows (they are not necessary after the completion of the previous handover that brought the MT in the BS₁ cell).
- 2) No wireless link is available, neither between MT and BS₁, nor between MT and BS₂—newly generated

upstream cells are stored in the MT transmission buffer and the downstream ATM connection downstream virtual circuit (VC down) is being rerouted at the CS toward BS₂.

3) MT is disconnected from BS₁ and not yet connected to BS₂, the rerouted downstream ATM connection is available—upstream cells are stored in the MT transmission buffer; downstream cells are stored in the downstream handover buffer of BS₂.

4) MT and BS₂ exchange cells via the newly established wireless link and the upstream ATM connection is being rerouted at the CS—upstream cells are stored in the upstream handover buffer of BS₂; downstream cells are delivered to MT.

5) MT and BS₂ exchange cells via the newly established wireless link and the rerouted ATM connection is available—the BS₂ handover buffers are being emptied.

6) MT transmits (receives) to (from) BS₂ and handover buffering is not any longer necessary, since the handover is completed (all buffers have been emptied).

Two alternative phase sequences are possible during a handover cycle, depending on whether the rerouting of the ATM connection is completed before or after the new wireless link between the MT and BS₂ is established. If the connection rerouting is completed first, the phase sequence is *a, b, c, e, a'*, and buffering is required for downstream cells only, during

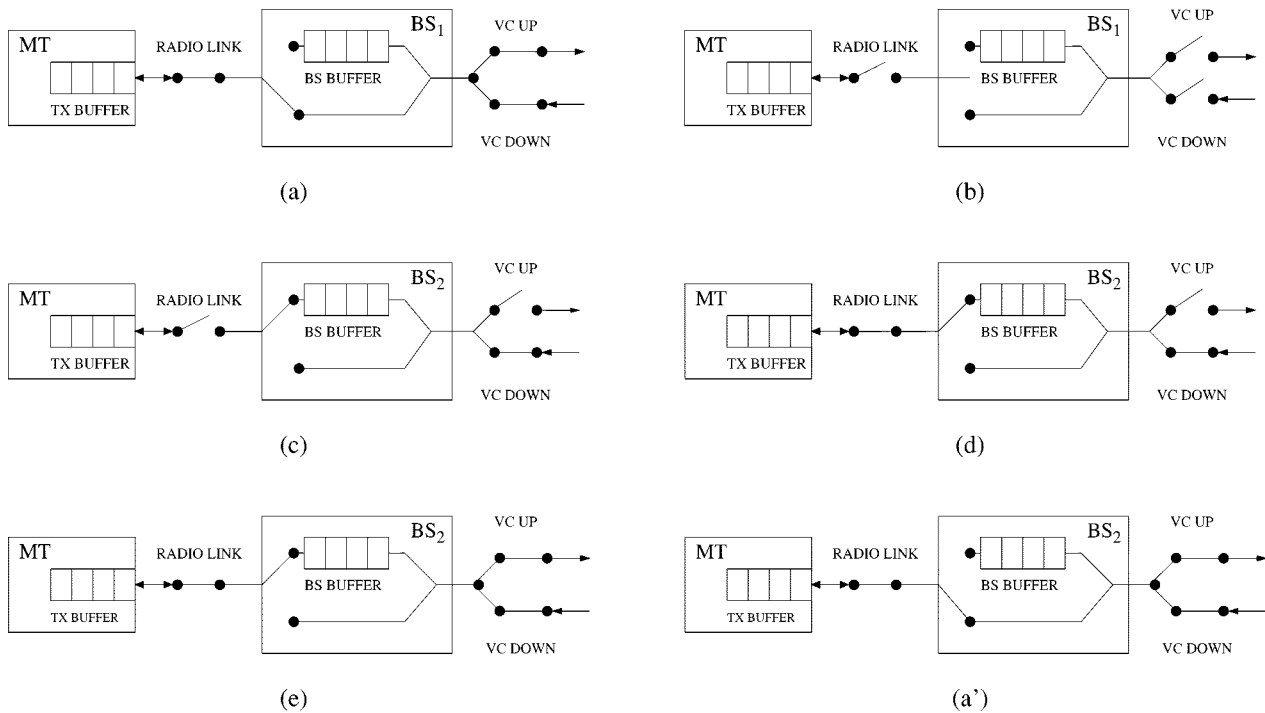


Fig. 4. Sequence of phases of the handover buffers during the execution of one handover.

phases *c* and *e*. If the wireless link with BS₂ is established first, the phase sequence is *a*, *b*, *d*, *e*, *a'*, and buffering is required for upstream cells only, during phases *d* and *e*. It is important to notice that during phase *e*, the handover buffer is filled at the remote terminal transmission rate (that on the average equals SCR) and emptied at PCR. During this phase, the handover buffer is therefore emptied at the differential rate PCR – SCR.

Under the DB policy, one distinct upstream (downstream) buffer is reserved at the BS for every upstream (downstream) connection requesting handover from phase *b* to the end of phase *e*. The buffer is released at the end of phase *e*. Under the B²S² policy there is only one upstream (downstream) buffer always available at the BS that is shared by the upstream (downstream) connections concurrently requesting handover toward the same destination BS. Notice that each handover is characterized by its own phase sequence and phases of different handovers are not necessarily synchronized.

No head-of-the-line blocking is assumed for cells of different connections: the available handover buffer space is shared by all connections, but each connection has its own logical queue (if a cell of connection *X* arrives at the buffer before a cell of connection *Y*, and the rerouting of connection *X* is not yet completed while connection *Y* has already been rerouted, the cell of connection *Y* needs not wait).

B. Modeling Assumptions

The distinct phases of the handover procedure are modeled using the GSPN formalism [14], a tool that allows a concise representation of the system to be modeled at the desired level of abstraction. The formalism consists of the following basic components.

Places that represent the state of the system; they are indicated by circles.

Transitions that describe events that modify the system state. They can be immediate (black bars) or timed (white rectangles): the former represent instantaneous events (i.e., logic actions) that have null activation time; the latter represent time consuming actions with associated exponentially distributed random firing times.

Arcs that define the relation between states and events by means of input and output functions represented by arrow-headed arcs; or *inhibitor* functions identified by circle-headed arcs.

Tokens—that identify the state of the system; they are represented as markers in places and move from one place to another via the firing of some transitions. Tokens can be *colored*, i.e., have an identity that allows their individual behavior to be distinguished; in this case we refer to the formalism as colored GSPN (CGSPN).

The exponential distribution of firing delays associated with the timed transitions implies that the stochastic process described by a GSPN is Markovian.

As already mentioned, GSPN and CGSPN models have been successfully used in a number of studies of complex telecommunication systems, providing a simple but powerful paradigm for the construction of stochastic models of several different networking scenarios.

In constructing the four GSPN models some assumptions are introduced.

- 1) All delays in the system must be taken to be exponentially distributed random variables, as required by GSPN models. While this assumption entails an approximation, since delays often are far from being exponentially distributed, the comparison of the performance estimates generated by the GSPN models and the results obtained by detailed simulations indicate that the shape of the dis-

tribution is not critical for the accurate prediction of the system performance. On the other hand, some statistical analyses of field measurement data indicate that some of the delays that we consider in the model are actually close to being exponentially distributed; for example, the handover traffic is Poissonian in a nonblocking environment [38], [39].

- 2) The time between two handover requests from the same MT is assumed to be longer than the time required to complete the handover cycle. Thus, when the MT requests a new handover, the handover buffers utilized in the previous handover of the same MT are empty. In other words, two handover cycles of the same MT cannot overlap. This assumption should be realistic in most cases, since the cellular network should be designed to avoid an excessive handover frequency in order to limit the amount of signaling and control. The few handover instances which do actually violate this assumption should thus only have a marginal impact on the overall system performance. Note that the results produced by the GSPN model can be instrumental in the selection of the network parameters (minimum cell size, maximum MT speed, etc.) that indeed do avoid an excessive handover frequency by providing estimates of the time necessary to complete the handover procedure.
- 3) The propagation delay in the WATM network is negligible with respect to the time necessary to the rerouting of the ATM connection. The latter time is a random variable that depends on the congestion level at the CS.
- 4) The MT average offered load is denoted by L_o Mb/s, a varying system parameter. Bursty traffic is modeled generating messages whose length is a random number of cells. The number of cells in each message ν_b is taken to be geometrically distributed with given mean. This is a fairly realistic assumption, specially if referred to communication services exploiting the internet protocol (IP) protocol, whose message size can vary from very few cells to tens and even hundreds of cells. Notice, however, that if a message length distribution other than the geometric distribution is considered, the GSPN model needs to be modified and its complexity may significantly increase. For instance, in the case of Erlang or hyperexponential distributions, the message transmission time from MT to the BS has to be approximated by using several transitions with exponentially distributed firing times. A similar consideration holds for the transmission time of messages stored in the upstream handover buffer at the BS. We also highlight that the accuracy of the GSPN model has been tested for geometrically distributed message lengths only.
- 5) A representation at the message level (instead of cell level) is used to limit the growth of the number of states in the underlying Markovian process. This approach has the advantage of generating a state space whose size is not a function of the number of cells in the messages. Cell-level performance parameters can nevertheless be derived from the message-level description, as explained later.

In spite of these simplifying assumptions, as shown in Section V, the proposed GSPN models are as accurate as detailed simulation experiments, which require much longer CPU times and cannot provide satisfactory results when the observed events are rare.

C. The GSPN Models

The models developed for the dedicated and shared buffer policies are described next. Since the behaviors of the upstream and downstream cell transmissions are decoupled, for each policy two distinct, but in principle similar GSPN models are developed, one for the upstream buffer and the other for the downstream buffer.

Each model comprises two interacting components (or subnets) that respectively describe: 1) the phases of the handover cycle and 2) the impact of each phase on the flow of cells in the BS handover buffer.

1) Dedicated Buffering (DB):

a) *Upstream Handover Buffer*: The model focuses on the flow of messages generated by the MT and sent over the connection via BS₁ before the handover, and via BS₂ after the handover.

The first GSPN model component is shown in the left part of Fig. 5. The handover cycle begins when transition *start_HO* fires removing the token from place CONNECTED. The firing of this transition, with rate μ_{DC} , indicates that the handover is moving from phase *a* to phase *b*. The firing of *start_HO* generates a token in two places: NO_VC and NO_RADIO_LINK. The marking of place NO_RADIO_LINK indicates that MT is disconnected from BS₁ and not yet connected to BS₂; the marking of place NO_VC indicates that the (upstream) ATM connection rerouting has not yet been completed. The marking of these two places enables the two concurrent and timed transitions *get_radio_link* and *get_VC*, respectively. The former transition has rate μ_{NC} , and its firing represents the establishment of the wireless link between MT and BS₂. The latter has rate μ_{UT} , and represents the rerouting of the ATM connection from BS₁ to BS₂.

If transition *GET_RADIO_LINK* fires first, a token is generated in place RADIO_LINK, to indicate the availability of the wireless link from MT to BS₂ (this marking is equivalent to phase *d*, when the upstream cells are stored into the upstream handover buffer). When transition *GET_VC* then fires, the token in place NO_VC is moved into place VC, to indicate the availability of the ATM connection toward BS₂.

Alternatively, if transition *get_VC* fires first, a token is generated in place VC, before the marking of place RADIO_LINK, and no message is buffered in the upstream handover buffer because the ATM connection is already open when MT establishes the wireless link to BS₂.

When a token is present in both places RADIO_LINK and VC, transition *got_both* fires, and a token is generated in place RECONNECTED (phase *e*). The token is removed from place RECONNECTED through the firing of transition *end_HO* after that the upstream handover buffer has been emptied and the handover cycle has thus been completed.

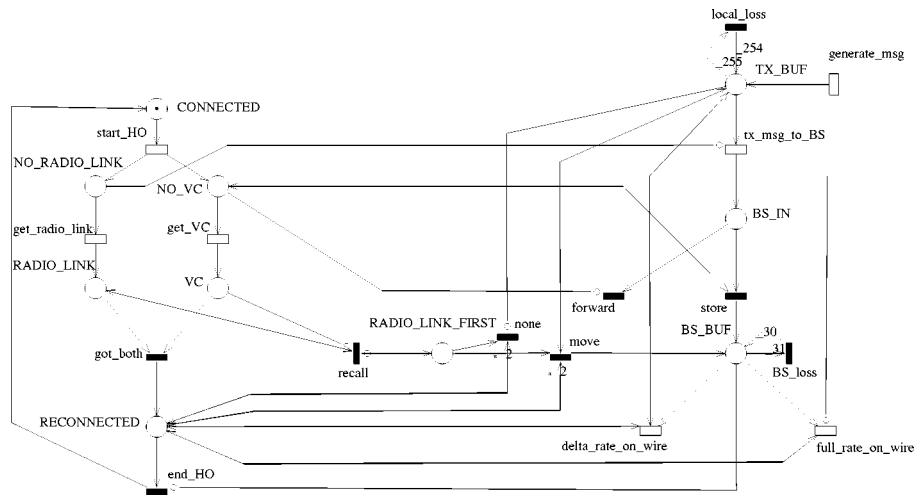


Fig. 5. The GSPN model for the design of the dedicated upstream handover buffer.

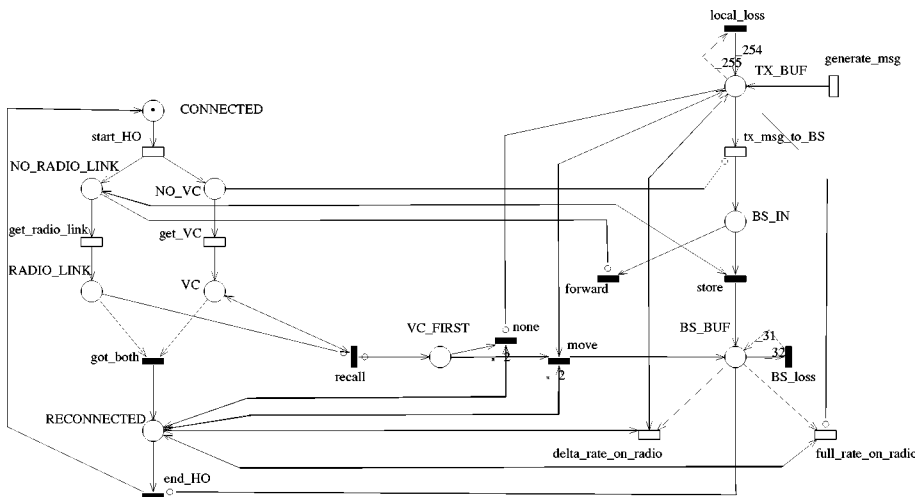


Fig. 6. The GSPN model for the design of the dedicated downstream handover buffer.

The second GSPN model component is shown in the right side of Fig. 5. Transition *generate_msg* represents the generation of user messages according to a Poisson process with rate μ_{gen} . Each token generated into place *tx_BUF* represents one message stored in the MT transmission buffer. The capacity of this buffer is limited, and overflow messages are discarded by the firing of transition *local_loss* (the arc weights in Fig. 5 indicate that the buffer can hold at most 254 messages; when a 255th token enters place *tx_BUF*, transition *local_loss* fires, destroying it). Rate μ_{gen} is computed from the MT average offered load (L_o Mb/s) and the mean number of cells per message (ν_b) using the relation: $\mu_{gen} = (L_o/424\nu_b)$.

Transition *tx_msg_to_BS* models the transmission of messages from MT to the BS to which it is connected (either BS_1 or BS_2); this transmission is interrupted in the periods when no connection is available between the MT and the BS (after disconnecting from BS_1 and before connecting to BS_2). For this reason an inhibitor arc from place *NO_RADIO_LINK* to transition *tx_msg_to_BS* is used. The firing rate of transition *tx_msg_to_BS*, that represents the MT transmitting on the wireless link at rate SCR , is $\mu_r = (SCR/424\nu_b)$.

The firing of transition *tx_msg_to_BS* indicates that all cells belonging to the same message have been transmitted by the MT. For each transmitted message, a token is deposited into place *BS_IN*. When a token reaches place *BS_IN* and place *NO_VC* is empty (i.e., the ATM connection is available), transition *forward* fires and removes the token from the place (the message's cells need not be stored in the upstream handover buffer.) When a token reaches place *BS_IN* and place *NO_VC* is marked (i.e., the ATM connection is not available), transition *store* fires and moves the token into place *BS_BUF* that represents the upstream handover buffer where the message is stored. The capacity of this buffer is assumed finite and overflow messages are discarded through the firing of transition *BS_loss*. In this case, the arc weights in Fig. 5 indicate that the buffer capacity is 30 messages (the maximum marking of both places *BS_BUF* and *TX_BUF* is a parameter of the model, defined with a trade off between message loss avoidance and state-space reduction).

Transmission of messages stored in the upstream handover buffer is represented by two mutually exclusive transitions *delta_rate_on_wire* and *full_rate_on_wire*, enabled when place *RECONNECTED* is marked (phase *e*). If place *TX_BUF* is

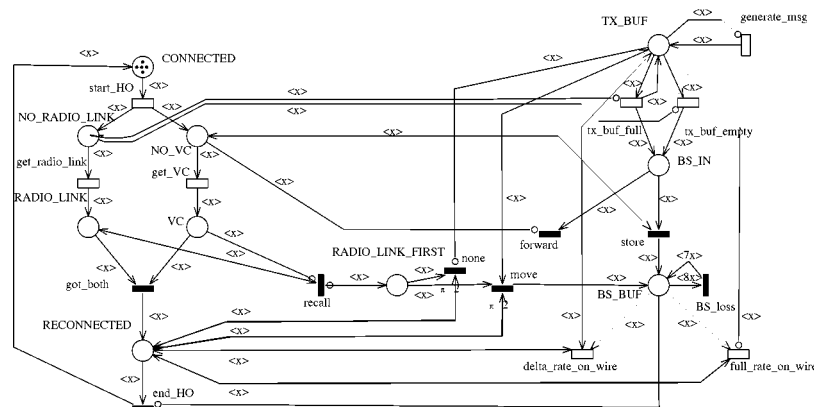


Fig. 7. The CGSPN model for the shared upstream handover buffer.

empty (the MT transmission buffer contains no messages), transition *full_rate_on_wire* is enabled, indicating that the number of messages in the handover buffer decreases at the message transmission rate determined by the connection PCR. The firing rate of transition *full_rate_on_wire* is therefore $\mu_{\phi} = (PCR/424\nu'_b)$, where ν'_b is the average number of ATM cells comprising a message stored in the handover buffer.² Alternatively, if place TX_BUF is marked (the MT transmission buffer contains messages), transition *delta_rate_on_wire* is enabled, indicating that the number of messages in the handover buffer decreases at the *relative* message transmission rate determined by the difference between the cell departure rate, i.e., PCR, and the cell arrival rate, i.e., SCR. The firing rate of transition *delta_rate_on_wire* is therefore $\mu_{\delta} = (PCR - SCR/424\nu'_b)$. Notice that in the latter case tokens flowing to place BS_IN and modeling the messages transmitted over the wireless link are discarded through transition *forward* to guarantee conservation of flow.

When place BS_BUF is eventually emptied the handover terminates and transition *end_HO* fires.

Due to the message level nature of the GSPN, three transitions (*recall*, *none*, *move*) and one place (RADIO_LINK_FIRST) are added to the model. If the MT connects to BS₂ before the ATM connection is rerouted, it may happen that at the time the ATM connection is rerouted (firing of *get_VC*) a portion of a message (subset of cells) is still at the MT while the rest is stored in the handover buffer. This is modeled in the GSPN using two distinct tokens, one representing the message portion transferred to the handover buffer (in place BS_BUF), the other representing the portion left in the MT buffer (in place TX_BUF). [The splitting of the (message) token into two is possible due to the memoryless property of the exponentially distributed length of the message.] Place RADIO_LINK_FIRST and immediate transition *recall* keep track of whether the wireless link to BS₂ is established before the rerouting of the ATM connection. Immediate transitions *none* and *move* generate the (second) token in place BS_BUF if the MT transmission buffer is not empty when transition *get_VC* fires. (Notice that immediate transitions *move* and *none* have higher priority than transition *end_HO*, as indicated by labels $\pi 2$, which set their priority to two, higher than the default value of 1 given to the other immediate transitions.)

²The value of ν'_b is derived in Section IV-D.

The initial marking of the GSPN model consists of just one token in place CONNECTED.

b) *Downstream Handover Buffer*: The model focuses on the flow of messages generated by the remote terminal and reaching the MT via BS₁ before the handover, and via BS₂ after the handover.

The GSPN model is shown in Fig. 6. The left side of the model is identical to the model in Fig. 5. The right side of the model is different because in this case messages arrive from the remote terminal:

- 1) transition *tx_msg_to_BS* is now disabled by the presence of a token in place NO_VC since cells cannot reach the BS while the ATM connection is being rerouted;
- 2) mutually exclusive transitions *forward* and *store* depend now on the marking of place NO_RADIO_LINK since cell buffering at BS₂ occurs while the new wireless link is not yet established;
- 3) transition *recall* fires if the ATM connection rerouting takes place *before* the establishment of the wireless link between BS₂ and MT.

2) *Shared Buffering (B²S²)*: When the handover buffer is shared, all concurrent connections requesting handover toward the same BS must be modeled simultaneously to take into account the interaction between the buffered cells. Although this is possible with the GSPN model introduced for the DB approach, its complexity becomes unacceptably high with already few connections. This is due to the fact that the DB model must be duplicated for each concurrent handover.

To overcome this problem, the CGSPN formalism is used to model the B²S² approach. With CGSPN, the complexity of the model layout remains substantially the same of the GSPN model proposed for the DB case, with the difference that each handover is associated with a distinct token color. In addition to the compact representation of the model, the CGSPN approach allows to automatically fold statistical equivalent states into a single state, thus significantly reducing the number of state probabilities that must be solved numerically to compute the performance estimates.

The models of the upstream and downstream shared buffers are similar to the nets built to represent the dedicated buffer case. Fig. 7 shows the net representing the upstream shared buffer. (Due to space limitation, the model of the downstream

shared buffer is not shown as its construction is straightforward.) Place BS_IN represents the entire buffer capability available at the BS that is shared among the MTs performing concurrent handover procedures. In order to decrease the number of states generated by the CGSPN model, a slightly different representation of the MT transmission buffer is used in this case. In the CGSPN model the marking of place TX_BUF can never be greater than one, and transition $tx_msg_to_BS$ is replaced by transitions tx_buf_full and tx_buf_empty . By firing, either transitions move one token from place TX_BUF to place BS_IN, the former replacing the token in TX_BUF, the latter emptying the input place. The firing rate of transition tx_buf_full is equal to the firing rate of transition $tx_msg_to_BS$ in the GSPN model conditioned to the markings that have a number of tokens in place TX_BUF greater than one. The firing rate of transition tx_buf_empty is equal to the firing rate of transition $tx_msg_to_BS$ in the GSPN model conditioned to the markings with one token in place TX_BUF. Transition $generate_msg$ is enabled only when place TX_BUF becomes empty.

Besides the aforementioned changes, each set of colored tokens flows in the net as illustrated in the GSPN model.

D. Cell Level Analysis

The results provided by the solution of GSPN and CGSPN models (i.e., the steady-state probabilities of the individual markings) refer to messages. This section describes how to derive cell-related performance parameters from these results. For the sake of brevity, we only present the analysis for the GSPN model of Fig. 5. The same approach can be applied to the dedicated downstream handover buffer, the shared upstream and downstream handover buffers.

Any token in place TX_BUF represents a message (or the tail portion of a message) generated by MT. The average number of cells associated with each token is thus ν_b . Any token in place BS_BUF represents a message (or the head portion of a message) that has been transmitted over the wireless link while the ATM connection is not yet reestablished. The average size of this message is thus conditioned to the fact that the message is transmitted over a time span that is bounded from above by the rerouting time of the ATM connection. In the GSPN model this situation is represented by the concurrent enabling of transition $tx_msg_to_BS$ and transition get_VC . The average number of cells associated with a token in place BS_BUF is therefore

$$\nu'_b = \frac{SCR}{424(\mu_r + \mu_{UT})}$$

where SCR is the cell transmission rate over the wireless link.

As all timed transitions in the GSPN model have exponentially distributed firing times, irrespective of the average message length, any token associated with a message or a portion of a message represents a geometrically distributed number of cells.

Let $e(i)$, $\forall i \geq 0$, denote the probability that a token represents i cells [with $e(0) = 0$, i.e., any token represents at least

one cell]. Then, $E_k(i)$ defined as the probability that k tokens all together represent i cells is

$$E_0(0) = 1 \quad (1)$$

$$E_1(i) = e(i) \quad \forall i > 0 \quad (2)$$

$$E_k(i) = 0 \quad \forall i, k: i < k \quad (3)$$

$$E_k(i) = \sum_{j=1}^{i-k+1} E_{k-1}(i-j)E_1(j) \quad \forall k > 1, i \geq k. \quad (4)$$

Using the above expressions, once the GSPN model is numerically solved, and the steady-state probabilities of the markings are obtained, the cell performance metrics of interest can be derived. We limit our analysis to one handover cycle since the obtained results can be easily extended to the case of concurrent handovers.

The probability that during phases d and e of the handover cycle the number of tokens in place BS_BUF is equal to k is defined as the probability that the number of tokens in place BS_BUF is equal to k , given that either the number of tokens in place RADIO_LINK is equal to one or the number of tokens in place RECONNECTED is equal to one: $P\{\#BS_BUF = k | \#RADIO_LINK = 1 \text{ or } \#RECONNECTED = 1\}$. This probability is simply the sum of the steady-state probabilities of all the markings of the GSPN that satisfy this condition and can be analytically derived. The probability that the upstream handover buffer contains n ATM cells is then

$$B(n) = \sum_{k=0}^n P\{\#BS_BUF = k | \#RADIO_LINK = 1 \text{ or } \#RECONNECTED = 1\} E_k(n) \quad \forall n \geq 0. \quad (5)$$

An upper bound to the cell loss probability due to buffer overflow is

$$P_{\text{loss}}(m) = \sum_{k=m+1}^{\infty} B(k) \quad (6)$$

where m is the buffer capacity.

The average time spent by a cell in the handover buffer depends on the phase in which the cell reaches the buffer. A cell that reaches the handover buffer during phase d leaves the buffer only after the upstream connection has been rerouted and all the cells already in the buffer have been transmitted. A cell that reaches the buffer during phase e remains in the buffer only for the time necessary to transmit the cells already present in the buffer. Therefore, distinct computations are required for the two phases: d and e . The probabilities that a cell reaching the handover buffer finds n cells in it are, respectively,

$$Q_d(n) = \sum_{k=0}^n P\{\#BS_BUF = k | \#RADIO_LINK = 1 \text{ or } \#TX_BUF > 0\} E_k(n) \quad (7)$$

$$Q_e(n) = \sum_{k=0}^n P\{\#BS_BUF = k | \#RECONNECTED = 1 \text{ or } \#TX_BUF > 0\} E_k(n). \quad (8)$$

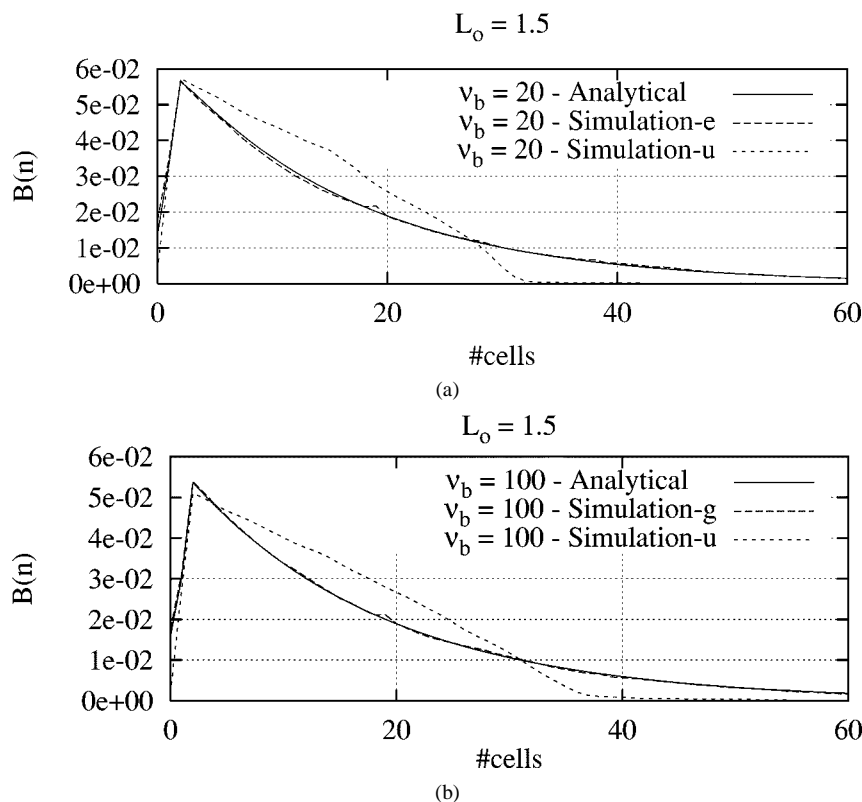


Fig. 8. Probability density function of the number of cells stored in the upstream handover buffer with mean rerouting time $\mu_{UT}^{-1} = 4$ ms, SCR = 1.9 Mb/s, average offered load $L_o = 1.5$ Mb/s, and for different values of mean message length (ν_b [cells]). Simulation results are obtained for an upstream rerouting time geometrically (g) and uniformly (u) distributed. (a) $\nu_b = 20$. (b) $\nu_b = 100$.

Then, the average delay encountered by a cell stored in the handover buffer is

$$\begin{aligned}
 M_d = & \left[\frac{1}{\mu_{UT}} + \left(\sum_n n Q_d(n) \right) \tau_{PCR} \right] \\
 & \cdot P\{\# \text{RADIO_LINK} = 1, \# \text{TX_BUF} > 0\} / P_T \\
 & + \left(\sum_n n Q_e(n) \right) \tau_{PCR} \\
 & \cdot P\{\# \text{RECONNECTED} = 1, \# \text{TX_BUF} > 0\} / P_T \\
 & \forall n \geq 0
 \end{aligned} \quad (9)$$

being τ_{PCR} the cell transmission time at rate PCR and

$$\begin{aligned}
 P_T = & P\{\# \text{RADIO_LINK} = 1, \# \text{TX_BUF} > 0\} \\
 & + P\{\# \text{RECONNECTED} = 1, \# \text{TX_BUF} > 0\}.
 \end{aligned} \quad (10)$$

To compute the worst case MBS we consider the maximum acceptable value of cell loss probability and using (6), we derive m ; that is also the maximum number of cells stored in the buffer. Then, we have

$$\text{MBS}_{wc} = \frac{424m}{\text{PCR} - \text{SCR}}. \quad (11)$$

Finally, we denote as $h_{UP}(\epsilon)$ the maximum frequency for the upstream connection to request handover. This frequency is chosen so as to avoid that consecutive handover requests for the same connection overlap with probability higher than ϵ . We define the two intervals RD and

RT as $P\{\text{radio interruption time} > RD\} \leq \epsilon$ and $P\{\text{upstream rerouting time} > RT\} \leq \epsilon$. Then, we have

$$h_{UP}(\epsilon) = \min \left(\frac{1}{RD}, \frac{1}{RT + \text{MBS}_{wc}} \right). \quad (12)$$

The possibility of deriving accurate (as we shall see) cell level performance metrics from a message level analysis is quite a relevant result since it yields a significant reduction of the stochastic model state-space size, hence, of the complexity of the model solution.

V. NUMERICAL RESULTS

To validate the accuracy of the GSPN models, this section presents first a comparison of the performance estimates obtained by both numerically solving the GSPN models and running simulation experiments. The numerical analysis of GSPN models exploits the GreatSPN software [40], a standard tool for performance evaluation with GSPN and CGSPN. Simulation runs are based on the CLASS software tool (cell level ATM services simulator) [41], that allows the simulation of ATM networks at the time scale of individual cells and groups of cells. Observe that while GSPN models operate at the message level, CLASS operates at the cell level. The two descriptions of the system dynamics are thus quite different. CLASS adopts rather sophisticated statistical techniques for the estimation of the confidence level and accuracy of its performance estimates [41]. A good match between simulation and GSPN performance predictions thus constitutes a reasonable validation of the proposed modeling approach.

The network setup used in the validation process consists of two BSs connected to the crossover switch. The propagation delay between either BS and the CS is set equal to 0.2 ms. The time needed to establish the radio link between MT and BS₂ is taken as a uniformly distributed random variable with mean value equal to 500 μ s. All simulation results presented in this section have 5% accuracy and 99% confidence level. On a Pentium Pro II PC, simulation runs last approximately 5 h, as opposed to the few seconds required to numerically solve the GSPN model.

In this validation study, some system parameters are fixed to reduce the number of validation experiments. The numerical values used for the validation are selected aiming at typical cases, rather than providing an exhaustive evaluation of all potential system scenarios. Since our main objective in this paper is the illustration and validation of the modeling approach, we do not delve very deep in the investigation of the impact of the various parameter values on the system performance.

Fixed values are used for the PCR, set equal to 2.0 Mb/s—a traffic rate currently considered for WATM systems—the average burst length, set to $\nu_b = 20$ ATM cells (unless a different value is explicitly declared)—approximately 1 Kb, a size comparable to Ethernet packets—and the average time necessary to establish the wireless link between MT and BS₂, $\mu_{NC}^{-1} = 500$ μ s. In the case of dedicated buffer, we assume $\mu_{DC}^{-1} = 1$ s, i.e., the MT generates a new handover request on average 1 s after the completion of the previous handover. The interval between two successive handover requests is set artificially small to generate a large number of handover requests in the models. In the case of shared buffering, we assume that μ_{DC} is a varying parameter that controls the average number of concurrent handovers. The other varying system parameters are L_o , SCR, and μ_{UT} .

Note that we always take L_o to be less than SCR. If this is not the case, the GSPN model still produces meaningful results, but the probability of overflow of the MT buffer becomes non negligible as the average input rate exceeds the average output rate. We consider this situation of significant loss probability at the MT not to be of interest.

First, the GSPN model defined for the dedicated upstream handover buffer is validated through comparison with simulation results. In particular, the probability density function of the number of cells stored in the dedicated upstream handover buffer, i.e., the probability $B(n)$ that the upstream handover buffer contains n cells, is used to carry out the comparison.³ In Fig. 8, $B(n)$ is plotted for L_o equal to 1.5 Mb/s, SCR = 1.9 Mb/s, and $\mu_{UT}^{-1} = 4$ ms [42]. Two possible average message lengths are considered, i.e., 20 and 100 cells. Simulation results are obtained considering two possible distributions of the upstream rerouting time: geometric and uniform. In either case, the mean value equals 4 ms. Results obtained by the GSPN model are accurate in all shown cases. In particular, we notice that even when the rerouting time is assumed to be uniformly distributed, the model is able to capture the behavior of the buffer occupancy fairly well. Simulation results presented in the rest of the paper

³To estimate the probability density function of the number of cells in the upstream handover buffer, the buffer size is chosen to be sufficiently large to avoid cell losses.

TABLE I
AVERAGE CELL DELAY IN THE UPSTREAM HANDOVER BUFFER FOR
VARYING VALUES OF L_o , SCR, AND μ_{UT}

L_o (Mbps)	SCR (Mbps)	$\mu_{UT}(\text{ms}^{-1})$	M_d (ms)	
			Analys.	Simul.
0.5	1.9	0.25	4.02	4.05
1.5	1.9	0.25	3.97	3.98
1.8	1.9	0.25	4.03	4.09
1.5	2.0	0.25	3.98	4.06
0.5	1.9	0.1	9.69	9.92

are thus obtained assuming a geometrically distributed upstream rerouting time.

Table I shows values obtained for M_d , the average delay experienced by the cells in the upstream handover buffer for varying values of L_o , SCR, and μ_{UT} . The GSPN estimates always fall within the confidence interval of the simulator point estimates.

With respect to the dedicated downstream handover buffer, Fig. 9 shows $B(n)$ values obtained for SCR = 1.9 Mb/s, $L_o = 0.5$ Mb/s and $L_o = 1.8$ Mb/s, respectively. The GSPN results are obtained assuming that transition get_VC in Fig. 6 is immediate, hence, supposing that the rerouting of the downstream ATM connection always takes place before the establishment of the wireless link between MT and BS₂, as normally happens with reasonable system parameters. The match between simulation results and GSPN performance predictions is not as accurate as in the previous model. However, it should be noted that the number of cells stored in the downstream handover buffer is quite small, thus the problem of correctly dimensioning this buffer is less critical when compared to the correct sizing of the upstream handover buffer.

Likewise, the validation of the CGSPN model of the shared handover buffers is done considering $B(n)$, i.e., the probability density function of the number of cells stored in either handover buffers. Two scenarios are considered: one with an average number of concurrent handovers equal to $N_{HO} = 1.5$ and the other with an average number of concurrent handovers equal to $N_{HO} = 3$. In either cases, no more than five concurrent handovers are allowed. The values for the other parameters are SCR = 1.9 Mb/s, and $\mu_{UT}^{-1} = 4$ ms. Fig. 10 shows the good agreement between the distributions of the number of cells in the shared upstream handover buffer resulting from the analytical and simulation studies when $L_o = 1.0$ Mb/s and 1.5 Mb/s. Likewise, Fig. 11 illustrates the comparison between the numerical and simulation results obtained for the shared downstream handover buffer when $L_o = 1.5$ Mb/s.

As shown in the figures, with symmetrical traffic, the required size of the downstream buffer is always smaller than the required size of the upstream buffer in both the dedicated and shared buffer approaches. We therefore limit the rest of this analysis to the latter buffer.

Results obtained using the GSPN models allow us to optimally determine the size of the dedicated and shared buffers as a function of several system parameters, even when the occurrence of the measured events is rare. Fig. 12 shows the cell loss probability versus the size of the dedicated upstream handover

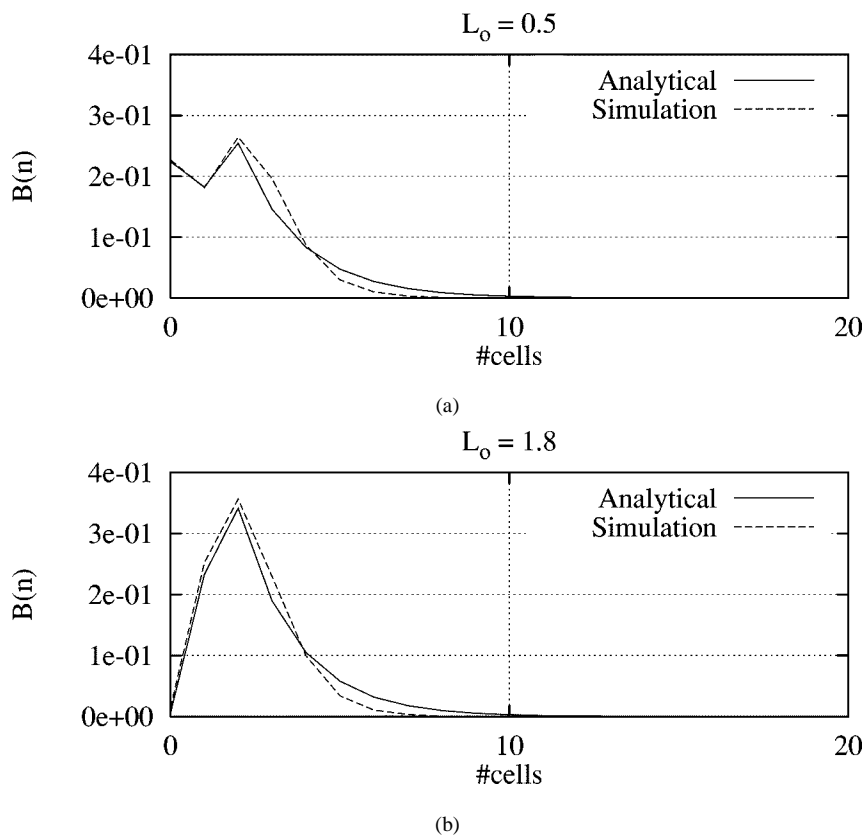


Fig. 9. Probability density function of the number of cells stored in the downstream handover buffer, for SCR = 1.9 Mb/s, and variable values of average offered load (L_o [Mb/s]). (a) $L_o = 0.5$. (b) $L_o = 1.8$.

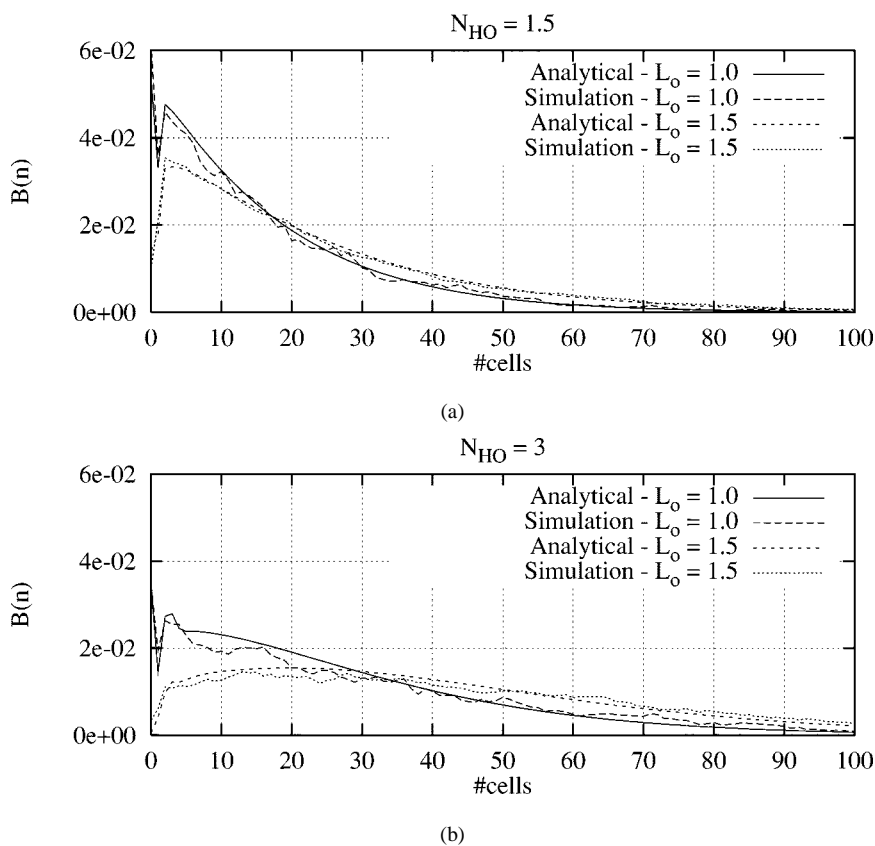


Fig. 10. Probability density function of the number of cells stored in the shared upstream handover buffer, for average offered load $L_o = 1.0$ and 1.5 Mb/s, SCR = 1.9 Mb/s, mean rerouting time $\mu_{UT}^{-1} = 4$ ms, and different average number of concurrent handovers. (a) $N_{HO} = 1.5$. (b) $N_{HO} = 3$.

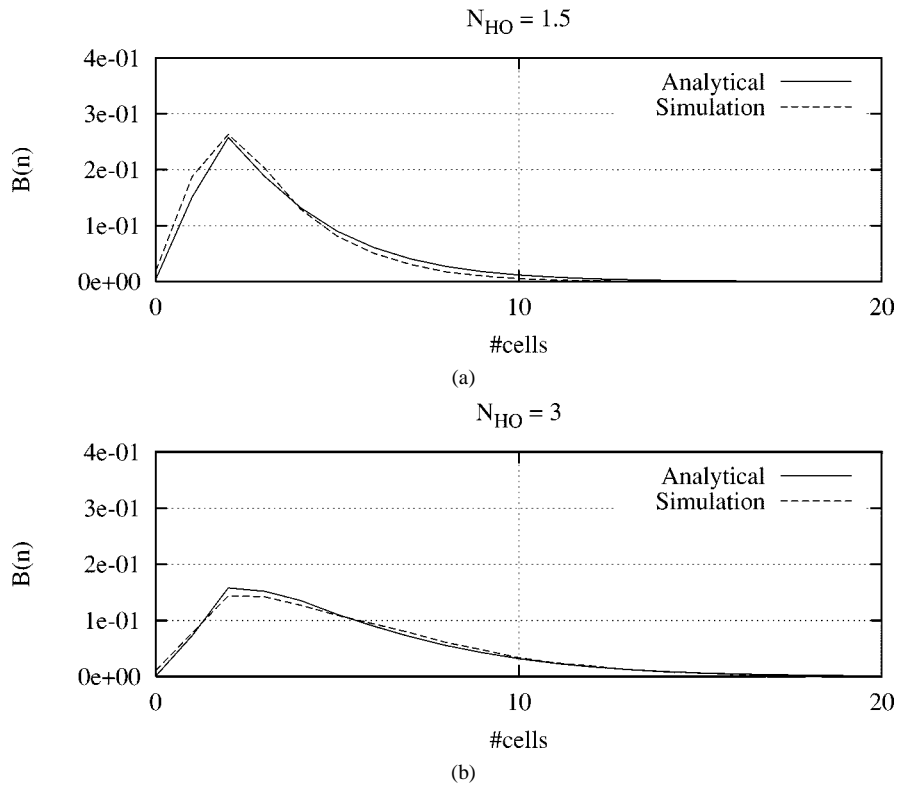


Fig. 11. Probability density function of the number of cells stored in the shared downstream handover buffer, for average offered load $L_o = 1.5$ Mb/s, SCR = 1.9 Mb/s, mean rerouting time $\mu_{UT}^{-1} = 4$ ms, and different average number of concurrent handovers. (a) $N_{HO} = 1.5$. (b) $N_{HO} = 3$.

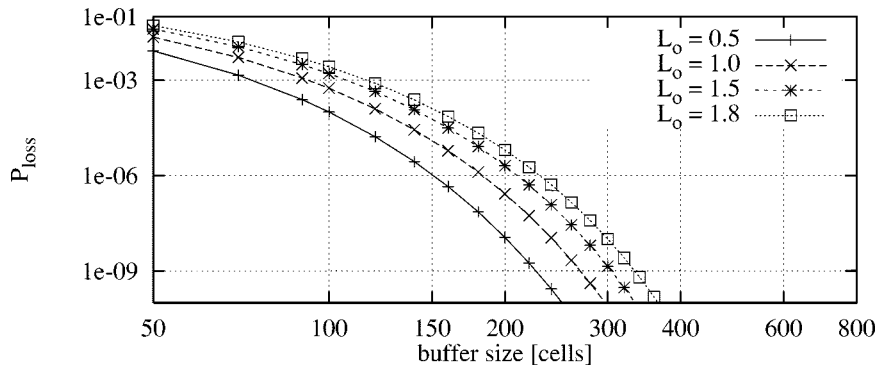


Fig. 12. Cell loss probability versus dedicated upstream handover buffer size, for SCR = 1.9 Mb/s, mean rerouting time $\mu_{UT}^{-1} = 4$ ms, and variable values of average offered load (L_o [Mb/s]).

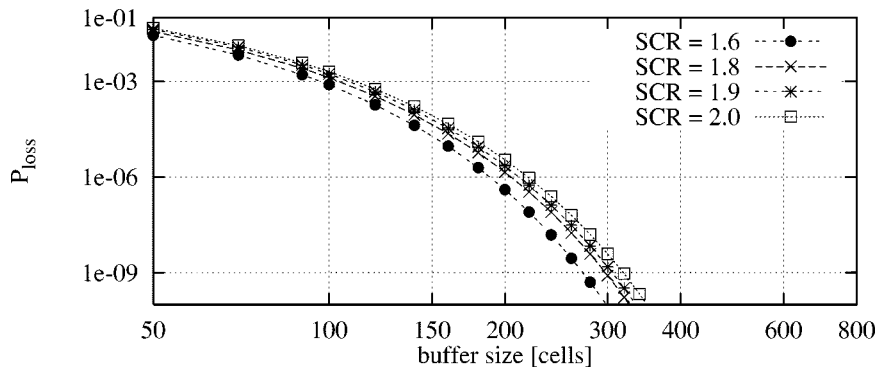


Fig. 13. Cell loss probability versus dedicated upstream handover buffer size for average offered load $L_o = 1.5$ Mb/s, mean rerouting time $\mu_{UT}^{-1} = 4$ ms, and variable values of SCR [Mb/s].

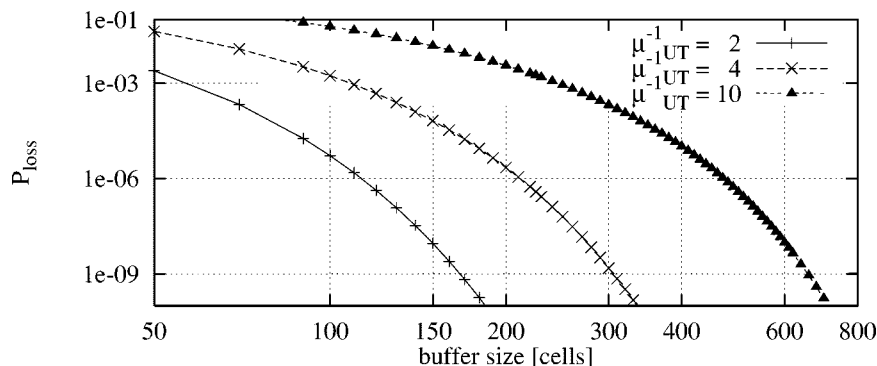


Fig. 14. Cell loss probability versus dedicated upstream handover buffer size for average offered load $L_o = 1.5$ Mb/s, $SCR = 1.9$ Mb/s, and variable values of mean rerouting time μ_{UT}^{-1} [ms].

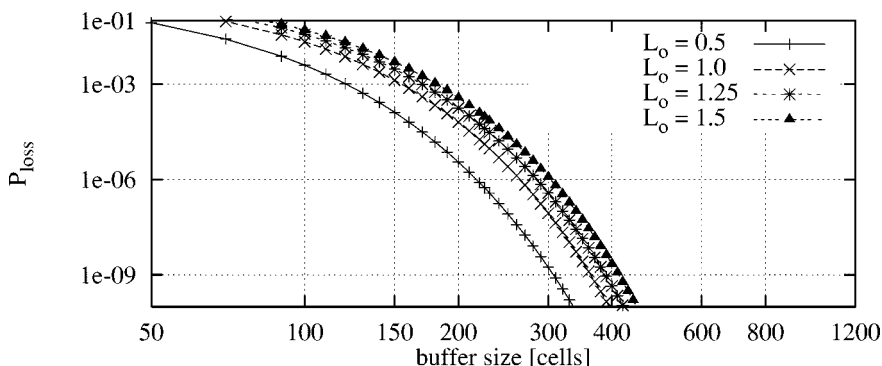


Fig. 15. Cell loss probability versus shared upstream handover buffer size, for $SCR = 1.9$ Mb/s, mean rerouting time $\mu_{UT}^{-1} = 4$ ms, and variable values of average offered load (L_o [Mb/s]).

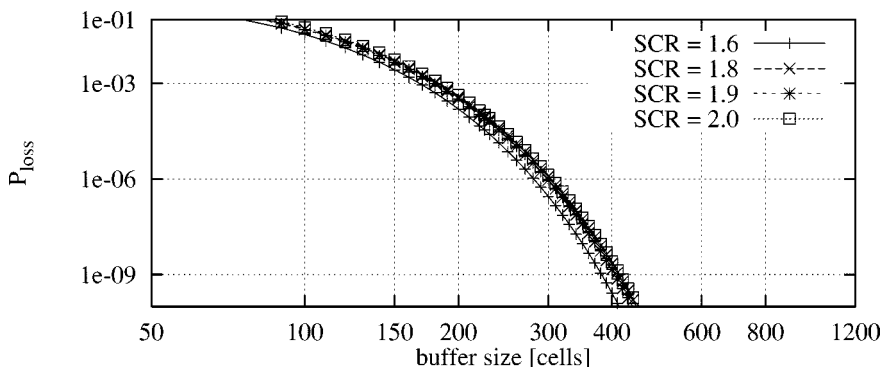


Fig. 16. Cell loss probability versus shared upstream handover buffer size for average offered load $L_o = 1.5$ Mb/s, mean rerouting time $\mu_{UT}^{-1} = 4$ ms, and variable values of SCR [Mb/s].

buffer for $SCR = 1.9$ Mb/s, $\mu_{UT}^{-1} = 4$ ms, and varying values of L_o . Fig. 13 shows the same performance parameter when $L_o = 1.5$ Mb/s, $\mu_{UT}^{-1} = 4$ ms, and SCR varies. Fig. 14 shows similar curves for $SCR = 1.9$ Mb/s, $L_o = 1.5$ Mb/s, and varying values of μ_{UT} . These curves quantify the heavy dependency of the loss probability on the value of μ_{UT} , i.e., the time necessary to reroute the ATM connection.

Notice that once the maximum number of cells stored in the buffer is determined, MBS_{wc} and $h_{UT}(\epsilon)$ can be easily derived using (11) and (12).

Fig. 15 shows the cell loss probability versus the size of the shared upstream buffer when $SCR = 1.9$ Mb/s, and $\mu_{UT}^{-1} = 4$ ms. Varying values of L_o are used to derive the curves. The value for N_{HO} is approximately 3.5. In Fig. 16 similar results

are shown for $L_o = 1.5$ Mb/s, $\mu_{UT}^{-1} = 4$ ms, varying values of SCR , and N_{HO} approximately equal to 3.5.

Finally, Fig. 17 presents the cell loss probability versus the size of the shared upstream buffer for varying values of μ_{UT} , using $L_o = 1.5$ Mb/s, $SCR = 1.9$ Mb/s, with $N_{HO} = 3.5$. Notice that in order to guarantee the required cell loss probability using the dedicated buffer technique, we have to multiply by $N_{HO} = 3.5$ the corresponding size of the dedicated buffer shown in Fig. 14. From this comparison, we see that a cell loss probability in the 10^{-6} range can be achieved by the shared buffer approach using less than half the size of the buffer required in the dedicated buffer approach.

Finally, we point out that the model allows to quantitatively characterize the dependency of the cell loss probability on the

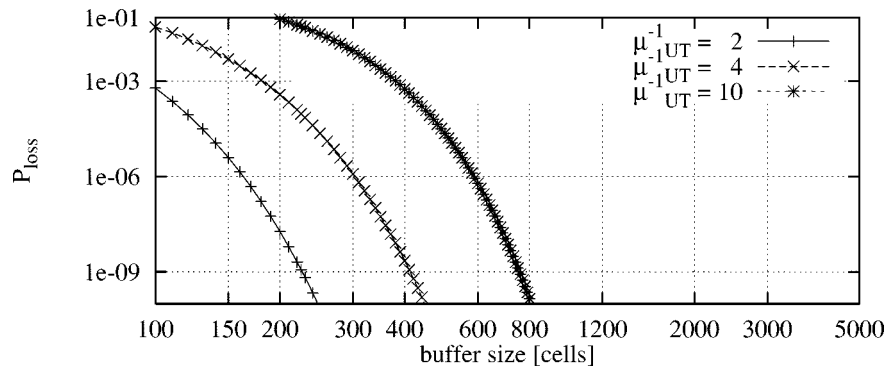


Fig. 17. Cell loss probability versus shared upstream handover buffer size for average offered load $L_o = 1.5$ Mb/s, SCR = 1.9 Mb/s, and variable values of mean rerouting time (μ_{UT}^{-1} [ms]).

connection reestablishment time μ_{UT}^{-1} . With the chosen system configuration and requested cell loss rate, the buffer size necessary at the BS can be significantly reduced if μ_{UT}^{-1} is kept below 4 ms.

VI. SUMMARY

This paper described an approximate and accurate analytical modeling approach to evaluating the performance of hard handover procedures based on cell buffering at the destination base station in WATM networks.

In order to demonstrate the versatility and accuracy of the modeling approach, the performance of two buffering policies were evaluated and results were compared with those produced by detailed simulation experiments.

The accuracy of the analytical models and the considerable number of system parameters taken into account, make the proposed models a practical and flexible framework for estimating: 1) the minimum buffer requirement necessary to satisfy the contracted QoS in WATM networks; 2) the worst-case MBS for the ATM connection; and 3) the maximum handover rate allowed by the system to prevent the cumulative effect of consecutive handovers on the connection delay.

With minor modifications, the proposed models can be adapted to investigate soft handover protocols.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments on an earlier version of this manuscript. They would also like to thank G. Franceschinis and R. Gaeta for their help with the GreatSPN software.

REFERENCES

- [1] D. Raychaudhuri, "Wireless ATM networks: Architecture, system design and prototyping," *IEEE Personal Commun.*, vol. 3, pp. 42–49, Aug. 1996.
- [2] J. Porter, A. Hopper, D. Gilmurray, O. Mason, and J. Naylon, "The ORL radio ATM system, architecture and implementation," Olivetti Oracle Res. Lab., Tech. Rep. 96-5, Jan. 1996.
- [3] M. J. Karol, K. Y. Eng., M. Veeraghavan, and E. Ayanoglu, "BAHAMA: A broadband *ad-hoc* wireless ATM local-area network," in *Proc. ICC'95*, Seattle, WA, June 1995, pp. 1216–1223.
- [4] M. Cheng, B. Rajagopalan, L. F. Chang, and G. P. Pollini, "PCS mobility support over fixed ATM networks," *IEEE Commun. Mag.*, vol. 37, no. 11, Nov. 1997.
- [5] A. Acharya, J. Li, B. Rajagopalan, and D. Raychaudhuri, "Mobility management in wireless ATM networks," *IEEE Commun.*, vol. 35, pp. 100–109, Nov. 1997.
- [6] 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," *ACM/Baltzer MONET*, vol. 1, no. 3, pp. 287–298, Dec. 1996.
- [7] "Requirements for soft handover," in *ATM Forum/97-0696/WATM*, Paris, France, Sept. 1997.
- [8] M. Ajmone Marsan, C. F. Chiasserini, A. Fumagalli, R. Lo Cigno, and M. Munafò, "Local and global handovers based on in-band signaling in wireless ATM networks," *Mobile Computing Commun. Rev. (ACM MC²R)*, vol. 2, no. 3, July 1998.
- [9] B. Banh, G. Anido, and E. Dutkiewicz, "Handover re-routing schemes for connection oriented services in mobile ATM networks," in *Infocom'98*, San Francisco, CA, Mar. 1998, pp. 1139–1146.
- [10] M. Ajmone Marsan, C. F. Chiasserini, A. Fumagalli, R. L. Cigno, and M. Munafò, "Local and global handovers for mobility management in wireless ATM networks," *IEEE Personal Commun.*, vol. 4, pp. 16–24, Oct. 1997.
- [11] C.-K. Toh, "A hybrid handover protocol for local area wireless ATM networks," *ACM/Baltzer MONET*, vol. 1, no. 3, pp. 313–334, Dec. 1996.
- [12] H. Mitts, H. Hansen, J. Immonen, and S. Veikkolainen, "Lossless handover for wireless ATM," *ACM/Baltzer MONET*, vol. 1, no. 3, pp. 299–312, Dec. 1996.
- [13] C. Petri, "Communication with automata," Rome Air Dev. Ctr., New York, Tech. Rep. RADC-TR-65-377, 1966.
- [14] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis, *Modeling with Generalized Stochastic Petri Nets*. New York: Wiley, 1995.
- [15] G. Chiola, C. Dutheillet, G. Franceschinis, and S. Haddad, "Stochastic well-formed colored nets and symmetric modeling applications," *IEEE Trans. Comput.*, vol. 42, pp. 1343–1359, Nov. 1993.
- [16] M. Ajmone Marsan and R. Gaeta, "Modeling ATM systems with GSPNs and SWNs," *ACM Performance Evaluation Rev.*, vol. 26, no. 2, pp. 28–37, Aug. 1998.
- [17] M. Ajmone Marsan, F. Neri, C. Scarpato Cioffari, and A. Vasco, "GSPN models of bridged LAN configurations," *J. Syst. Architecture*, vol. 46, no. 2, pp. 105–130, Jan. 2000.
- [18] M. Ajmone Marsan, C. Casetti, R. Gaeta, and M. Meo, "Performance analysis of TCP connections sharing a congested Internet link," *Performance Evaluation*, vol. 42, no. 2–3, Sept. 2000.
- [19] M. Ajmone Marsan and R. Gaeta, "GSPN models of ATM switches," in *Proc. PNPMP '97*, Saint Malo, France, June 1997, pp. 237–246.
- [20] M. Ajmone Marsan, K. Begain, R. Gaeta, and M. Telek, "GSPN analysis of ABR in ATM LANs," in *Proc. PNPMP '97*, Saint Malo, France, June 1997, pp. 227–236.
- [21] M. Ajmone Marsan, C. F. Chiasserini, and A. Fumagalli, "Dimensioning handover buffers in wireless ATM networks with GSPN models," in *19th Int. Conf. Application and Theory of Petri Nets*, Lisbon, Portugal, June 1998, pp. 44–63.
- [22] M. Ajmone Marsan, M. Meo, and M. Sereno, "GSPN analysis of dual-band mobile telephony networks," in *Proc. PNPMP'99*, Zaragoza, Spain, Sept. 1999, pp. 237–246.
- [23] M. Ajmone Marsan, C. Casetti, R. Gaeta, and M. Meo, "An approximate GSPN model for the accurate performance analysis of correlated TCP connections," in *Proc. SPECTS'00*, Vancouver, BC, Canada, July 2000, pp. 154–162.

- [24] N. A. Anisimov and M. Koutny, "On compositionality and Petri nets in protocol engineering," *Protocol Specification, Testing, Verification XV*, pp. 71–86, 1996.
- [25] F. Bause, H. Kabutz, P. Kemper, and P. Krintzinger, "SDL and Petri net performance analysis of communicating system," *Protocol Specification, Testing, Verification XV*, pp. 269–282, 1996.
- [26] M. Bosch and G. Schmid, "Generic Petri net models of protocol mechanisms in communication systems," *Computer Commun.*, vol. 14, no. 3, pp. 143–156, Apr. 1991.
- [27] M. Diaz, "Petri net based models for the specification and validation of protocols," in *Petri Net Newsletter no. 17*. Bonn, Germany, June 1984, pp. 21–39.
- [28] M. El-Karakasy, M. Reda, A. S. Nouh, and A. R. Al-Obaidan, "Performance analysis of timed Petri net models for communication protocols: A methodology and a package," *Computer Commun.*, vol. 13, no. 2, pp. 73–82, Mar. 1990.
- [29] S. Gordon and J. Billington, "Modeling the WAP transaction service using colored Petri nets," in *Proc. 1st Int. Conf. Mobile Data Access (MDA'99)*. Berlin, Germany, Dec. 1999, vol. 1748, Lecture Notes Comput. Sci., pp. 105–114.
- [30] G. Juanole, Y. Atamna, and R. L. R. Carmo, "On the stochastic timed Petri nets model and its application to the DQDB protocol," *Annales des Telecommunications—Annals of Telecommunications*, vol. 49, no. 5–6, pp. 324–336, 1994.
- [31] M. Li and N. D. Georganas, "Colored generalized stochastic Petri nets for integrated systems protocol performance modeling," *Computer Commun.*, vol. 13, no. 7, pp. 414–424, Sept. 1990.
- [32] S. M. Shatz, T. Suzuki, and T. Murata, "Automated protocol modeling and verification combining an entity-based specification language and Petri nets," in *Proc. 13th Annu. Int. Computer Software and Applications Conf.*, Orlando, FL, Sept. 1989, pp. 580–587.
- [33] R. Y. Awdeh and H. T. Mouftah, "Survey of ATM switch architectures," *Comput. Networks ISDN Syst.*, vol. 27, pp. 1567–1613, 1995.
- [34] Z. Tao and S. Cheng, "A new way to share buffer-grouped input queueing in ATM switching," in *Proc. IEEE Globecom*, San Francisco, CA, Nov. 1994, pp. 475–479.
- [35] The ATM Forum, *ATM UNI Specification, Version 3.1*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
- [36] ATM Forum STR-TM-41.00, "ATM forum traffic management specification," Draft Version 4.1, Feb. 1999.
- [37] M. Proglor, "MBS air interface principles," in *Proc. RACE Mobile Telecom Summit*, Cascais, Portugal, Nov. 1995.
- [38] E. Chlebus, "Analytical grade of service evaluation in cellular mobile systems with respect to subscribers' velocity distribution," in *Proc. Australian Teletraffic Research Seminar*, Melbourne, Dec. 1992, pp. 90–101.
- [39] E. Chlebus and W. Ludwin, "Is handoff traffic really Poissonian?," in *Proc. IEEE ICUPC'95*, Tokyo, Japan, Nov. 1995, pp. 348–353.
- [40] G. Chiola, G. Franceschinis, R. Gaeta, and M. Ribaud, "GreatSPN 1.7: Graphical editor and analyzer for timed and stochastic Petri nets," *Performance Evaluation*, vol. 24, pp. 47–68, Nov. 1995.
- [41] M. Ajmone Marsan, A. Bianco, T. V. Do, L. Jereb, R. Lo Cigno, and M. Munafò, "ATM simulation with CLASS," *Performance Evaluation*, vol. 24, pp. 137–159, Nov. 1995.

- [42] R. Ramjee, "Supporting connection mobility in wireless networks," Ph.D. dissertation, Univ. Massachusetts, Amherst, May 1997.



Marco Ajmone Marsan (S'76–M'78–SM'86–F'99) holds degrees from the Politecnico di Torino, Italy, and University of California, Los Angeles.

He is a Full Professor at the Electronics Department of Politecnico di Torino, in Italy. He has coauthored over 250 journal and conference papers in the areas of communications and computer science, as well as *Performance Models of Multiprocessor Systems* (Cambridge, MA: MIT Press, 1986) and *Modeling with Generalized Stochastic Petri Nets* (New York: Wiley, 1995). His current interests are in the

fields of performance evaluation of communication networks and their protocols.

Dr. Marsan received the Best Paper Award at the 3rd International Conference on Distributed Computing Systems, Miami, FL, in 1982.



Carla-Fabiana Chiasserini (S'98–M'00) received the Laurea degree in electrical engineering from the University of Florence, Italy, in 1996 and the Ph.D. degree from Politecnico di Torino, Italy, in 1999.

Since then, she has been with the Department of Electrical Engineering, Politecnico di Torino, where she is currently an Assistant Professor. She was with 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 wire-

less networks.



Andrea Fumagalli (M'99) received the Ph.D. degree 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, Amherst. From 1992 to 1998, he was on the Faculty of the Electrical Engineering Department, Politecnico di Torino. Since 1997, he has been Associate Professor of electrical engineering at the University of Texas at Dallas. Primarily dealing with network architectures and protocols, his 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.